Potentially translated sequences determine protein-coding potential of RNAs

in cellular organisms

Yusuke Suenaga^{1,†,*}, Mamoru Kato^{2,†}, Momoko Nagai², Kazuma Nakatani^{1,3,4}, Hiroyuki Kogashi¹, Miho Kobatake¹, Takashi Makino⁵

¹Department of Molecular Carcinogenesis, Chiba Cancer Centre Research Institute, 666-2 Nitona, Chuo-ku, Chiba 260-8717, Japan

²Division of Bioinformatics, National Cancer Centre Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

³Department of Molecular Biology and Oncology, Chiba University School of Medicine, 666-2 Nitona, Chuo-ku, Chiba 260-8717, Japan

⁴Innovative Medicine CHIBA Doctoral WISE Program

⁵Laboratory of Evolutionary Genomics, Graduate School of Life Sciences, Tohoku University, Aobayama Campus 6-3, Aramaki Aza-Aoba, Aoba-ku, Sendai 980-8578, Japan

[†]These authors equally contributed to this work

***Correspondence to:**

Yusuke Suenaga

E-mail: ysuenaga@chiba-cc.jp

23 Abstract

24 RNA sequence characteristics determine whether their transcripts are coding or
 25 noncoding. Recent studies have shown that, paradoxical to the definition of noncoding
 26 RNA, several long noncoding RNAs (lncRNAs) translate functional peptides/proteins.
 27 However, the characteristics of RNA sequences that distinguish such newly identified
 28 coding transcripts from lncRNAs remain largely unknown. In this study, we found that
 29 potentially translated sequences in RNAs determine the protein-coding potential of
 30 RNAs in cellular organisms. We defined the potentially translated island (PTI) score as
 31 the fraction of the length of the longest potentially translated region among all regions.
 32 To analyze its relationship with protein-coding potential, we calculated the PTI scores
 33 in 3.4 million RNA transcripts from 100 cellular organisms, including 5 bacteria, 10
 34 archaea, and 85 eukaryotes, as well as 105 positive-sense single-strand RNA virus
 35 genomes. In bacteria and archaea, coding and noncoding transcripts exclusively
 36 presented high and low PTI scores, respectively, whereas those of eukaryotic coding
 37 and noncoding transcripts showed relatively broader distributions. The relationship
 38 between the PTI score and protein-coding potential was sigmoidal in most eukaryotes;
 39 however, it was linear passing through the origin in three distinct eutherian lineages,
 40 including humans. The RNA sequences of virus genomes appeared to adapt to
 41 translation systems of host organisms by maximizing protein-coding potential in host
 42 cells. Hence, the PTIs determined the protein-coding potential of RNAs in cellular
 43 organisms. Additionally, coding and noncoding RNA do not exhibit dichotomous

sequence characteristics in eukaryotes, instead they exhibit a gradient of protein-coding potential.

Introduction

Recent advances in sequencing technology have revealed that most of the eukaryotic genome is transcribed, resulting in production of noncoding RNAs (Okazaki et al. 2002; Djebali et al. 2012; Ulitsky and Bartel 2013; Kopp et al. 2018). Noncoding RNAs that are more than 200 nucleotides in length are referred to as long noncoding RNAs (lncRNAs) and are not translated into proteins (Ulitsky and Bartel 2013; Kopp et al. 2018). LncRNAs have been reported to function in various biological phenomena, such as the regulation of transcription, modulation of protein or RNA functions, and organization of the nuclear architecture (Ulitsky and Bartel 2013; Kopp et al. 2018). However, paradoxical to its definition, a large fraction of lncRNAs are associated with ribosomes and translated into peptides (Frith et al. 2006; Ingolia et al. 2011; Bazzini et al. 2014; Ingolia et al. 2014; Ruiz-Orera et al. 2014). Peptides translated from transcripts annotated as lncRNAs have been shown to have biological functions in various eukaryotes (Li and Liu 2019). Moreover, known protein-coding genes, such as *TP53*, act as functional RNAs (Candeias 2011; Kloc et al. 2011). Detection of these RNAs with binary functions has blurred the distinction between coding and noncoding RNAs, and the characteristics of RNA sequences that explain the continuity between noncoding and coding transcripts remain unclear.

During evolution, new genes originated from pre-existing genes via gene duplication or from non-genic regions via the generation of new open reading frames (ORFs) (Ohno 1970; Chen et al. 2013; Zhang and Long 2014; McLysaght and Guerzoni 2015; MyLysaght and Hurst 2016; Holland et al. 2017). Newly evolved genes from non-genic regions are known as *de novo* genes (Begun et al. 2006; Levine et al. 2006; Begun et al. 2007; Knowles and McLysaght 2009; Li et al. 2009; Toll-Riera et al. 2009; Li et al. 2010), and regulate phenotypes and diseases (McLysaght and Guerzoni 2015; Chen et al. 2013; Zhang and Long 2014), such as brain function and carcinogenesis in humans (Li et al. 2010; Suenaga et al. 2014). LncRNAs serve as sources of newly evolving gene products (Ruiz-Orera et al. 2014), including *de novo* evolved proteins. Not only ORFs exposed to natural selection, neutrally evolving ORFs are also translated from lncRNAs to stably express peptides (Ruiz-Orera et al. 2018), providing a foundation for the development of new functional peptides/proteins. High levels of lncRNA expression (Ruiz-Orera et al. 2018), hexamer frequencies of ORFs (Sun et al. 2013; Wang et al. 2013; Ruiz-Orera et al. 2014), and high flexibility of peptides (Wilson et al. 2017) have been proposed as determinants of coding potential; however, the molecular mechanisms of how lncRNAs evolve into new coding transcripts remain unclear (Van Oss and Carvunis 2019).

This study, therefore, sought to identify a new indicator to determine the protein-coding potential of RNAs. Within an RNA sequence, we termed sequence segments that start with AUG (the base sequence of the start codon) and end with UAG, UGA, or UAA (base sequences of stop codons) as potentially translated islands (PTIs). First, we

defined this indicator using PTI lengths and subsequently examined associations between the indicator and protein-coding potential. We also present analyses of more than 3.4 million transcripts in 100 cellular organisms belonging to all three domains of life to investigate the evolution of the relationship between the PTI score and protein-coding potential. Finally, we examined whether virus RNA genomes have differentially evolved to maximize the protein-coding potential in different host organisms.

Results

Coding transcripts show higher PTI scores in humans and mice

We defined PTI score as described in the Materials and Methods and illustrated in Figure 1A to B, and analyzed human transcripts registered in the nucleotide database of the National Centre for Biotechnology Information for coding (RefSeq ID starting with NM) and noncoding (RefSeq ID starting with NR) RNA transcripts. Sample PTI score calculation is presented in Supplementary Figure 1. The nucleic acid sequences and IDs were downloaded from Table browser (<https://genome.ucsc.edu/cgi-bin/hgTables>) after setting the track tab as “RefSeq Genes.” A total of 50,052 coding (NM) and 13,550 noncoding (NR) RNAs were registered in 2018 (Supplementary Table 1). To analyze putative lncRNAs with protein-coding potential, we excluded small RNAs (shorter than 200 bp) or RNAs with a short primary PTI (pPTI; shorter than 20 amino acids) from the NR transcripts, focusing on the remaining 12,827 transcripts.

We analyzed relative frequencies of NM and NR transcripts, designated as $f(x)$ and $g(x)$ respectively (Figure 1C), where x indicates PTI score. See Materials and Methods for the definition of $f(x)$ and $g(x)$.

In human transcripts, $g(x)$ showed a distribution that shifted to the left with an apex of 0.15; in contrast, distribution of $f(x)$ shifted to the right with an apex of 0.55 (Figure 1C, upper panel). As a control, we used randomly generated nucleic acid sequences where A/T/G/C bases were randomly assigned with equal probabilities. In the controls, the relative frequencies of PTI scores were shifted to the left in both coding and noncoding transcripts (Figure 1C, bottom panel). The controls that randomly shuffled the original sequence without affecting the number of A/T/C/G bases in each transcript also had relative frequencies of PTI scores shifted to the left in both coding and noncoding transcripts (Supplementary Figure 2A). Moreover, similar results were obtained using a data set from the Ensembl database (Supplementary Figure 2B). We also calculated the PTI scores of mouse transcripts registered in RefSeq and Ensembl and found that the distribution of $f(x)$ was shifted to the right with an apex of 0.55 (Supplementary Figure 2C), similar to that of human transcripts. These results suggest that the sequences, not lengths, of the coding transcripts increase PTI scores in mice and humans.

PTI scores correlate with protein-coding potential in humans and mice

Next, we examined the relationship between the PTI score and protein-coding potential of genes. Protein-coding potential $F(x)$ is defined in the Materials and Methods and a sample $F(0.15)$ calculation in human transcripts is presented in Figure 1D. This result

indicates that any given human RNA transcript with a calculated PTI score of 0.15 has a protein-coding potential $F(x)$ of 0.183. $F(x)$ was correlated with PTI score at PTI scores ≤ 0.65 (Figure 1E and Supplementary Figure 3A). The protein-coding potential of sequences in RefSeq data slightly decreased after peaking at 0.65 (Figure 1E) whereas that of sequences in the Ensembl data remained high (Supplementary Figure 3A). $F(x)$ of human transcripts was well approximated by the linear regression as follows:

Based on Ensembl data

$$F(x) = 1.301x + 0.0072 \quad (x \leq 0.65), R^2 = 0.984$$

Based on RefSeq data

$$F(x) = 1.313x + 0.0189 \quad (x \leq 0.65), R^2 = 0.990$$

The intercepts were near zero, and slope was approximately 1.3. By using these formulas, we can calculate the protein-coding potential $F(x)$ of any given human transcript with a PTI score ≤ 0.65 . For example, the protein-coding potential $F(x)$ of *NCYM* was calculated as 0.746 or 0.765 based on Ensembl or RefSeq databases, respectively (Supplementary Figure 1C). In contrast, $F(x)$ in the controls had no positive correlation with PTI scores (Figure 1E, bottom panel, and Supplementary Figure 3A). Similar results were obtained for mouse transcripts (Supplementary Figures 3B). $F(x)$ of mouse transcripts (PTI score ≤ 0.65) was approximated as follows.

Based on Ensembl data

$$F(x) = 1.142x + 0.067, R^2 = 0.982$$

Based on RefSeq data

$$F(x) = 1.482x - 0.061, R^2 = 0.990$$

In both human and mouse transcripts, the PTI score correlated linearly with the protein-coding potential at PTI scores ≤ 0.65 . Moreover, when the PTI score limit approached 0, the probability of the transcript being a coding RNA was 0.

PTIs affect the protein-coding potential predicted by Ka/Ks

To examine the relationship between the PTI score and natural selection in the prediction of protein-coding potential, we calculated the ratio of nonsynonymous (Ka) to synonymous (Ks) values by comparing human transcripts with syntenic genome regions of chimpanzee or mouse (Figure 1F). Transcripts were selected based on the syntenically conserved regions, 44,593 (vs. chimp) and 14,016 (vs. mouse). The results revealed linear relationships between the $F(x)$ and PTI score in the conserved transcripts (Figure 1F, left panels). As predicted, coding transcripts contained transcripts with $Ka/Ks < 0.5$ at a higher frequency than noncoding transcripts, with the largest difference observed when the PTI score was > 0.9 , and the smallest difference at PTI scores of approximately 0.45 (Figure 1F, right panels).

Therefore, noncoding transcripts showing both negative selection ($Ka/Ks < 0.5$) and the highest PTI score may include new coding transcript candidates.

We have listed 23 such transcripts (Supplementary Table 2), which include four transcript variants of a previously identified lncRNA that encodes a tumor suppressive small peptide, HOXB-AS3 (Huang et al 2017).

Characterization of high PTI score lncRNAs in humans

Next, we investigated whether the PTI score is useful for identifying coding RNAs among NR transcripts. From the 7,144 transcripts registered as noncoding genes until 2015, we excluded small RNAs (< 200 nucleotides) and those with short primary PTIs (< 20 amino acids). Among the remaining 6,617 NR genes, 219 were reassigned as NM over the past 3 years (Supplementary Table 3), including the previously identified *de novo* gene *MYCNOS/NCYM* (Suenaga et al. 2014). The percentage of reclassification increased among NR transcripts with high PTI scores (Figure 1G). Thus, a high PTI score is a useful indicator of coding transcripts. NR transcripts with high protein-coding potential ($0.6 \leq \text{PTI score} < 0.8$) were then extracted, and the domain structure of the pPTI amino acid sequence was estimated using BLASTP (See the Materials and Methods for details). A total of 217 transcripts showed a putative domain structure(s) in pPTI, whereas 310 transcripts showed none (Supplementary Table 4). Transcripts with a domain structure are often derived from transcript variants, pseudogenes, or readthrough of coding genes; transcripts without domain structures are often derived from antisense RNA or long intergenic noncoding RNA (lincRNA) (Table 1).

We next examined the functions of genes generating NR transcripts with high coding probabilities ($0.6 \leq \text{PTI score} < 0.8$). We divided the NR transcripts into those with or without putative domains to investigate candidates of novel coding genes, either originating from pre-existing genes or created from non-genic regions, respectively. Analysis using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) Functional Annotation Tool (Huang et al. 2009a, 2009b) showed that NR transcripts without domain structures are derived from original genes related to transcriptional regulation, multicellular organismal process, and developmental process (Supplementary Table 5; see Materials and Methods for the definition of original genes). Among the target genes of transcription factors, NMYC, TGIF, and ZIC2 were ranked in the top three and are all necessary for forebrain development (Supplementary Table 5) (Brown et al. 1998; Gripp et al. 2000; van Bokhoven et al. 2005). We observed that NR transcripts with domain structures originated from genes that undergo alternative splicing which are related to organelle function and are expressed in various cancers, including respiratory tract tumors, gastrointestinal tumors, retinoblastomas, and medulloblastomas (Supplementary Table 6).

Relationship between PTI score and relative frequencies of coding/noncoding transcripts in 100 cellular organisms

To analyze the relationship between PTI scores and protein-coding potential in a broad lineage of cellular organisms, we selected 100 organisms, consisting of 5 bacteria, 10 archaea and 85 eukaryotes (Supplementary Table 1), and calculated the PTI score for a

total of more than 3.4 million transcripts (Supplementary Table 1). Phylogenetic trees of the cellular organisms are presented on a logarithmic time scale along with the number of species in each lineage used in the analyses (Figure 2). To examine the evolutionary conservation of the linear relationship between PTI score and protein-coding potential in humans and mice, we selected a relatively large number of species (36) from mammals. The species with fewer than three lncRNAs were not used to calculate $g(x)$ and were not included in the histograms illustrating the relationship with PTI score (Figure 3). In all organisms, the relative frequency of coding transcripts $f(x)$ was shifted to the right (higher PTI score) compared to random or random shuffling controls (Figure 3, Figure 4, Supplementary Figure 4 and Supplementary Figure 5).

In bacteria and archaea, $f(x)$ and $g(x)$ exclusively presented high and low PTI scores, respectively, indicating a clear boundary between the coding transcripts and lncRNAs in terms of PTI scores (Figure 3 and Supplementary Figure 4). In addition, the highest frequency of coding transcript $f(x)$ presenting a PTI score was 0.75 in all examined bacteria (Figure 3) and ≥ 0.75 in archaea (Supplementary Figure 4). Among eukaryotes, unicellular organisms and non-vertebrates showed the highest frequency of coding transcripts at 0.65 or 0.75 (Figure 3), while most vertebrates showed highest values ≤ 0.65 (Figure 3 and 4). In addition, the $f(x)$ distribution in vertebrates was broad and shifted to the left (lower PTI scores) compared to bacteria and archaea (Figure 3 and 4). In sharp contrast to $f(x)$, the relative frequency of lncRNAs $g(x)$ was shifted to the right (higher PTI scores) in eukaryotes, including *G. lamblia* that belongs to the earliest diverging eukaryotic lineage and lacks mitochondria (Figure 3). Since the distribution

of $f(x)$ in the Excavata, including *G. lamblia*, showed a similar pattern to bacteria, the right shift of $g(x)$ seems to be an earlier event than the left shift of $f(x)$ in the evolution of eukaryotes. Collectively, the right and left shifts of $f(x)$ and $g(x)$ contribute to blurring the boundary between coding and noncoding transcripts in eukaryotes.

Relationship between the PTI score and protein-coding potential

The overlapping of relative frequencies in $f(x)$ and $g(x)$ led us to examine the relationship between PTI score and protein-coding potential $F(x)$ in eukaryotes. To avoid misleading data obtained by small sampling numbers, we selected 32 species with more than 1000 lncRNAs that contain pPTIs for calculation of $F(x)$ (Figure 5 and Supplementary Figure 6). In humans and mice, the relationship between PTI score and $F(x)$ was approximated with the linear function passing through the origin of the PTI score ≤ 0.65 . Therefore, we used linear approximation in the $F(x)$ of 32 species and divided them into two groups, linear and sigmoidal, based on the shape and formula of the approximated function (Figure 5). We defined the linear group as $R^2 > 0.9$ with an absolute value of the intercept < 0.1 ; the sigmoidal group had a slope > 1.0 and intercept < -0.1 . In *U. americanus*, *C. canadensis*, and *G. gorilla*, fewer than five lncRNAs exhibited PTI scores of 0.05; thus, we eliminated the $F(0.05)$ in these species for the approximation by linear function (indicated with asterisks in Figure 5). The five species that did not fit in the linear or sigmoidal group were characterized by high $F(x)$ in low PTI scores, and belonged to plants (*Z. mays*), reptiles (*A. Carolinensis*) and mammals (*O. anatinu*, *S. boliviensis*, and *G. gorilla*) (Figure 5). In these species, PTI scores

showed weaker association with protein-coding potential. Sigmoidal relationships were observed in 18/32 species, while linear relationships were apparent in nine species within three mammalian lineages, Cetartiodactyla, Rodentia and Primates. Since the sigmoidal relationship was the most broadly observed across the examined lineages, this relationship appears to be of an ancestral type.

Characteristics of RNA virus genomes in human and bacterial cells

In sharp contrast to the coding transcripts of bacteria and archaea, the PTI score in coding transcripts of eukaryotes exhibited overlap with that of noncoding RNA by its broad distribution of low PTI scores. To investigate the molecular mechanism underlying the distinct distribution of coding transcripts between bacteria and eukaryotes, we analyzed genome sequences of RNA viruses that infect human or bacterial cells. Positive-sense single-strand RNAs, (+) ssRNAs, are parts of virus genomes that generate mRNAs and are translated into viral proteins via the host translation system. Therefore, efficient translation in host cells contributes to the replication of the (+) ssRNA viruses. We speculated that PTIs, other than bona-fide ORFs, affect the coding potential of the viral genome in host cells. In a viral genome, multiple bona-fide ORFs exist. We thus extended the concept of PTIs (Figure 6A) to the multiple ORFs in a viral RNA genome and set the viral ORF (vORF) score (see Materials and Methods for the definition of the vORF score). Among the positive-sense ssRNA viruses registered in the NCBI database, 198 were human viruses and 13 were bacteriophages. We eliminated the viruses that translate

virus proteins by exceptional translation mechanisms such as ribosome frameshifting, alternative initiation sites, ribosome slippage, and RNA editing, and focused on the remaining 95 human viruses including 9 retroviruses (Supplementary Table 7) and 10 bacteriophages (Supplementary Table 8). The relative frequency of the human viruses and bacteriophages showed distinct peaks at PTI scores of 0.65 and 0.75, respectively (Figure 6B). These values correspond to the PTI scores of the highest protein-coding potential in humans (Figure 1E and Supplementary Figure 3A) and the highest frequency of coding transcripts in bacteria (Figure 3), respectively. In addition, the relative frequency of human viruses showed broader distribution of low PTI scores compared to bacteriophages, particularly in human retroviruses (Figure 6B). Therefore, the RNA genomes of viruses appear to have sequence characteristics that maximize protein-coding potential in host cells.

Discussion

Here, we showed that the PTI is associated with protein-coding potential in cellular organisms. In bacteria and archaea, the distribution of noncoding and coding transcripts separately presented at low and high PTI scores, whereas they were merged in eukaryotes. The overlapping distribution of noncoding and coding RNA in eukaryotes is caused by the right and left distribution shifts of noncoding and coding transcripts, respectively. The right shifts in the distribution of noncoding RNA occurred for *G. lamblia*, one of the earliest diverging eukaryotes that contain two nuclei and lack mitochondria,

peroxisomes, and a typical Golgi apparatus (Ankarklev et al. 2010; Bartelt et al. 2015; Buret et al. 2020), and were commonly observed in all examined eukaryotes. The most prominent difference between eukaryotes and bacteria/archaea is the existence of nuclei in cells, as this is the definition of eukaryotes. Moreover, functional noncoding RNAs, by definition, should not be translated by ribosomes in cells. However, in bacteria and archaea, newly transcribed RNAs are immediately bound by ribosomes (Miller et al. 1970; French et al. 2007) and do not have the chance to escape from translation. Alternatively, in eukaryotes, the existence of nuclei prevents the immediate binding of lncRNAs by ribosomes and cytosolic translocation is required for translation. Therefore, the eukaryotic lncRNAs have a chance to function in the nucleus regardless of PTI scores or translation efficiencies. This nuclear localization likely promotes diversification of RNA sequences, resulting in the right shift in the distribution of noncoding RNAs. Thus, the pervasive transcription of the genome seems to be advantageous for eukaryotes to create new functional nuclear RNAs while being disadvantageous for bacteria/archaea by increasing the risk of transcription of a high PTI score transcript, leading to immediate translation of wasteful and/or toxic proteins (Monsellier et al. 2007).

The left shifts observed in distribution of coding RNAs were not observed in *G. lamblia* with its peak at 0.75, however, the majority of fungi, plants and animals, including humans, exhibited this shift with peaks at PTI scores ≤ 0.65 . Thus, the left shift of coding transcripts likely occurred after the right shift of noncoding transcripts in evolution of eukaryotes.

326 Relative frequencies of positive-sense ssRNA viruses exhibit sharp peaks at vORF
327 scores of 0.75 in bacteriophages and 0.65 in human viruses, indicating the adaptation of
328 RNA viruses to host cells by maximizing the protein-coding potential of their genomes.
329 Immediately after viral infection, the viral (+) ssRNA genomes, save for those of
330 retroviruses, are used as templates for translation in host cytosol. Thus, the distinct
331 translation systems between humans and bacteria likely affected the left shift in the viral
332 genome peak, as well as the left shift in coding RNA distribution in eukaryotes. In a
333 retrovirus, reverse transcriptase produces double-stranded DNA using the viral genome
334 as a template, which is then inserted into the host genome. The viral genome is
335 subsequently transcribed within the nucleus and the mRNA is transported to the
336 cytoplasm where protein products become translated in a manner similar to host
337 proteins. Therefore, the relatively lower vORF score distribution in human retrovirus
338 genomes is likely caused by the nuclear localization of the RNA genome, avoiding
339 immediate binding by ribosomes.

340 In eukaryotes, we calculated protein-coding potential $F(x)$ based on the overlapped
341 relative frequencies of noncoding $g(x)$ and coding transcripts $f(x)$. The relationship
342 between protein-coding potential and PTI score was divided into three groups,
343 sigmoidal, linear, and others. Among them, switch-type sigmoidal relationships seem to
344 be of the ancestral-type based on their conservation in eukaryotes, and to emerge after
345 the all-or-none-type relationships in bacteria and archaea. Meanwhile, the linear group
346 showed relatively high protein-coding potential at low or intermediate PTI scores,
347 further blurring the boundary between noncoding and coding transcripts. For example,

the PTI score of human *TP53* (NM_000546.5) is 0.476, and the $F(x)$ calculated by the linear function is 0.644. The PTI score of *TP53* indicates that the transcript has a high probability of functioning as both a coding and noncoding RNA, explaining the bifunction of *TP53* (Candeias 2011; Kloc et al., 2011). Further studies are required to explain the molecular mechanisms underlying the existence of these three relationships in eukaryotes.

The PTI score is a value uniquely calculated based on any given RNA sequence, whereas the Ka/Ks value requires comparison with a species and varies across different species. Therefore, the Ka/Ks value cannot be uniquely calculated from a given RNA sequence. Hence, PTI scores, but not Ka/Ks values, exhibit species-specific relationships with unique distributions of coding potential. Due to the advantages of the PTI score, we can investigate evolutionary changes in the relationship between PTI score and coding potential. Furthermore, $Ka/Ks < 0.5$ did not show different relative frequencies of noncoding and coding transcripts at a PTI score of approximately 0.45, suggesting that Ka/Ks does not predict the coding potential at the PTI score and that the prediction is dependent on the PTI score. In addition, to calculate Ka/Ks , orthologous transcripts or genomes should be conserved among species. Therefore, Ka/Ks cannot be applied for coding prediction of species-specific transcripts such as *NCYM*. On the other hand, by using approximate functions, we can calculate the coding potential $F(x)$ of any given transcript with a PTI score ≤ 0.65 in the nine species classified to linear group.

Human NR transcripts with high PTI scores have been reclassified as coding genes over the past 3 years, including the human *de novo* gene *NCYM* (Suenaga et al. 2014;

Suenaga et al. 2020). Because *de novo* gene products have no known domain structures, high PTI score-NR transcripts without putative domains, may be good candidates as novel *de novo* genes in eukaryotes. *NCYM* is an antisense gene of *MYCN* whose protein product stabilizes *MYCN* (Suenaga et al. 2014). *MYCN* directly stimulates *NCYM* and *OCT4* transcription, whereas *OCT4* induces *MYCN* (Kaneko et al. 2015). This functional interplay forms a positive feedback loop for these genes to induce each other in human neuroblastomas (Suenaga et al. 2014; Islam et al. 2015; Kaneko et al. 2015; Shoji et al. 2015). Functional annotation of NR genes without putative domains was related to transcriptional regulation, and the target genes of transcription factors, including *MYCN*, *TGIF*, and *ZIC2*, were enriched. As *de novo* emergence of *NCYM* occurred in homininae, *NCYM*-mediated *MYCN* activation may modulate human *de novo* gene births during evolution, regulating the transcription of *MYCN* target genes. Both *TGIF* and *ZIC2* are causative genes of holoprosencephaly, a disorder caused by a failure in embryonic forebrain development (Brown et al. 1998; Gripp et al. 2000), whereas *MYCN* is the causative gene of Feingold syndrome and megalencephaly syndrome, which are associated with a reduced and increased brain size, respectively (van Bokhoven et al. 2005; Kato et al. 2019). Thus, the present study also provides a list of candidate human *de novo* genes possibly involved in brain development and brain-related diseases.

In conclusion, we identified a novel determinant of protein-coding potential in cellular organisms. The relationship between PTI score and protein-coding potential revealed that the boundary between coding and noncoding transcripts in bacteria and archaea is

blurred in eukaryotes. Therefore, when a eukaryotic transcript possesses a moderate PTI score, bifunctional characterization as both a coding and noncoding transcript may be essential for a full understanding of the biological roles of RNAs.

Materials and Methods

Potentially translated islands

Definition

PTIs are defined as sequence segments beginning at AUG and ending with any of UAA, UAG, or UGA in the 5' to 3' direction within an RNA sequence in all three possible reading frames (Figure 1A).

Example

As an example, the PTIs in the human *de novo* gene *NCYM* (Suenaga et al. 2014) are identified using the cDNA sequence (Supplementary Figure 1A) and are shown as bold characters (Supplementary Figure 1B). Further information is included in Supplementary Notes.

The length of PTI and primary/secondary PTIs

Definition

The length of PTI is the length of the amino acid sequence excluding the stop codon and is represented by l (Figure 1A). In an RNA sequence, the longest PTI is designated as the primary PTI (pPTI), whereas the others are termed secondary PTIs (sPTIs). The lengths of a pPTI and sPTI are described as l_{pPTI} and l_{sPTI} , respectively (Figure 1A).

Example

Based on these definitions, the shortest PTI is “AUGUAA,” “AUGUAG,” or “AUGUGA,” and its amino acid length is 1. For example, the NCYM transcript has a pPTI with a length of 109 at frame 1; three sPTIs with lengths of 69, 8 and 6, respectively, at frame 2; and no PTIs at frame 3 (Supplementary Figure 1B and C).

Characteristics

Therefore, the following relationship between the lengths of pPTI and sPTI is held:

$$1 \leq l_{\text{sPTI}} \leq l_{\text{pPTI}} \quad (1)$$

PTI score

Definition

The definition of PTI score is motivated by our hypothetical concept that translation of pPTI is limited by alternate competing sPTIs. We defined the PTI score (Figure 1A) according to Equation 2 and 3:

$$\sum_{i=1}^n l_{\text{sPTI}i} = l_{\text{sPTI}1} + l_{\text{sPTI}2} + \cdots + l_{\text{sPTI}k} + \cdots + l_{\text{sPTI}n} \quad (2)$$

$$\text{PTI score} = \frac{l_{\text{pPTI}}}{l_{\text{pPTI}} + \sum_{i=1}^n l_{\text{sPTI}i}} \quad (3)$$

where, $l_{\text{pPTI}} + \sum_{i=1}^n l_{\text{sPTI}i}$ represents the sum of all PTI lengths.

Example

If an RNA sequence has only one PTI, the PTI score is 1 (Figure 1B). An RNA sequence with many sPTIs tends to have a score close to 0 (Figure 1B). If the sum of all sPTI length is equal to pPTI length, the PTI score is 0.5 (Figure 1B). The PTI score of

the NCYM transcript is calculated as 0.568 (Supplementary Figure 1C). Further information is included in Supplementary Notes.

Characteristics

Therefore, the range of the PTI score is as follows:

$$0 < \text{PTI score} \leq 1 \quad (4)$$

Relative frequencies $f(x)$ and $g(x)$

Definition

We defined the relative frequencies of coding and noncoding transcripts, $f(x)$ and $g(x)$, as follows (Figure 1C):

$$f(x) = \frac{NM(x)}{TNM} \quad (5)$$

$$g(x) = \frac{NR(x)}{TNR} \quad (6)$$

where, TNM and TNR represent the total number of coding and noncoding transcripts, respectively, excluding the transcripts lacking PTIs. $NM(x)$ or $NR(x)$ is the number of coding or noncoding transcripts having PTI score = x , respectively.

To define coding/non-coding transcripts with a PTI score of x , we made histograms divided by ten classes, and the median values of the classes were used to represent the PTI score (Figure 1C). Therefore, in Equation 5 and 6, the PTI score x is restricted as follows:

$$x = 0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, \text{ or } 0.95 \quad (7)$$

Characteristics

Thus, the $f(x)$ and $g(x)$ follow Equation 8 to 11:

$$0 \leq f(x) \leq 1 \quad (8)$$

$$0 \leq g(x) \leq 1 \quad (9)$$

$$\sum_x f(x) = 1 \quad (10)$$

$$\sum_x g(x) = 1 \quad (11)$$

Protein-coding potential $F(x)$

Definition

The protein-coding potential, $F(x)$, was calculated according to Equation 12:

$$F(x) = \frac{f(x)}{f(x) + g(x)} \quad (12)$$

Example

As an example, $F(0.15)$ in human transcripts is depicted in Figure 1D. $F(0.15)$ is calculated using Equation (12):

$$f(0.15) = 0.060$$

$$g(0.15) = 0.268$$

$$F(0.15) = \frac{f(0.15)}{f(0.15) + g(0.15)} = \frac{0.060}{0.060 + 0.268} = 0.18292 = 0.183$$

Identification of human noncoding transcripts with high protein-coding potential

To identify the noncoding transcripts with high protein-coding potential, NR transcripts with high $F(x)$ ($0.6 \leq x < 0.8$) were identified from total NR transcripts registered in

the nucleotide database of the National Centre for Biotechnology Information. NR transcripts shorter than 200 nucleotides or with pPTI encoding putative peptides with fewer than 20 amino acids were excluded. The amino acid sequences of pPTIs in these transcripts were subjected to a BLASTP search to examine the presence of putative domain structures. In the BLASTP search, non-redundant protein sequences (nr) were applied as the search set, and Quick BLASTP (accelerated protein-protein BLAST) was used as the algorithm. In the search results, putative conserved domains or the message “No putative conserved domains have been detected” were shown in the Graphical Summary tab. CDSEARCH/cdd was the database employed to search conserved domain structures using the default settings; low complexity filter: no; composition based adjustment: yes; E-value threshold: 0.01; maximum number of hits: 500. Based on the results, transcripts with or without putative conserved domain structure were indicated as + or -, respectively.

Functional annotation of original genes

Original genes are defined as the genes referred in the official gene name of NR transcripts, including sense genes for antisense transcripts, homologous genes for pseudogenes, coding genes for noncoding transcript variants, as well as read through, divergent, or intronic transcripts. For lincRNA, miRNA host gene, small nuclear RNA, and other lncRNA, the official gene symbol was used for the annotation. This information was manually checked using information available in the nucleotide database. The DAVID program (<https://www.david.ncifcrf.gov>) was used to identify

enriched molecular functions and pathways related to the original genes. Q -values (P -values adjusted for false discovery rate) were calculated using the Benjamini–Hochberg method in DAVID.

Ratio of nonsynonymous (K_a) to synonymous (K_s) nucleotide substitution rates

To identify orthologous regions between human transcripts and chimpanzee/mouse genomes, BLAT v. 36 (Kent 2002) was conducted using human transcript sequences with the estimated PTI score against chimpanzee (PtRV2) and mouse (GRCm38.p6) genomic sequences defined in the NCBI database. We defined the blat best-hit genomic regions of chimpanzee/mouse as orthologs for each human transcript. The human–chimpanzee (or human–mouse) sequences were aligned for each exon region and the sequences were combined for each transcript. Only orthologous sequence pairs more than 60 bp in length (encoding > 20 amino acid sequences) were extracted.

Nonsynonymous (K_a) and synonymous (K_s) nucleotide substitution rates were estimated by using the method described by Yang and Nielsen (Yang and Nielsen 2000) as implemented in PAML version 4.8a (Yang 1997). Note that transcripts with high K_a (> 1) or high K_s (> 1) were excluded from our datasets as outliers. Finally, we estimated both K_a and K_s for 47,228 NM human–chimpanzee, 14,116 NM human–mouse, 8,810 NR human–chimpanzee, and 1,561 NR human–mouse pairs.

Relative frequency of negatively selected genes

We defined the relative frequencies of negatively selected genes $h(x)$ in coding and noncoding transcripts (Figure 1F), as shown in Equation 13:

$$h(x) = \frac{Nns(x)}{TNor(x)} \quad (13)$$

where, $TNor(x)$ represents the total number of coding or noncoding transcripts with orthologous sequences at PTI score = x . $Nns(x)$ is the number of coding or noncoding transcripts with $Ka/Ks < 0.5$ at PTI score = x . The PTI score x is restricted as shown in Equation 7.

Phylogenetic tree

TimeTree (Hedges et al. 2006) was used to draw the phylogenetic tree using the official names of species.

Selection of viruses and identification of vORF

The complete genomes of positive-sense single-strand RNA viruses of human or bacteria (Supplementary table 7 and 8) were collected from the NCBI Virus (Hatcher et al. 2017). Based on the information of described in the Nucleotide database, viral ORFs were identified and sum of vORF lengths $\sum_{i=1}^n l_{vORFi}$ was manually calculated. We eliminated the viruses that translate virus proteins after splicing or by exceptional translation mechanisms such as ribosome frameshifting, alternative initiation sites, ribosome slippage, and RNA editing.

vORF score

Definition

vORF score was calculated according to Equation 14 to 16:

$$\sum_{i=1}^n l_{vORFi} = l_{vORF1} + l_{vORF2} + \dots + l_{vORFk} + \dots + l_{vORFn} \quad (14)$$

$$\sum_{i=1}^n l_{sPTIi} = l_{sPTI1} + l_{sPTI2} + \dots + l_{sPTIk} + \dots + l_{sPTIn} \quad (15)$$

$$vORF \text{ score} = \frac{\sum_{i=1}^n l_{vORFi}}{\sum_{i=1}^n l_{vORFi} + \sum_{i=1}^n l_{sPTIi}} \quad (16)$$

where, l_{vORFi} represents the length of bona-fide ORFs and $\sum_{i=1}^n l_{sPTIi}$ is the sum of the lengths of secondary PTI lengths. $\sum_{i=1}^n l_{vORFi} + \sum_{i=1}^n l_{sPTIi}$ represents the sum of the length of all PTIs including all ORFs.

Statistical analyses

Statistical analyses were performed using Excel and R software (R Project for Statistical Computing, Vienna, Austria).

Data availability

1. Source data for statistical analyses and figures (10 example datasets):

<https://figshare.com/s/498cb340a075284b2dbf>

2. Code associated with generating and analyzing these tables:

<https://figshare.com/s/0f1ed0954d5bd620eb59>

Funding

This work was partially supported by the Yasuda Medical Foundation and Interstellar Initiative (grant number 18jm0610006h0001 to Y.S.) from the Japan Agency for Medical Research and Development.

Acknowledgments

We thank Sana Yokoi, Akira Nakagawara, Yoshitaka Hippo, Yohko Yamaguchi for helpful comments, Isao Kurosaka and Miayamoto Joe Philips for assistance in calculating the PTI scores and Mayuko Seto, Ryo Ohtaka and Manami Shinbori for assistance in checking putative domain structure in BLASTP.

Author contributions

Y.S. conceived and developed the research plan; Y.S., M.K., M.N., K.N., H.K., M.K., and T.M. analyzed the data; and Y.S., M.K., and T.M. wrote the manuscript.

Additional information

Supplementary information is available.

Competing interests

The authors declare no competing financial interests.

References

- Ankarklev J, Jerlström-Hultqvist J, Ringqvist E, Troell K, Svärd SG. 2010. Behind the smile: cell biology and disease mechanisms of Giardia species. *Nat Rev Microbiol.* 8:413–422.
- Bailey TL, Grimmond SM. 2006. The abundance of short proteins in the mammalian proteome. *PLoS Genet.* 2, e52.
- Bartelt LA, Sartor RB. 2015. Advances in understanding Giardia: determinants and mechanisms of chronic sequelae. *F1000Prime Rep.* 7:62.
- Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 33:981–993.
- Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. *Genetics.* 172:1675–1681.
- Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for de novo evolution of testis-expressed genes in the drosophila yakuba/drosophila erecta clade. *Genetics.* 176:1131–1137.
- Brown SA, Warburton D, Brown LY, Yu CY, Roeder ER, Stengel-Rutkowski S, Hennekam RC, Muenke M. 1998. Holoprosencephaly due to mutations in ZIC2, a homologue of Drosophila odd-paired. *Nat Genet.* 20:180–183.
- Buret AG, Cacciò SM, Favennec L, Svärd S. 2020. Update on Giardia: Highlights from the seventh International Giardia and Cryptosporidium Conference. Mise à jour sur Giardia et la giardiase: faits saillants de la Septième Conférence Internationale sur Giardia et Cryptosporidium. *Parasite* 27:49.
- Candeias MM. 2011. The can and can't dos of p53 RNA. *Biochimie* 93:1962–1965.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet.* 14:645–660.

Djebali S, Davis CA, Merkel A, Dobin A, Lassermann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, et al. 2012. Landscape of transcription in human cells. *Nature* 489:101–108.

French SL, Santangelo TJ, Beyer AL, Reeve JN. 2007. Transcription and translation are coupled in Archaea. *Mol Biol Evol.* 24:893–895.

Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kawai J, Carninci P, Hayashizaki Y, Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147:789–802.

Gripp KW, Wotton D, Edwards MC, Roessler E, Ades L, Meinecke P, Richieri-Costa A, Zackai EH, Massagué J, Muenke M, et al. 2000. Mutations in TGIF cause holoprosencephaly and link NODAL signalling to human neural axis determination. *Nat Genet.* 25:205–208.

Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuk Y, Schäffer AA, Brister JR. 2017. Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Res.* 45:D482-D490.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics.* 22: 2971-2972.

Holland PW, Marlétaz F, Maeso I, Dunwell TL, Paps J. 2017. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philos Trans R Soc Lond B Biol Sci.* 372:20150480.

Huang DW, Sherman BT, Lempicki RA. 2009a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 4:44–57.

Huang DW, Sherman BT, Lempicki RA. 2009b. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37:1–13.

Huang JZ, Chen M, Chen D, Gao X-C, Zhu S, Huang H, Hu M, Zhu H, Yan G-R. 2017. A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. *Mol Cell* 68: 171–184.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell.* 147: 789-802.

Ingolia NT. 2014. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet.* 15:205–213.

Islam SM, Suenaga Y, Takatori A, Ueda Y, Kaneko Y, Kawana H, Itami M, Ohira M, Yokoi S, Nakagawara A. 2015. Sendai virus-mediated expression of reprogramming factors promotes plasticity of human neuroblastoma cells. 106:1351–1361.

Kaneko Y, Suenaga Y, Islam SM, Matsumoto D, Nakamura Y, Ohira M, Yokoi S, Nakagawara A. 2015. Functional interplay between MYCN, NCYM and OCT4 promotes aggressiveness of human neuroblastomas. *Cancer Sci.* 106:840–847.

Kato K, Miya F, Hamada N, Negishi Y, Narumi-Kishimoto Y, Ozawa H, Ito H, Hori I, Hattori A, Okamoto N, et al. 2019. *MYCN* de novo gain-of-function mutation in a patient with a novel megalencephaly syndrome. *J Med Genet.* 56:388-395.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res.* 12:656–664.

Kloc M, Foreman V, Reddy SA. 2011. Binary function of mRNA. *Biochimie* 93:1955–1961.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res.* 19:1752–1759.

Kopp F, Mendell JT. 2018. Functional classification and experimental dissection of long noncoding RNAs. *Cell.* 172:393–407.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A.* 103:9935–9939.

Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. 2009. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *Plant J.* 58:485–498.

Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. 2010. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* 20:408–420.

Li C-Y, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li X, Yu Q, Zheng X, et al. 2010. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput Biol.* 6:e1000734.

Li J, Liu C. 2019. Coding or noncoding, the converging concepts of RNAs. *Front Genet.* 10:496.

McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of *de novo* protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci.* 370:20140332.

McLysaght A, Hurst LD. 2016. Open questions in the study of *de novo* genes: what, how and why. *Nat Rev Genet* 17:567–578.

Miller OL Jr, Hamkalo BA, Thomas CA Jr. 1970. Visualization of bacterial genes in action. *Science* 169:392–395.

Monsellier E, Chiti F. 2007. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.* 8:737–742.

Neme R, Tautz D. 2014. Evolution: dynamics of *de novo* gene emergence. *Curr Biol.* 24:R238–R240.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group Phase I & II Team. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature.* 420:563–573.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. *Elife* 3:e03523.

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Albà MM. 2018. Translation of neutrally evolving peptides provides a basis for *de novo* gene evolution. *Nat Ecol Evol.* 2:890–896.

Shoji W, Suenaga Y, Kaneko Y, Islam SM, Alagu J, Yokoi S, Nio M, Nakagawara A. 2015. NCYM promotes calpain-mediated Myc-nick production in human

MYCN-amplified neuroblastoma cells. *Biochem Biophys Res Commun.* 461:501–506.

Suenaga Y, Islam SM, Alagu J, Kaneko Y, Kato M, Tanaka Y, Kawana H, Hossain S, Matsumoto D, Yamamoto M, et al. 2014. *NCYM*, a Cis-antisense gene of *MYCN*, encodes a de novo evolved protein that inhibits GSK3 β resulting in the stabilization of MYCN in human neuroblastomas. *PLoS Genet.* 10:e1003996.

Suenaga Y, Nakatani K, Nakagawara A. 2020. De novo evolved gene product NCYM in the pathogenesis and clinical outcome of human neuroblastomas and other cancers. *Jpn J Clin Oncol.* 50: 839-846.

Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y. 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41:e166.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Albà MM. 2009 Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol.* 26:603–612.

Ulitsky I, Bartel DP. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell.* 154:26–46.

van Bokhoven H, Celli J, van Reeuwijk J, Rinne T, Glaudemans B, van Beusekom E, Rieu P, Newbury-Ecob RA, Chiang C, Brunner HG. 2005. MYCN haploinsufficiency is associated with reduced brain size and intestinal atresias in Feingold syndrome. *Nat Genet.* 37:465–467.

Van Oss SB, Carvunis AR. 2019. De novo gene birth. *PLoS Genet.* 15:e1008160.

Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41:e74.

776
777 Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as
778 predicted by the preadaptation hypothesis of de novo gene birth. *Nat Ecol Evol.*
779 1:0146.
780
781 Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution
782 rates under realistic evolutionary models. *Mol Biol Evol.* 17:32–43.
783
784 Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum
785 likelihood. *Comput Appl Biosci.* 13:555–556
786
787 Zhang YE, Long M. 2014. New genes contribute to genetic and phenotypic
788 novelties in human evolution. *Curr Opin Genet Dev.* 29:90–96.
789

Table 1. Number of original transcripts that produced NR transcripts with high coding frequency ($0.6 \leq \text{PTI score} < 0.8$)

Transcript	Domain		Total	P-value
	With	Without		
Antisense	4	61	65	7.79E-08
LincRNA	3	65	68	7.60E-09
Pseudogene	50	17	67	4.32E-07
Readthrough	7	0	7	6.00E-03
Transcript variant of coding gene	146	35	181	1.05E-19
Divergent	0	2	2	N.S.
Intronic	0	6	6	N.S.
Small nuclear RNA	0	3	3	N.S.
miRNA host gene	0	3	3	N.S.
Other lncRNA	7	118	125	1.12E-13
Total	217	310	527	

P-values were calculated by Yate's continuity correction. N.S., not significant.

Figure legends

Figure 1. Potentially translated island (PTI) score predicts protein-coding potential

of human transcripts. (A) Conceptual explanation of PTIs in an RNA in the three frames and the definition of PTI score. Black and white rectangles indicate primary and secondary PTIs, respectively. The primary PTI is the longest PTI, while secondary PTIs are all others. l is the length of PTIs. (B) Visual image of the PTIs in the RNAs at low, medium (0.5), and high (1) PTI scores. (C) Histogram of relative frequency of PTI scores in coding $f(x)$ and noncoding $g(x)$ transcripts (upper) and in random controls (bottom). (D) Visual explanation of $F(x)$ at the PTI score of 0.15. (E) The PTI score correlates with protein-coding potential, $F(x)$, at PTI scores ≤ 0.65 (upper) and those in random controls (bottom). (F) Relationship between PTI score and $F(x)$ in syntenic human transcript to chimpanzee (upper left) or mouse (bottom left). The relative frequency of transcripts with negative selection $h(x)$ were plotted for each PTI score (upper and bottom right). See the Methods for the definition of $h(x)$. The transcripts are syntenic to the genome of chimpanzee (upper right) or mouse (bottom right). The open circles indicate NR transcripts, and the closed circles indicate NM transcripts. (G) Relationship between PTI scores and percentage of NR transcripts re-registered as NM in the past 3 years. N.D., not detected.

Figure 2. Phylogenetic tree for the cellular organisms. The numbers of species are indicated in each lineage. The lineage of five species including one archaea

(*Nitrososphaera viennensis* EN76), two fungi (*Puccinia graminis* f. sp. *Tritici* and *Pyricularia oryzae*), and two animals (*Strongylocentrotus purpuratus* and *Lingula anatine*) are unknown, and thus, excluded from the figure.

Figure 3. Relationship between PTI score and relative frequencies of coding and noncoding transcripts from bacteria to mammals. Histogram of $f(x)$ (white) or $g(x)$ (black) in observed data (left) and in nucleic-acid-scrambled controls (right). PTI scores with highest $f(x)$ are presented in the histograms.

Figure 4. Relationship between PTI score and relative frequencies of coding $f(x)$ and noncoding transcripts in Primates (A), Glires (B), and Laurasiatheria (C).

Figure 5. Relationship between PTI score and protein-coding potential $F(x)$ in 32 eukaryotes. Phylogenetic tree of the 32 species (left), the dot plots and shape and formula of approximate functions. L, S, and O indicate linear (in red), sigmoidal (in black) and other (in blue) functions. Less than five lncRNAs had a PTI score of 0.05 in *U. americanus*, *C. canadensis*, and *G. gorilla*. Therefore, we eliminated the $F(0.05)$ in these species for linear function approximations (asterisks).

Figure 6. Adaptation of (+) ssRNA viruses to host cells by maximizing protein-coding potential of the RNA genomes. (A) Conceptual explanation of the length of sPTIs and bona-fide viral ORFs in a genome of (+) ssRNA virus and the definition of

viral ORF (vORF) score. Black and white rectangles indicate viral ORFs and secondary PTIs, respectively. l is the length of the ORFs and PTIs. (B) Histograms of relative frequencies of human (+) ssRNA viruses (red) and bacteriophages (black).

Supplementary figure legends

Supplementary Figure 1. PTIs of NCYM and an example of PTI score calculation.

(A) cDNA sequence of *NCYM* transcript. (B) Translated amino acid sequence of *NCYM* in 3 frames in 5' to 3' (sense) direction. Red characters indicate a primary PTI and blue characters indicate secondary PTIs. Stop codons are shown as asterisks. (C) Calculation of PTI score and $F(x)$ in the *NCYM* transcripts. The length of pPTI is 109, and the sum of sPTI lengths is 83. Therefore, the PTI score is 0.568.

Supplementary Figure 2. Relative frequency of coding and noncoding transcripts

in human and mouse PTI scores. (A) Histogram of PTI score relative frequency in coding $f(x)$ and noncoding $g(x)$ human transcripts with random shuffling controls using human data sets from RefSeq. (B) Relative frequency of coding $f(x)$ and noncoding $g(x)$ transcripts calculated using human data sets from Ensembl. (B) Relative frequency of coding $f(x)$ and noncoding $g(x)$ transcripts calculated using mouse data sets from RefSeq (upper panels) or Ensembl (lower panels).

Supplementary Figure 3. PTI score correlates with protein-coding potential, $F(x)$,

at PTI scores ≤ 0.65 in human and mouse transcripts. (A) Relationship between

PTI score and $F(x)$ in a human data set from Ensemble and random controls (center).

Random shuffling controls (right) were generated from a human data set from both

Ensemble and RefSeq. (B) Relationship between PTI score and $F(x)$ in mouse

transcripts using data sets from RefSeq (upper panels) or Ensembl (lower panels).

Supplementary Figure 4. Relationship between PTI score and relative frequencies

of coding and noncoding transcripts in archaea. Phylogenetic tree for 9 archaeal

species and histogram of $f(x)$ (white) or $g(x)$ (black) in the data (left) and in nucleic-

acid-scrambled controls (right). PTI scores with highest $f(x)$ are indicated in the

histograms. The lineage of one archaea (*Nitrososphaera viennensis* EN76) is unknown

and thus excluded from the phylogenetic tree.

Supplementary Figure 5. Relationship between PTI score and relative frequencies

of coding and noncoding transcripts in plants. Phylogenetic tree for 12 plants and

histogram of $f(x)$ (white) or $g(x)$ (black) in the data (left) and in nucleic-acid-scrambled

controls (right). PTI scores with highest $f(x)$ are indicated.

Supplementary Figure 6. Relationship between PTI score and protein-coding

potential $F(x)$ in 32 eukaryotes. Relationship between PTI score and $F(x)$ in data sets

from Ensemble (left, used in Figure 5) and those in random controls (right). Mouse and

human data are identical to Supplementary Figure 3. Shape of approximate functions

are shown as L, S, or O, indicating linear (in red), sigmoidal (in black), and other (in

blue) functions, respectively. The number of lncRNAs at PTI score 0.05 were < 5 in *U. americanus*, *C. canadensis*, and *G. gorilla*. Therefore, we eliminated the $F(0.05)$ in these species for the approximation by linear functions (asterisks).

Supplementary Table 1. Cellular organisms with the official name, taxonomy ID, lineage information, and number of coding or noncoding transcripts.

Supplementary Table 2. Twenty-three human noncoding transcripts showing both negative selection ($Ka/Ks < 0.5$) and the highest PTI score.

Supplementary Table 3. Human NR transcripts reassigned as NM over the past three years.

Supplementary Table 4. Human NR transcripts with high protein-coding potential ($0.6 \leq \text{PTI score} < 0.8$).

Supplementary Table 5. Functional annotation of human NR transcripts with high protein-coding potential and without putative domain structure(s).

Supplementary Table 6. Functional annotation of human NR transcripts with high protein-coding potential and with putative domain structure(s).

905 **Supplementary Table 7. Positive-sense single-stranded human viruses with official**

906 **name, taxonomy ID, lineage information, as well as genome length and sequence.**

907

908 **Supplementary Table 8. Positive-sense single-stranded bacteriophages with official**

909 **name, taxonomy ID, lineage information, source information, and genome length**

910 **and sequence.**

Figure 1

Figure 2

Figure 3

Figure 5

Figure 6