Potentially translated sequences determine protein-coding potential of RNAs in cellular organisms

Yusuke Suenaga[1,‡]*, Mamoru Kato[2,‡], Momoko Nagai[2], Kazuma Nakatani[1,3,4], Hiroyuki Kogashi[1], Miho Kobatake[1], Takashi Makino[5]

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

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

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

[4]Innovative Medicine CHIBA Doctoral WISE Program

[5]Laboratory 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

Abstract

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

44 sequence characteristics in eukaryotes, instead they exhibit a gradient of protein-coding
45 potential.
46
47 **Introduction**
48 Recent advances in sequencing technology have revealed that most of the eukaryotic
49 genome is transcribed, resulting in production of noncoding RNAs (Okazaki et al. 2002;
50 Djebali et al. 2012; Ulitsky and Bartel 2013; Kopp et al. 2018). Noncoding RNAs that
51 are more than 200 nucleotides in length are referred to as long noncoding RNAs
52 (lncRNAs) and are not translated into proteins (Ulitsky and Bartel 2013; Kopp et al.
53 2018). LncRNAs have been reported to function in various biological phenomena, such
54 as the regulation of transcription, modulation of protein or RNA functions, and
55 organization of the nuclear architecture (Ulitsky and Bartel 2013; Kopp et al. 2018).
56 However, paradoxical to its definition, a large fraction of lncRNAs are associated with
57 ribosomes and translated into peptides (Frith et al. 2006; Ingolia et al. 2011; Bazzini et
58 al. 2014; Ingolia et al. 2014; Ruiz-Orera et al. 2014). Peptides translated from
59 transcripts annotated as lncRNAs have been shown to have biological functions in
60 various eukaryotes (Li and Liu 2019). Moreover, known protein-coding genes, such as
61 *TP53*, act as functional RNAs (Candeias 2011; Kloc et al. 2011). Detection of these
62 RNAs with binary functions has blurred the distinction between coding and noncoding
63 RNAs, and the characteristics of RNA sequences that explain the continuity between
64 noncoding and coding transcripts remain unclear.

65 During evolution, new genes originated from pre-existing genes via gene duplication or
66 from non-genic regions via the generation of new open reading frames (ORFs) (Ohno
67 1970; Chen et al. 2013; Zhang and Long 2014; McLysaght and Guerzoni 2015;
68 MyLysaght and Hurst 2016; Holland et al. 2017). Newly evolved genes from non-genic
69 regions are known as *de novo* genes (Begun et al. 2006; Levine et al. 2006; Begun et al.
70 2007; Knowles and McLysagtht 2009; Li et al. 2009; Toll-Riera et al. 2009; Li et al.
71 2010), and regulate phenotypes and diseases (McLysaght and Guerzoni 2015; Chen et
72 al. 2013; Zhang and Long 2014), such as brain function and carcinogenesis in humans
73 (Li et al. 2010; Suenaga et al. 2014). LncRNAs serve as sources of newly evolving gene
74 products (Ruiz-Orera et al. 2014), including *de novo* evolved proteins. Not only ORFs
75 exposed to natural selection, neutrally evolving ORFs are also translated from lncRNAs
76 to stably express peptides (Ruiz-Orera et al. 2018), providing a foundation for the
77 development of new functional peptides/proteins. High levels of lncRNA expression
78 (Ruiz-Orera et al. 2018), hexamer frequencies of ORFs (Sun et al. 2013; Wang et al.
79 2013; Ruiz-Orera et al. 2014), and high flexibility of peptides (Wilson et al. 2017) have
80 been proposed as determinants of coding potential; however, the molecular mechanisms
81 of how lncRNAs evolve into new coding transcripts remain unclear (Van Oss and
82 Carvunis 2019).
83 This study, therefore, sought to identify a new indicator to determine the protein-coding
84 potential of RNAs. Within an RNA sequence, we termed sequence segments that start
85 with AUG (the base sequence of the start codon) and end with UAG, UGA, or UAA
86 (base sequences of stop codons) as potentially translated islands (PTIs). First, we

87 defined this indicator using PTI lengths and subsequently examined associations
88 between the indicator and protein-coding potential. We also present analyses of more
89 than 3.4 million transcripts in 100 cellular organisms belonging to all three domains of
90 life to investigate the evolution of the relationship between the PTI score and protein-
91 coding potential. Finally, we examined whether virus RNA genomes have differentially
92 evolved to maximize the protein-coding potential in different host organisms.
93

94 **Results**

95 **Coding transcripts show higher PTI scores in humans and mice**

96 We defined PTI score as described in the Materials and Methods and illustrated in
97 Figure 1A to B, and analyzed human transcripts registered in the nucleotide database of
98 the National Centre for Biotechnology Information for coding (RefSeq ID starting with
99 NM) and noncoding (RefSeq ID starting with NR) RNA transcripts. Sample PTI score
100 calculation is presented in Supplementary Figure 1. The nucleic acid sequences and IDs
101 were downloaded from Table browser (https://genome.ucsc.edu/cgi-bin/hgTables) after
102 setting the track tab as "RefSeq Genes." A total of 50,052 coding (NM) and 13,550
103 noncoding (NR) RNAs were registered in 2018 (Supplementary Table 1). To analyze
104 putative lncRNAs with protein-coding potential, we excluded small RNAs (shorter than
105 200 bp) or RNAs with a short primary PTI (pPTI; shorter than 20 amino acids) from the
106 NR transcripts, focusing on the remaining 12,827 transcripts.

107 We analyzed relative frequencies of NM and NR transcripts, designated as $f(x)$ and $g(x)$
108 respectively (Figure 1C), where x indicates PTI score. See Materials and Methods for
109 the definition of $f(x)$ and $g(x)$.
110 In human transcripts, $g(x)$ showed a distribution that shifted to the left with an apex of
111 0.15; in contrast, distribution of $f(x)$ shifted to the right with an apex of 0.55 (Figure 1C,
112 upper panel). As a control, we used randomly generated nucleic acid sequences where
113 A/T/G/C bases were randomly assigned with equal probabilities. In the controls, the
114 relative frequencies of PTI scores were shifted to the left in both coding and noncoding
115 transcripts (Figure 1C, bottom panel). The controls that randomly shuffled the original
116 sequence without affecting the number of A/T/C/G bases in each transcript also had
117 relative frequencies of PTI scores shifted to the left in both coding and noncoding
118 transcripts (Supplementary Figure 2A). Moreover, similar results were obtained using a
119 data set from the Ensembl database (Supplementary Figure 2B). We also calculated the
120 PTI scores of mouse transcripts registered in RefSeq and Ensembl and found that the
121 distribution of $f(x)$ was shifted to the right with an apex of 0.55 (Supplementary Figure
122 2C), similar to that of human transcripts. These results suggest that the sequences, not
123 lengths, of the coding transcripts increase PTI scores in mice and humans.
124

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

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

129 indicates that any given human RNA transcript with a calculated PTI score of 0.15 has a
130 protein-coding potential $F(x)$ of 0.183. $F(x)$ was correlated with PTI score at PTI scores
131 ≤ 0.65 (Figure 1E and Supplementary Figure 3A). The protein-coding potential of
132 sequences in RefSeq data slightly decreased after peaking at 0.65 (Figure 1E) whereas
133 that of sequences in the Ensembl data remained high (Supplementary Figure 3A). $F(x)$
134 of human transcripts was well approximated by the linear regression as follows:
135
136 Based on Ensembl data
137 $F(x) = 1.301x + 0.0072$ ($x \leq 0.65$), $R^2 = 0.984$
138
139 Based on RefSeq data
140 $F(x) = 1.313x + 0.0189$ ($x \leq 0.65$), $R^2 = 0.990$
141
142 The intercepts were near zero, and slope was approximately 1.3. By using these
143 formulas, we can calculate the protein-coding potential $F(x)$ of any given human
144 transcript with a PTI score ≤ 0.65. For example, the protein-coding potential $F(x)$ of
145 *NCYM* was calculated as 0.746 or 0.765 based on Ensembl or RefSeq databases,
146 respectively (Supplementary Figure 1C). In contrast, $F(x)$ in the controls had no positive
147 correlation with PTI scores (Figure 1E, bottom panel, and Supplementary Figure 3A).
148 Similar results were obtained for mouse transcripts (Supplementary Figures 3B). $F(x)$ of
149 mouse transcripts (PTI score ≤ 0.65) was approximated as follows.
150

151 Based on Ensembl data

152 $F(x) = 1.142x + 0.067$, $R^2 = 0.982$

153

154 Based on RefSeq data

155 $F(x) = 1.482x - 0.061$, $R^2 = 0.990$

156

157 In both human and mouse transcripts, the PTI score correlated linearly with the protein-

158 coding potential at PTI scores ≤ 0.65. Moreover, when the PTI score limit approached

159 0, the probability of the transcript being a coding RNA was 0.

160

161 **PTIs affect the protein-coding potential predicted by *K*a/*K*s**

162 To examine the relationship between the PTI score and natural selection in the

163 prediction of protein-coding potential, we calculated the ratio of nonsynonymous (*K*a)

164 to synonymous (*K*s) values by comparing human transcripts with syntenic genome

165 regions of chimpanzee or mouse (Figure 1F). Transcripts were selected based on the

166 syntenically conserved regions, 44,593 (vs. chimp) and 14,016 (vs. mouse). The results

167 revealed linear relationships between the $F(x)$ and PTI score in the conserved transcripts

168 (Figure 1F, left panels). As predicted, coding transcripts contained transcripts with

169 *K*a/*K*s < 0.5 at a higher frequency than noncoding transcripts, with the largest difference

170 observed when the PTI score was > 0.9, and the smallest difference at PTI scores of

171 approximately 0.45 (Figure 1F, right panels).

172 Therefore, noncoding transcripts showing both negative selection (*K*a/*K*s < 0.5) and the
173 highest PTI score may include new coding transcript candidates.
174 We have listed 23 such transcripts (Supplementary Table 2), which include four
175 transcript variants of a previously identified lncRNA that encodes a tumor suppressive
176 small peptide, HOXB-AS3 (Huang et al 2017).
177
178 **Characterization of high PTI score lncRNAs in humans**
179 Next, we investigated whether the PTI score is useful for identifying coding RNAs
180 among NR transcripts. From the 7,144 transcripts registered as noncoding genes until
181 2015, we excluded small RNAs (< 200 nucleotides) and those with short primary PTIs
182 (< 20 amino acids). Among the remaining 6,617 NR genes, 219 were reassigned as NM
183 over the past 3 years (Supplementary Table 3), including the previously identified *de*
184 *novo* gene *MYCNOS/NCYM* (Suenaga et al. 2014). The percentage of reclassification
185 increased among NR transcripts with high PTI scores (Figure 1G). Thus, a high PTI
186 score is a useful indicator of coding transcripts. NR transcripts with high protein-coding
187 potential (0.6 ≤ PTI score < 0.8) were then extracted, and the domain structure of the
188 pPTI amino acid sequence was estimated using BLASTP (See the Materials and
189 Methods for details). A total of 217 transcripts showed a putative domain structure(s) in
190 pPTI, whereas 310 transcripts showed none (Supplementary Table 4). Transcripts with a
191 domain structure are often derived from transcript variants, pseudogenes, or
192 readthrough of coding genes; transcripts without domain structures are often derived
193 from antisense RNA or long intergenic noncoding RNA (lincRNA) (Table 1).

194 We next examined the functions of genes generating NR transcripts with high coding
195 probabilities (0.6 ≤ PTI score < 0.8). We divided the NR transcripts into those with or
196 without putative domains to investigate candidates of novel coding genes, either
197 originating from pre-existing genes or created from non-genic regions, respectively.
198 Analysis using the Database for Annotation, Visualization, and Integrated Discovery
199 (DAVID) Functional Annotation Tool (Huang et al. 2009a, 2009b) showed that NR
200 transcripts without domain structures are derived from original genes related to
201 transcriptional regulation, multicellular organismal process, and developmental process
202 (Supplementary Table 5; see Materials and Methods for the definition of original
203 genes). Among the target genes of transcription factors, NMYC, TGIF, and ZIC2 were
204 ranked in the top three and are all necessary for forebrain development (Supplementary
205 Table 5) (Brown et al. 1998; Gripp et al. 2000; van Bokhoven et al. 2005). We observed
206 that NR transcripts with domain structures originated from genes that undergo
207 alternative splicing which are related to organelle function and are expressed in various
208 cancers, including respiratory tract tumors, gastrointestinal tumors, retinoblastomas, and
209 medulloblastomas (Supplementary Table 6).
210
211 **Relationship between PTI score and relative frequencies of coding/noncoding**
212 **transcripts in 100 cellular organisms**
213 To analyze the relationship between PTI scores and protein-coding potential in a broad
214 lineage of cellular organisms, we selected 100 organisms, consisting of 5 bacteria, 10
215 archaea and 85 eukaryotes (Supplementary Table 1), and calculated the PTI score for a

216 total of more than 3.4 million transcripts (Supplementary Table 1). Phylogenic trees of
217 the cellular organisms are presented on a logarithmic time scale along with the number
218 of species in each lineage used in the analyses (Figure 2). To examine the evolutionary
219 conservation of the linear relationship between PTI score and protein-coding potential
220 in humans and mice, we selected a relatively large number of species (36) from
221 mammals. The species with fewer than three lncRNAs were not used to calculate $g(x)$
222 and were not included in the histograms illustrating the relationship with PTI score
223 (Figure 3). In all organisms, the relative frequency of coding transcripts $f(x)$ was shifted
224 to the right (higher PTI score) compared to random or random shuffling controls (Figure
225 3, Figure 4, Supplementary Figure 4 and Supplementary Figure 5).
226 In bacteria and archaea, $f(x)$ and $g(x)$ exclusively presented high and low PTI scores,
227 respectively, indicating a clear boundary between the coding transcripts and lncRNAs in
228 terms of PTI scores (Figure 3 and Supplementary Figure 4). In addition, the highest
229 frequency of coding transcript $f(x)$ presenting a PTI score was 0.75 in all examined
230 bacteria (Figure 3) and ≥ 0.75 in archaea (Supplementary Figure 4). Among eukaryotes,
231 unicellular organisms and non-vertebrates showed the highest frequency of coding
232 transcripts at 0.65 or 0.75 (Figure 3), while most vertebrates showed highest values $\leq$
233 0.65 (Figure 3 and 4). In addition, the $f(x)$ distribution in vertebrates was broad and
234 shifted to the left (lower PTI scores) compared to bacteria and archaea (Figure 3 and 4).
235 In sharp contrast to $f(x)$, the relative frequency of lncRNAs $g(x)$ was shifted to the right
236 (higher PTI scores) in eukaryotes, including *G. lamblia* that belongs to the earliest
237 diverging eukaryotic lineage and lacks mitochondria (Figure 3). Since the distribution

238 of $f(x)$ in the Excavata, including *G. lamblia,* showed a similar pattern to bacteria, the
239 right shift of $g(x)$ seems to be an earlier event than the left shift of $f(x)$ in the evolution
240 of eukaryotes. Collectively, the right and left shifts of $f(x)$ and $g(x)$ contribute to
241 blurring the boundary between coding and noncoding transcripts in eukaryotes.
242
243 **Relationship between the PTI score and protein-coding potential**
244 The overlapping of relative frequencies in $f(x)$ and $g(x)$ led us to examine the
245 relationship between PTI score and protein-coding potential $F(x)$ in eukaryotes. To
246 avoid misleading data obtained by small sampling numbers, we selected 32 species with
247 more than 1000 lncRNAs that contain pPTIs for calculation of $F(x)$ (Figure 5 and
248 Supplementary Figure 6). In humans and mice, the relationship between PTI score and
249 $F(x)$ was approximated with the linear function passing through the origin of the PTI
250 score ≤ 0.65. Therefore, we used linear approximation in the $F(x)$ of 32 species and
251 divided them into two groups, linear and sigmoidal, based on the shape and formula of
252 the approximated function (Figure 5). We defined the linear group as $R^2 > 0.9$ with an
253 absolute value of the intercept < 0.1; the sigmoidal group had a slope > 1.0 and intercept
254 < -0.1. In *U. americanus*, *C. canadensis*, and *G. gorilla*, fewer than five lncRNAs
255 exhibited PTI scores of 0.05; thus, we eliminated the $F(0.05)$ in these species for the
256 approximation by linear function (indicated with asterisks in Figure 5). The five species
257 that did not fit in the linear or sigmoidal group were characterized by high $F(x)$ in low
258 PTI scores, and belonged to plants (*Z. mays*), reptiles (*A. Carolinensis*) and mammals
259 (*O. anatinu*, *S. boliviensis*, and *G. gorilla*) (Figure 5). In these species, PTI scores

260 showed weaker association with protein-coding potential. Sigmoidal relationships were
261 observed in 18/32 species, while linear relationships were apparent in nine species
262 within three mammalian lineages, Cetartiodactyla, Rodentia and Primates. Since the
263 sigmoidal relationship was the most broadly observed across the examined lineages, this
264 relationship appears to be of an ancestral type.
265
266 **Characteristics of RNA virus genomes in human and bacterial cells**
267 In sharp contrast to the coding transcripts of bacteria and archaea, the PTI score in
268 coding transcripts of eukaryotes exhibited overlap with that of noncoding RNA by its
269 broad distribution of low PTI scores. To investigate the molecular mechanism
270 underlying the distinct distribution of coding transcripts between bacteria and
271 eukaryotes, we analyzed genome sequences of RNA viruses that infect human or
272 bacterial cells. Positive-sense single-strand RNAs, (+) ssRNAs, are parts of virus
273 genomes that generate mRNAs and are translated into viral proteins via the host
274 translation system. Therefore, efficient translation in host cells contributes to the
275 replication of the (+) ssRNA viruses. We speculated that PTIs, other than bona-fide
276 ORFs, affect the coding potential of the viral genome in host cells. In a viral genome,
277 multiple bona-fide ORFs exist. We thus extended the concept of PTIs (Figure 6A) to the
278 multiple ORFs in a viral RNA genome and set the viral ORF (vORF) score (see
279 Materials and Methods for the definition of the vORF score).
280 Among the positive-sense ssRNA viruses registered in the NCBI database, 198 were
281 human viruses and 13 were bacteriophages. We eliminated the viruses that translate

282 virus proteins by exceptional translation mechanisms such as ribosome frameshifting,
283 alternative initiation sites, ribosome slippage, and RNA editing, and focused on the
284 remaining 95 human viruses including 9 retroviruses (Supplementary Table 7) and 10
285 bacteriophages (Supplementary Table 8). The relative frequency of the human viruses
286 and bacteriophages showed distinct peaks at PTI scores of 0.65 and 0.75, respectively
287 (Figure 6B). These values correspond to the PTI scores of the highest protein-coding
288 potential in humans (Figure 1E and Supplementary Figure 3A) and the highest
289 frequency of coding transcripts in bacteria (Figure 3), respectively. In addition, the
290 relative frequency of human viruses showed broader distribution of low PTI scores
291 compared to bacteriophages, particularly in human retroviruses (Figure 6B). Therefore,
292 the RNA genomes of viruses appear to have sequence characteristics that maximize
293 protein-coding potential in host cells.
294

295 **Discussion**

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

304 peroxisomes, and a typical Golgi apparatus (Ankarklev et al. 2010; Bartelt et al. 2015;
305 Buret et al. 2020), and were commonly observed in all examined eukaryotes. The most
306 prominent difference between eukaryotes and bacteria/archaea is the existence of nuclei
307 in cells, as this is the definition of eukaryotes. Moreover, functional noncoding RNAs,
308 by definition, should not be translated by ribosomes in cells. However, in bacteria and
309 archaea, newly transcribed RNAs are immediately bound by ribosomes (Miller et al.
310 1970; French et al. 2007) and do not have the chance to escape from translation.
311 Alternatively, in eukaryotes, the existence of nuclei prevents the immediate binding of
312 lncRNAs by ribosomes and cytosolic translocation is required for translation. Therefore,
313 the eukaryotic lncRNAs have a chance to function in the nucleus regardless of PTI
314 scores or translation efficiencies. This nuclear localization likely promotes
315 diversification of RNA sequences, resulting in the right shift in the distribution of
316 noncoding RNAs. Thus, the pervasive transcription of the genome seems to be
317 advantageous for eukaryotes to create new functional nuclear RNAs while being
318 disadvantageous for bacteria/archaea by increasing the risk of transcription of a high
319 PTI score transcript, leading to immediate translation of wasteful and/or toxic proteins
320 (Monsellier et al. 2007).
321 The left shifts observed in distribution of coding RNAs were not observed in *G. lamblia*
322 with its peak at 0.75, however, the majority of fungi, plants and animals, including
323 humans, exhibited this shift with peaks at PTI scores ≤ 0.65. Thus, the left shift of
324 coding transcripts likely occurred after the right shift of noncoding transcripts in
325 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,

348 the PTI score of human *TP53* (NM_000546.5) is 0.476, and the $F(x)$ calculated by the
349 linear function is 0.644. The PTI score of *TP53* indicates that the transcript has a high
350 probability of functioning as both a coding and noncoding RNA, explaining the
351 bifunction of *TP53* (Candeias 2011; Kloc et al., 2011). Further studies are required to
352 explain the molecular mechanisms underlying the existence of these three relationships
353 in eukaryotes.
354 The PTI score is a value uniquely calculated based on any given RNA sequence,
355 whereas the *K*a/*K*s value requires comparison with a species and varies across different
356 species. Therefore, the *K*a/*K*s value cannot be uniquely calculated from a given RNA
357 sequence. Hence, PTI scores, but not *K*a/*K*s values, exhibit species-specific
358 relationships with unique distributions of coding potential. Due to the advantages of the
359 PTI score, we can investigate evolutionary changes in the relationship between PTI
360 score and coding potential. Furthermore, *K*a/*K*s < 0.5 did not show different relative
361 frequencies of noncoding and coding transcripts at a PTI score of approximately 0.45,
362 suggesting that *K*a/*K*s does not predict the coding potential at the PTI score and that the
363 prediction is dependent on the PTI score. In addition, to calculate *K*a/*K*s, orthologous
364 transcripts or genomes should be conserved among species. Therefore, *K*a/*K*s cannot be
365 applied for coding prediction of species-specific transcripts such as *NCYM*. On the other
366 hand, by using approximate functions, we can calculate the coding potential $F(x)$ of any
367 given transcript with a PTI score ≤ 0.65 in the nine species classified to linear group.
368 Human NR transcripts with high PTI scores have been reclassified as coding genes over
369 the past 3 years, including the human *de novo* gene *NCYM* (Suenaga et al. 2014;

370 Suenaga et al. 2020). Because *de novo* gene products have no known domain structures,
371 high PTI score-NR transcripts without putative domains, may be good candidates as
372 novel *de novo* genes in eukaryotes. *NCYM* is an antisense gene of *MYCN* whose protein
373 product stabilizes MYCN (Suenaga et al. 2014). MYCN directly stimulates NCYM and
374 OCT4 transcription, whereas OCT4 induces MYCN (Kaneko et al. 2015). This
375 functional interplay forms a positive feedback loop for these genes to induce each other
376 in human neuroblastomas (Suenaga et al. 2014; Islam et al. 2015; Kaneko et al. 2015;
377 Shoji et al. 2015). Functional annotation of NR genes without putative domains was
378 related to transcriptional regulation, and the target genes of transcription factors,
379 including MYCN, TGIF, and ZIC2, were enriched. As *de novo* emergence of *NCYM*
380 occurred in homininae, NCYM-mediated MYCN activation may modulate human *de*
381 *novo* gene births during evolution, regulating the transcription of *MYCN* target genes.
382 Both TGIF and ZIC2 are causative genes of holoprosencephaly, a disorder caused by a
383 failure in embryonic forebrain development (Brown et al. 1998; Gripp et al. 2000),
384 whereas *MYCN* is the causative gene of Feingold syndrome and megalencephaly
385 syndrome, which are associated with a reduced and increased brain size, respectively
386 (van Bokhoven et al. 2005; Kato et al. 2019). Thus, the present study also provides a list
387 of candidate human *de novo* genes possibly involved in brain development and brain-
388 related diseases.
389 In conclusion, we identified a novel determinant of protein-coding potential in cellular
390 organisms. The relationship between PTI score and protein-coding potential revealed
391 that the boundary between coding and noncoding transcripts in bacteria and archaea is

392 blurred in eukaryotes. Therefore, when a eukaryotic transcript possesses a moderate PTI
393 score, bifunctional characterization as both a coding and noncoding transcript may be
394 essential for a full understanding of the biological roles of RNAs.
395

396 **Materials and Methods**

397 **Potentially translated islands**

398 ***Definition***

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

402 ***Example***

403 As an example, the PTIs in the human *de novo* gene *NCYM* (Suenaga et al. 2014) are
404 identified using the cDNA sequence (Supplementary Figure 1A) and are shown as bold
405 characters (Supplementary Figure 1B). Further information is included in
406 Supplementary Notes.
407

408 **The length of PTI and primary/secondary PTIs**

409 ***Definition***

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

414 ***Example***

415 Based on these definitions, the shortest PTI is "AUGUAA," "AUGUAG," or

416 "AUGUGA," and its amino acid length is 1. For example, the NCYM transcript has a

417 pPTI with a length of 109 at frame 1; three sPTIs with lengths of 69, 8 and 6,

418 respectively, at frame 2; and no PTIs at frame 3 (Supplementary Figure 1B and C).

419 ***Characteristics***

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

421

422 $$1 \leq l_{\mathrm{sPTI}} \leq l_{\mathrm{pPTI}} \quad (1)$$

423

424 **PTI score**

425 ***Definition***

426 The definition of PTI score is motivated by our hypothetical concept that translation of

427 pPTI is limited by alternate competing sPTIs. We defined the PTI score (Figure 1A)

428 according to Equation 2 and 3:

429 $$\sum_{i=1}^{n} l_{\mathrm{sPTI}i} = l_{\mathrm{sPTI1}} + l_{\mathrm{sPTI2}} + \cdots + l_{\mathrm{sPTI}k} + \cdots + l_{\mathrm{sPTI}n} \quad (2)$$

430 $$\text{PTI score} = \frac{l_{\mathrm{pPTI}}}{l_{\mathrm{pPTI}} + \sum_{i=1}^{n} l_{\mathrm{sPTI}i}} \quad (3)$$

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

432 ***Example***

433 If an RNA sequence has only one PTI, the PTI score is 1 (Figure 1B). An RNA

434 sequence with many sPTIs tends to have a score close to 0 (Figure 1B). If the sum of all

435 sPTI length is equal to pPTI length, the PTI score is 0.5 (Figure 1B). The PTI score of

436 the NCYM transcript is calculated as 0.568 (Supplementary Figure 1C). Further

437 information is included in Supplementary Notes.

438 ***Characteristics***

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

440 $$0 < \text{PTI score} \leq 1 \quad (4)$$

441

442 **Relative frequencies *f*(*x*) and *g*(*x*)**

443 ***Definition***

444 We defined the relative frequencies of coding and noncoding transcripts, *f*(*x*) and *g*(*x*),

445 as follows (Figure 1C):

446 $$f(x) = \frac{NM(x)}{TNM} \quad (5)$$

447 $$g(x) = \frac{NR(x)}{TNR} \quad (6)$$

448 where, TNM and TNR represent the total number of coding and noncoding

449 transcripts, respectively, excluding the transcripts lacking PTIs. *NM*(*x*) or *NR*(*x*) is the

450 number of coding or noncoding transcripts having PTI score = *x*, respectively.

451 To define coding/non-coding transcripts with a PTI score of *x*, we made histograms

452 divided by ten classes, and the median values of the classes were used to represent the

453 PTI score (Figure 1C). Therefore, in Equation 5 and 6, the PTI score *x* is restricted as

454 follows:

455 $$x = 0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, \text{or } 0.95 \quad (7)$$

456 ***Characteristics***

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

458

459 $$0 \leq f(x) \leq 1 \quad (8)$$

460 $$0 \leq g(x) \leq 1 \quad (9)$$

461 $$\sum_x f(x) = 1 \quad (10)$$

462 $$\sum_x g(x) = 1 \quad (11)$$

463

464 **Protein-coding potential *F*(*x*)**

465 ***Definition***

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

467 $$F(x) = \frac{f(\mathrm{x})}{f(\mathrm{x})+g(\mathrm{x})} \quad (12)$$

468 ***Example***

469 As an example, $F(0.15)$ in human transcripts is depicted in Figure 1D. $F(0.15)$ is

470 calculated using Equation (12):

471 $$f(0.15) = 0.060$$

472 $$g(0.15) = 0.268$$

473 $$F(0.15) = \frac{f(0.15)}{f(0.15)+g(0.15)} = \frac{0.060}{0.060+0.268} = 0.18292 = 0.183$$

474

475 **Identification of human noncoding transcripts with high protein-coding potential**

476 To identify the noncoding transcripts with high protein-coding potential, NR transcripts

477 with high $F(x)$ ($0.6 \leq x < 0.8$) were identified from total NR transcripts registered in

478 the nucleotide database of the National Centre for Biotechnology Information. NR
479 transcripts shorter than 200 nucleotides or with pPTI encoding putative peptides with
480 fewer than 20 amino acids were excluded. The amino acid sequences of pPTIs in these
481 transcripts were subjected to a BLASTP search to examine the presence of putative
482 domain structures. In the BLASTP search, non-redundant protein sequences (nr) were
483 applied as the search set, and Quick BLASTP (accelerated protein-protein BLAST) was
484 used as the algorithm. In the search results, putative conserved domains or the message
485 "No putative conserved domains have been detected" were shown in the Graphical
486 Summary tab. CDSEARCH/cdd was the database employed to search conserved
487 domain structures using the default settings; low complexity filter: no; composition
488 based adjustment: yes; E-value threshold: 0.01; maximum number of hits: 500. Based
489 on the results, transcripts with or without putative conserved domain structure were
490 indicated as + or -, respectively.
491
492 **Functional annotation of original genes**
493 Original genes are defined as the genes referred in the official gene name of NR
494 transcripts, including sense genes for antisense transcripts, homologous genes for
495 pseudogenes, coding genes for noncoding transcript variants, as well as read through,
496 divergent, or intronic transcripts. For lincRNA, miRNA host gene, small nuclear RNA,
497 and other lncRNA, the official gene symbol was used for the annotation. This
498 information was manually checked using information available in the nucleotide
499 database. The DAVID program (https://www.david.ncifcrf.gov) was used to identify

500 enriched molecular functions and pathways related to the original genes. *Q*-values (*P*-
501 values adjusted for false discovery rate) were calculated using the Benjamini–Hochberg
502 method in DAVID.
503
504 **Ratio of nonsynonymous (*K*a) to synonymous (*K*s) nucleotide substitution rates**
505 To identify orthologous regions between human transcripts and chimpanzee/mouse
506 genomes, BLAT v. 36 (Kent 2002) was conducted using human transcript sequences
507 with the estimated PTI score against chimpanzee (PtRV2) and mouse (GRCm38.p6)
508 genomic sequences defined in the NCBI database. We defined the blat best-hit genomic
509 regions of chimpanzee/mouse as orthologs for each human transcript. The human–
510 chimpanzee (or human–mouse) sequences were aligned for each exon region and the
511 sequences were combined for each transcript. Only orthologous sequence pairs more
512 than 60 bp in length (encoding > 20 amino acid sequences) were extracted.
513 Nonsynonymous (*K*a) and synonymous (*K*s) nucleotide substitution rates were
514 estimated by using the method described by Yang and Nielsen (Yang and Nielsen 2000)
515 as implemented in PAML version 4.8a (Yang 1997). Note that transcripts with high *K*a
516 (> 1) or high *K*s (> 1) were excluded from our datasets as outliers. Finally, we estimated
517 both *K*a and *K*s for 47,228 NM human–chimpanzee, 14,116 NM human–mouse, 8,810
518 NR human–chimpanzee, and 1,561 NR human–mouse pairs.
519
520 **Relative frequency of negatively selected genes**

521 We defined the relative frequencies of negatively selected genes $h(x)$ in coding and

522 noncoding transcripts (Figure 1F), as shown in Equation 13:

523 $$h(x) = \frac{Nns(x)}{TNor(x)} \quad (13)$$

524 where, $TNor(x)$ represents the total number of coding or noncoding transcripts with

525 orthologous sequences at PTI score = x. $Nns(x)$ is the number of coding or noncoding

526 transcripts with $Ka/Ks < 0.5$ at PTI score = x. The PTI score x is restricted as shown in

527 Equation 7.

528

529 **Phylogenic tree**

530 TimeTree (Hedges et al. 2006) was used to draw the phylogenic tree using the official

531 names of species.

532

533 **Selection of viruses and identification of vORF**

534 The complete genomes of positive-sense single-strand RNA viruses of human or

535 bacteria (Supplementary table 7 and 8) were collected from the NCBI Virus (Hatcher et

536 al. 2017). Based on the information of described in the Nucleotide database, viral ORFs

537 were identified and sum of vORF lengths $\sum_{i=1}^{n} l_{\mathrm{vORF}i}$ was manually calculated. We

538 eliminated the viruses that translate virus proteins after splicing or by exceptional

539 translation mechanisms such as ribosome frameshifting, alternative initiation sites,

540 ribosome slippage, and RNA editing.

541

vORF score

Definition

vORF score was calculated according to Equation 14 to 16:

$$\sum_{i=1}^{n} l_{\mathrm{vORF}i} = l_{\mathrm{vORF1}} + l_{\mathrm{vORF2}} + \cdots + l_{\mathrm{vORF}k} + \cdots + l_{\mathrm{vORF}n} \quad (14)$$

$$\sum_{i=1}^{n} l_{\mathrm{sPTI}i} = l_{\mathrm{sPTI1}} + l_{\mathrm{sPTI2}} + \ldots + l_{\mathrm{sPTI}k} + \ldots + l_{\mathrm{sPTI}n} \quad (15)$$

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

where, $l_{\mathrm{vORF}i}$ represents the length of bona-fide ORFs and $\sum_{i=1}^{n} l_{\mathrm{sPTI}i}$ is the sum of the lengths of secondary PTI lengths. $\sum_{i=1}^{n} l_{\mathrm{vORF}i} + \sum_{i=1}^{n} l_{\mathrm{sPTI}i}$ 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

585 Ankarklev J, Jerlström-Hultqvist J, Ringqvist E, Troell K, Svärd SG. 2010. Behind
586 the smile: cell biology and disease mechanisms of Giardia species. *Nat Rev*
587 *Microbiol*. 8:413–422.
588
589 Bailey TL, Grimmond SM. 2006. The abundance of short proteins in the
590 mammalian proteome. *PLoS Genet*. 2, e52.
591
592 Bartelt LA, Sartor RB. 2015. Advances in understanding Giardia: determinants and
593 mechanisms of chronic sequelae. *F1000Prime Rep*. 7:62.
594
595 Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming
596 ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of
597 small ORFs in vertebrates using ribosome footprinting and evolutionary
598 conservation. *EMBO J*. 33:981–993.
599
600 Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved
601 genes identified from Drosophila yakuba and D. erecta accessory gland expressed
602 sequence tags. *Genetics*. 172:1675–1681.
603
604 Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for de novo evolution
605 of testis-expressed genes in the drosophila yakuba/drosophila erecta clade. *Genetics*.
606 176:1131–1137.
607
608 Brown SA, Warburton D, Brown LY, Yu CY, Roeder ER, Stengel-Rutkowski S,
609 Hennekam RC, Muenke M. 1998. Holoprosencephaly due to mutations in ZIC2, a
610 homologue of Drosophila odd-paired. *Nat Genet.* 20:180–183.
611
612 Buret AG, Cacciò SM, Favennec L, Svärd S. 2020. Update on Giardia: Highlights
613 from the seventh International Giardia and Cryptosporidium Conference. Mise à
614 jour sur Giardia et la giardiase: faits saillants de la Septième Conférence
615 Internationale sur Giardia et Cryptosporidium. *Parasite* 27:49.
616
617 Candeias MM. 2011. The can and can't dos of p53 RNA. *Biochimie* 93:1962–1965.

618
619 Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution.
620 New genes as drivers of phenotypic evolution. *Nat Rev Genet*. 14:645–660.
621
622 Djebali S, Davis CA, Merkel A, Dobin A, Lassermann T, Mortazavi A, Tanzer A,
623 Lagarde J, Lin W, Schlesinger F, Xue C, et al. 2012. Landscape of transcription in
624 human cells. *Nature* 489:101–108.
625
626 French SL, Santangelo TJ, Beyer AL, Reeve JN. 2007. Transcription and translation
627 are coupled in Archaea. *Mol Biol Evol*. 24:893–895.
628
629 Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kawai J, Carninci P, Hayashizaki
630 Y, Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse
631 embryonic stem cells reveals the complexity and dynamics of mammalian
632 proteomes. *Cell* 147:789–802.
633
634 Gripp KW, Wotton D, Edwards MC, Roessler E, Ades L, Meinecke P, Richieri-
635 Costa A, Zackai EH, Massagué J, Muenke M, et al. 2000. Mutations in TGIF cause
636 holoprosencephaly and link NODAL signalling to human neural axis determination.
637 *Nat Genet*. 25**:**205–208.
638
639 Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y,
640 Schäffer AA, Brister JR. 2017. Virus Variation Resource - improved response to
641 emergent viral outbreaks. *Nucleic Acids Res*. 45:D482-D490.
642
643 Hedges SB, Dudley J, Kumar S. 2006. TimeTree: A public knowledge-base of
644 divergence times among organisms. *Bioinformatics.* 22: 2971-2972.
645
646 Holland PW, Marlétaz F, Maeso I, Dunwell TL, Paps J. 2017. New genes from old:
647 asymmetric divergence of gene duplicates and the evolution of development. *Philos*
648 *Trans R Soc Lond B Biol Sci*. 372:20150480.
649

650 Huang DW, Sherman BT, Lempicki RA. 2009a. Systematic and integrative analysis
651 of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 4:44–57.
652
653 Huang DW, Sherman BT, Lempicki RA. 2009b. Bioinformatics enrichment tools:
654 paths toward the comprehensive functional analysis of large gene lists. *Nucleic*
655 *Acids Res.* 37:1–13.
656
657 Huang JZ, Chen M, Chen D, Gao X-C, Zhu S, Huang H, Hu M, Zhu H, Yan G-R.
658 2017. A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer
659 growth. *Mol Cell* 68: 171–184.
660
661 Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse
662 embryonic stem cells reveals the complexity and dynamics of mammalian
663 proteomes. Cell. 147: 789-802.
664
665 Ingolia NT. 2014. Ribosome profiling: new views of translation, from single codons
666 to genome scale. *Nat Rev Genet*. 15:205–213.
667
668 Islam SM, Suenaga Y, Takatori A, Ueda Y, Kaneko Y, Kawana H, Itami M, Ohira
669 M, Yokoi S, Nakagawara A. 2015. Sendai virus-mediated expression of
670 reprogramming factors promotes plasticity of human neuroblastoma cells.
671 106:1351–1361.
672
673 Kaneko Y, Suenaga Y, Islam SM, Matsumoto D, Nakamura Y, Ohira M, Yokoi S,
674 Nakagawara A. 2015. Functional interplay between MYCN, NCYM and OCT4
675 promotes aggressiveness of human neuroblastomas. *Cancer Sci.* 106:840–847.
676

677 Kato K, Miya F, Hamada N, Negishi Y, Narumi-Kishimoto Y, Ozawa H, Ito H,
678 Hori I, Hattori A, Okamoto N, et al. 2019. *MYCN* de novo gain-of-function mutation
679 in a patient with a novel megalencephaly syndrome. *J Med Genet.* 56:388-395.
680
681 Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res*. 12:656–664.
682
683 Kloc M, Foreman V, Reddy SA. 2011. Binary function of mRNA. *Biochimie*
684 93:1955–1961.
685
686 Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding
687 genes. *Genome Res*. 19:1752–1759.
688
689 Kopp F, Mendell JT. 2018. Functional classification and experimental dissection of
690 long noncoding RNAs. *Cell*. 172:393–407.
691
692 Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes
693 derived from noncoding DNA in drosophila melanogaster are frequently X-linked
694 and exhibit testis-biased expression. *Proc Natl Acad Sci U S A*. 103:9935–9939.
695
696 Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. 2009.
697 Identification of the novel protein QQS as a component of the starch metabolic
698 network in Arabidopsis leaves. *Plant J*. 58:485–498.
699
700 Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. 2010. A de novo originated gene
701 depresses budding yeast mating pathway and is repressed by the protein encoded by
702 its antisense strand. *Cell Res*. 20:408–420.
703
704 Li C-Y, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li X, Yu Q, Zheng
705 X, et al. 2010. A human-specific de novo protein-coding gene associated with
706 human brain functions. *PLoS Comput Biol*. 6:e1000734.
707
708 Li J, Liu C. 2019. Coding or noncoding, the converging concepts of RNAs. *Front*
709 *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 GSK3β 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.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat Ecol Evol*. 1:0146.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* 17:32–43.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556

Zhang YE, Long M. 2014. New genes contribute to genetic and phenotypic novelties in human evolution. *Curr Opin Genet Dev*. 29:90–96.

790

791 **Table 1.** Number of original transcripts that produced NR transcripts with high coding

792 frequency (0.6 ≤ 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	

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

794

795 **Figure legends**

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

813

814

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

817 (*Nitrososphaera viennensis* EN76), two fungi (*Puccinia graminis f. sp. Tritici* and
818 *Pyricularia oryzae*), and two animals (*Strongylocentrotus purpuratus* and *Lingula*
819 *anatine*) are unknown, and thus, excluded from the figure.
820
821 **Figure 3**. **Relationship between PTI score and relative frequencies of coding and**
822 **noncoding transcripts from bacteria to mammals**. Histogram of $f(x)$ (white) or $g(x)$
823 (black) in observed data (left) and in nucleic-acid–scrambled controls (right). PTI scores
824 with highest $f(x)$ are presented in the histograms.
825
826 **Figure 4. Relationship between PTI score and relative frequencies of coding f(x)**
827 **and noncoding transcripts in Primates (A), Glires (B), and Laurasiatheria (C).**
828
829 **Figure 5. Relationship between PTI score and protein-coding potential F(x) in 32**
830 **eukaryotes.** Phylogenetic tree of the 32 species (left), the dot plots and shape and
831 formula of approximate functions. L, S, and O indicate linear (in red), sigmoidal (in
832 black) and other (in blue) functions. Less than five lncRNAs had a PTI score of 0.05 in
833 *U. americanus*, *C. canadensis*, and *G. gorilla*. Therefore, we eliminated the $F(0.05)$ in
834 these species for linear function approximations (asterisks).
835
836 **Figure 6. Adaptation of (+) ssRNA viruses to host cells by maximizing protein-**
837 **coding potential of the RNA genomes.** (A) Conceptual explanation of the length of
838 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 $\leq$ 0.65 in human and mouse transcripts. (A) Relationship between

861 PTI score and $F(x)$ in a human data set from Ensemble and random controls (center).
862 Random shuffling controls (right) were generated from a human data set from both
863 Ensemble and RefSeq. (B) Relationship between PTI score and $F(x)$ in mouse
864 transcripts using data sets from RefSeq (upper panels) or Ensembl (lower panels).
865
866 **Supplementary Figure 4. Relationship between PTI score and relative frequencies**
867 **of coding and noncoding transcripts in archaea**. Phylogenetic tree for 9 archaeal
868 species and histogram of $f(x)$ (white) or $g(x)$ (black) in the data (left) and in nucleic-
869 acid–scrambled controls (right). PTI scores with highest $f(x)$ are indicated in the
870 histograms. The lineage of one archaea (*Nitrososphaera viennensis* EN76) is unknown
871 and thus excluded from the phylogenic tree.
872
873 **Supplementary Figure 5. Relationship between PTI score and relative frequencies**
874 **of coding and noncoding transcripts in plants.** Phylogenetic tree for 12 plants and
875 histogram of $f(x)$ (white) or $g(x)$ (black) in the data (left) and in nucleic-acid–scrambled
876 controls (right). PTI scores with highest $f(x)$ are indicated.
877
878 **Supplementary Figure 6. Relationship between PTI score and protein-coding**
879 **potential $F(x)$ in 32 eukaryotes.** Relationship between PTI score and $F(x)$ in data sets
880 from Ensemble (left, used in Figure 5) and those in random controls (right). Mouse and
881 human data are identical to Supplementary Figure 3. Shape of approximate functions
882 are shown as L, S, or O, indicating linear (in red), sigmoidal (in black), and other (in

883 blue) functions, respectively. The number of lncRNAs at PTI score 0.05 were < 5 in *U.*
884 *americanus*, *C. canadensis*, and *G. gorilla*. Therefore, we eliminated the *F*(0.05) in
885 these species for the approximation by linear functions (asterisks).
886
887 **Supplementary Table 1. Cellular organisms with the official name, taxonomy ID,**
888 **lineage information, and number of coding or noncoding transcripts.**
889
890 **Supplementary Table 2. Twenty-three human noncoding transcripts showing both**
891 **negative selection (*K*a/*K*s < 0.5) and the highest PTI score.**
892
893 **Supplementary Table 3. Human NR transcripts reassigned as NM over the past**
894 **three years.**
895
896 **Supplementary Table 4. Human NR transcripts with high protein-coding potential**
897 **(0.6 ≤ PTI score < 0.8).**
898
899 **Supplementary Table 5. Funtional annotation of human NR transcripts with high**
900 **protein-coding potential and without putative domain structure(s).**
901
902 **Supplementary Table 6. Functional annotation of human NR transcripts with high**
903 **protein-coding potential and with putative domain structure(s).**
904

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

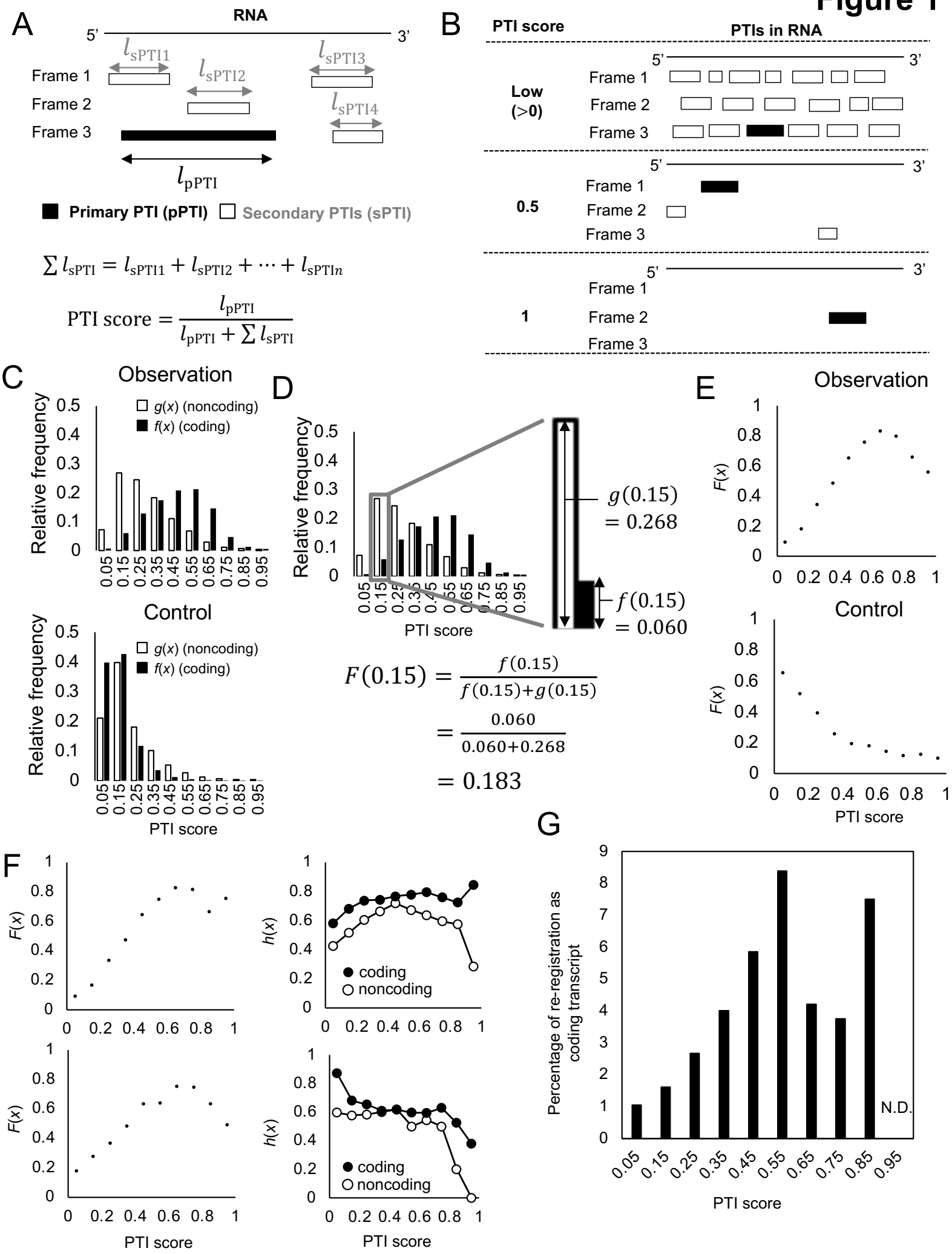

A
RNA
5'
3'
l_{sPTI1}
l_{sPTI2}
l_{sPTI3}
l_{sPTI4}
Frame 1
Frame 2
Frame 3
l_{pPTI}
Primary PTI (pPTI)
Secondary PTIs (sPTI)
$\sum l_{\mathrm{sPTI}} = l_{\mathrm{sPTI1}} + l_{\mathrm{sPTI2}} + \cdots + l_{\mathrm{sPTI}n}$
$\mathrm{PTI\ score} = \frac{l_{\mathrm{pPTI}}}{l_{\mathrm{pPTI}} + \sum l_{\mathrm{sPTI}}}$
B
PTI score
PTIs in RNA
Low (>0)
0.5
1
C
Observation
Control
Relative frequency
g(x) (noncoding)
f(x) (coding)
PTI score
D
$g(0.15) = 0.268$
$f(0.15) = 0.060$
$F(0.15) = \frac{f(0.15)}{f(0.15)+g(0.15)} = \frac{0.060}{0.060+0.268} = 0.183$
E
Observation
Control
F(x)
PTI score
F
F(x)
h(x)
coding
noncoding
PTI score
G
Percentage of re-registration as coding transcript
N.D.
PTI score

Figure 2

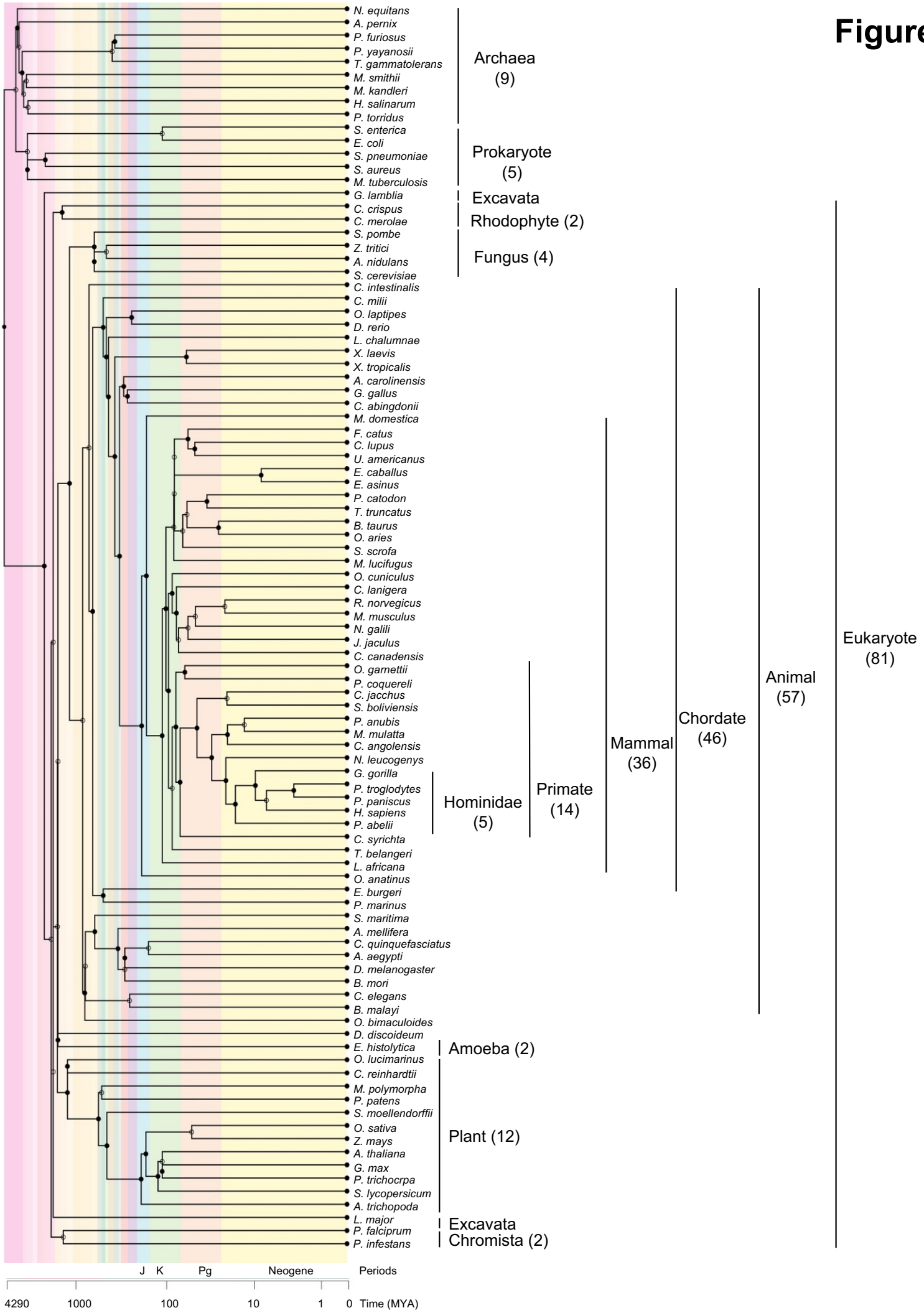

N. equitans
A. pernix
P. furiosus
P. yayanosii
T. gammatolerans
M. smithii
M. kandleri
H. salinarum
P. torridus
S. enterica
E. coli
S. pneumoniae
S. aureus
M. tuberculosis
G. lamblia
C. crispus
C. merolae
S. pombe
Z. tritici
A. nidulans
S. cerevisiae
C. intestinalis
C. milii
O. laptipes
D. rerio
L. chalumnae
X. laevis
X. tropicalis
A. carolinensis
G. gallus
C. abingdonii
M. domestica
F. catus
C. lupus
U. americanus
E. caballus
E. asinus
P. catodon
T. truncatus
B. taurus
O. aries
S. scrofa
M. lucifugus
O. cuniculus
C. lanigera
R. norvegicus
M. musculus
N. galili
J. jaculus
C. canadensis
O. garnettii
P. coquereli
C. jacchus
S. boliviensis
P. anubis
M. mulatta
C. angolensis
N. leucogenys
G. gorilla
P. troglodytes
P. paniscus
H. sapiens
P. abelii
C. syrichta
T. belangeri
L. africana
O. anatinus
E. burgeri
P. marinus
S. maritima
A. mellifera
C. quinquefasciatus
A. aegypti
D. melanogaster
B. mori
C. elegans
B. malayi
O. bimaculoides
D. discoideum
E. histolytica
O. lucimarinus
C. reinhardtii
M. polymorpha
P. patens
S. moellendorffii
O. sativa
Z. mays
A. thaliana
G. max
P. trichocrpa
S. lycopersicum
A. trichopoda
L. major
P. falciprum
P. infestans
Archaea (9)
Prokaryote (5)
Excavata
Rhodophyte (2)
Fungus (4)
Hominidae (5)
Primate (14)
Mammal (36)
Chordate (46)
Animal (57)
Eukaryote (81)
Amoeba (2)
Plant (12)
Excavata
Chromista (2)
J K Pg Neogene Periods
4290 1000 100 10 1 0 Time (MYA)

Figure 3

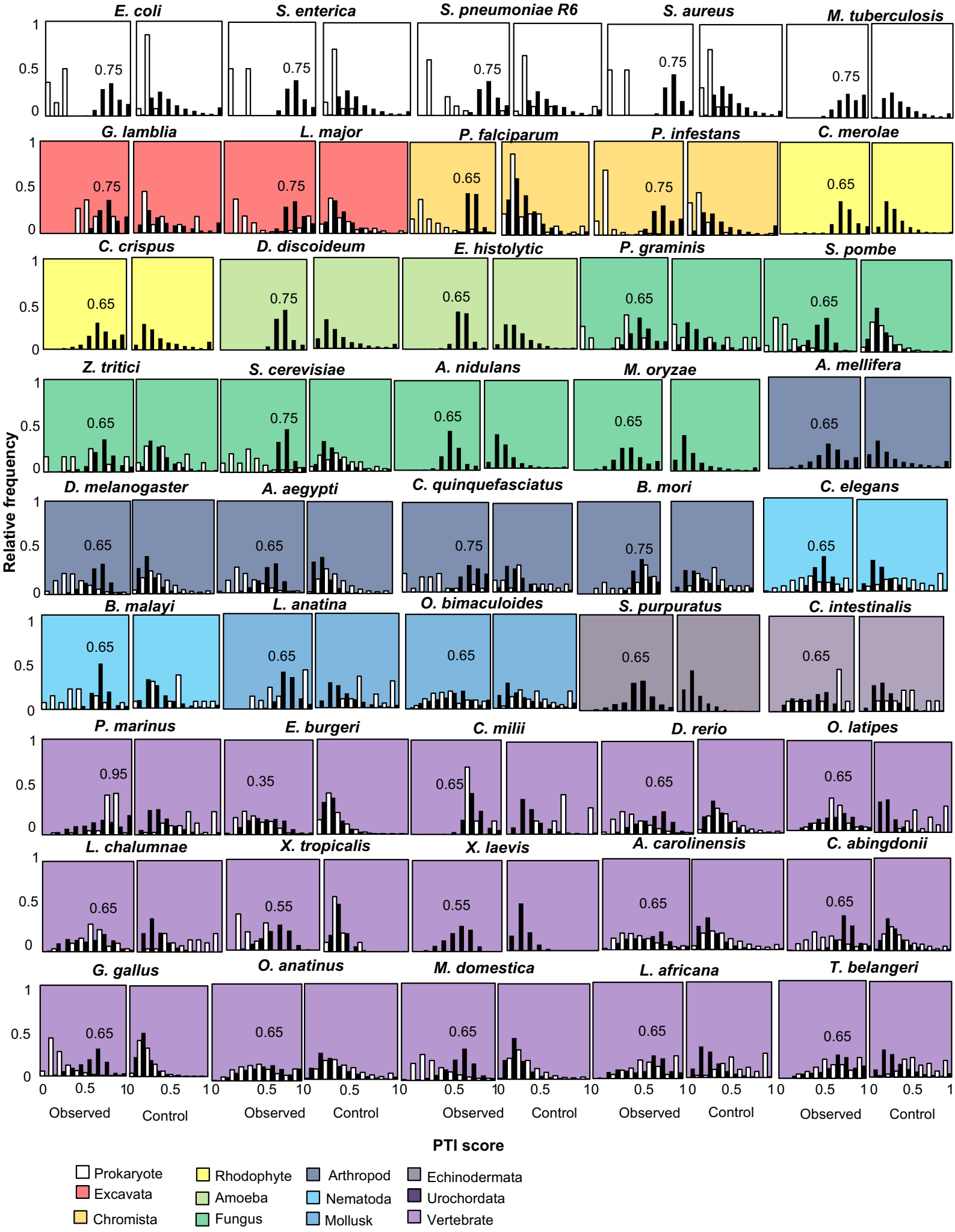

E. coli
S. enterica
S. pneumoniae R6
S. aureus
M. tuberculosis
G. lamblia
L. major
P. falciparum
P. infestans
C. merolae
C. crispus
D. discoideum
E. histolytic
P. graminis
S. pombe
Z. tritici
S. cerevisiae
A. nidulans
M. oryzae
A. mellifera
D. melanogaster
A. aegypti
C. quinquefasciatus
B. mori
C. elegans
B. malayi
L. anatina
O. bimaculoides
S. purpuratus
C. intestinalis
P. marinus
E. burgeri
C. milii
D. rerio
O. latipes
L. chalumnae
X. tropicalis
X. laevis
A. carolinensis
C. abingdonii
G. gallus
O. anatinus
M. domestica
L. africana
T. belangeri
Relative frequency
Observed
Control
PTI score
Prokaryote
Excavata
Chromista
Rhodophyte
Amoeba
Fungus
Arthropod
Nematoda
Mollusk
Echinodermata
Urochordata
Vertebrate

Figure 4

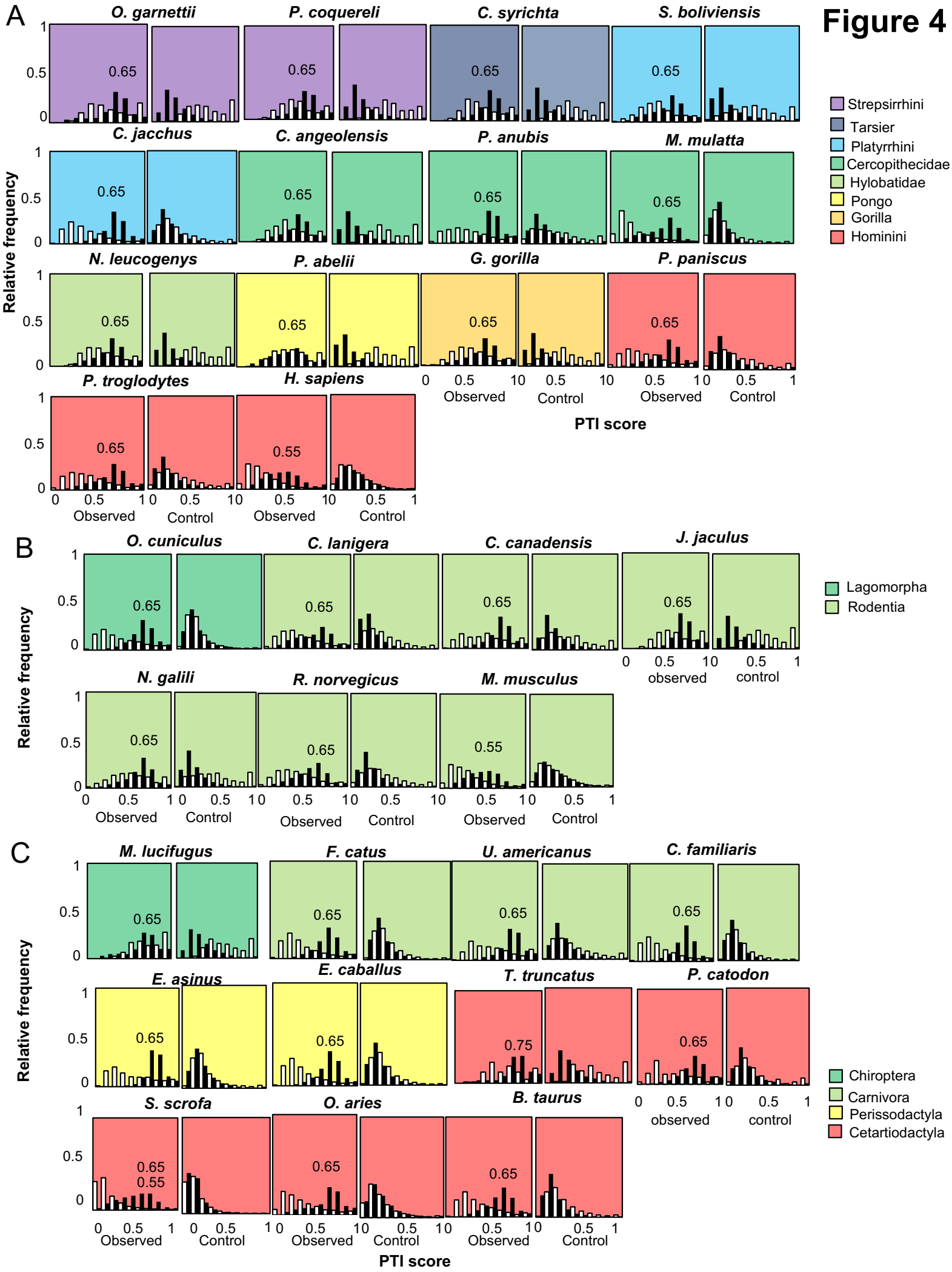

A
O. garnettii
P. coquereli
C. syrichta
S. boliviensis
C. jacchus
C. angeolensis
P. anubis
M. mulatta
N. leucogenys
P. abelii
G. gorilla
P. paniscus
P. troglodytes
H. sapiens
Relative frequency
PTI score
Observed
Control
Strepsirrhini
Tarsier
Platyrrhini
Cercopithecidae
Hylobatidae
Pongo
Gorilla
Hominini
B
O. cuniculus
C. lanigera
C. canadensis
J. jaculus
N. galili
R. norvegicus
M. musculus
observed
control
Lagomorpha
Rodentia
C
M. lucifugus
F. catus
U. americanus
C. familiaris
E. asinus
E. caballus
T. truncatus
P. catodon
S. scrofa
O. aries
B. taurus
Chiroptera
Carnivora
Perissodactyla
Cetartiodactyla

Figure 5

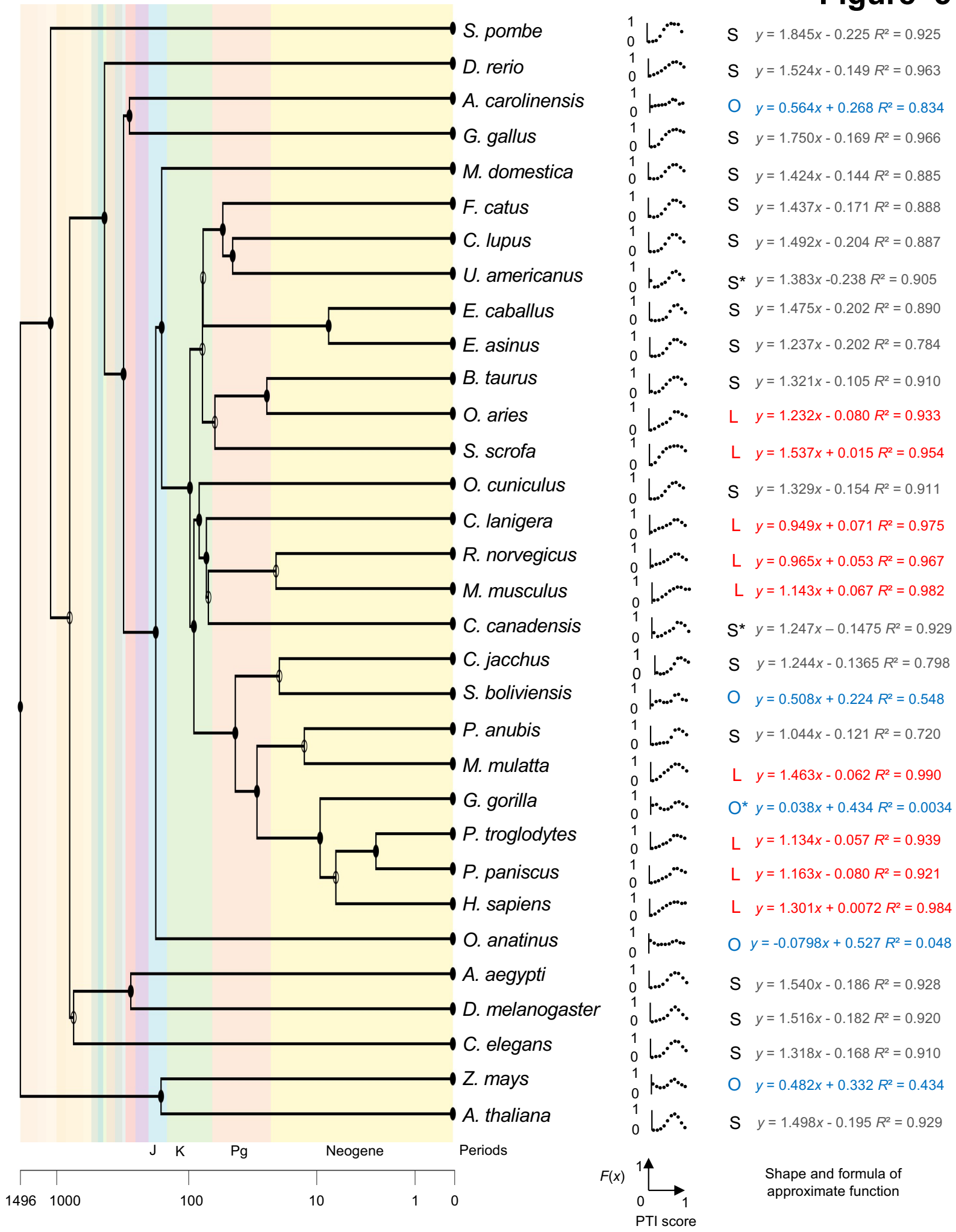
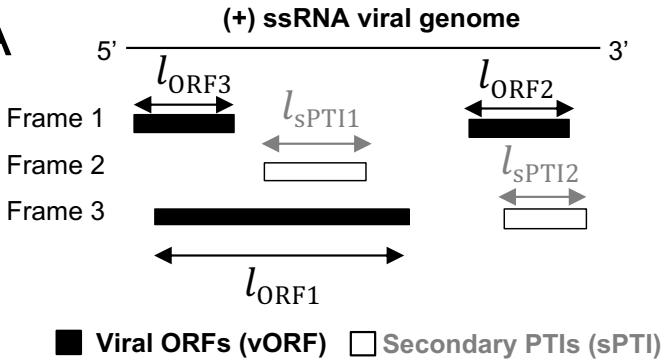

S. pombe
D. rerio
A. carolinensis
G. gallus
M. domestica
F. catus
C. lupus
U. americanus
E. caballus
E. asinus
B. taurus
O. aries
S. scrofa
O. cuniculus
C. lanigera
R. norvegicus
M. musculus
C. canadensis
C. jacchus
S. boliviensis
P. anubis
M. mulatta
G. gorilla
P. troglodytes
P. paniscus
H. sapiens
O. anatinus
A. aegypti
D. melanogaster
C. elegans
Z. mays
A. thaliana
S y = 1.845x - 0.225 R² = 0.925
S y = 1.524x - 0.149 R² = 0.963
O y = 0.564x + 0.268 R² = 0.834
S y = 1.750x - 0.169 R² = 0.966
S y = 1.424x - 0.144 R² = 0.885
S y = 1.437x - 0.171 R² = 0.888
S y = 1.492x - 0.204 R² = 0.887
S* y = 1.383x -0.238 R² = 0.905
S y = 1.475x - 0.202 R² = 0.890
S y = 1.237x - 0.202 R² = 0.784
S y = 1.321x - 0.105 R² = 0.910
L y = 1.232x - 0.080 R² = 0.933
L y = 1.537x + 0.015 R² = 0.954
S y = 1.329x - 0.154 R² = 0.911
L y = 0.949x + 0.071 R² = 0.975
L y = 0.965x + 0.053 R² = 0.967
L y = 1.143x + 0.067 R² = 0.982
S* y = 1.247x – 0.1475 R² = 0.929
S y = 1.244x - 0.1365 R² = 0.798
O y = 0.508x + 0.224 R² = 0.548
S y = 1.044x - 0.121 R² = 0.720
L y = 1.463x - 0.062 R² = 0.990
O* y = 0.038x + 0.434 R² = 0.0034
L y = 1.134x - 0.057 R² = 0.939
L y = 1.163x - 0.080 R² = 0.921
L y = 1.301x + 0.0072 R² = 0.984
O y = -0.0798x + 0.527 R² = 0.048
S y = 1.540x - 0.186 R² = 0.928
S y = 1.516x - 0.182 R² = 0.920
S y = 1.318x - 0.168 R² = 0.910
O y = 0.482x + 0.332 R² = 0.434
S y = 1.498x - 0.195 R² = 0.929
J K Pg Neogene Periods
1496 1000 100 10 1 0
F(x)
PTI score
Shape and formula of approximate function

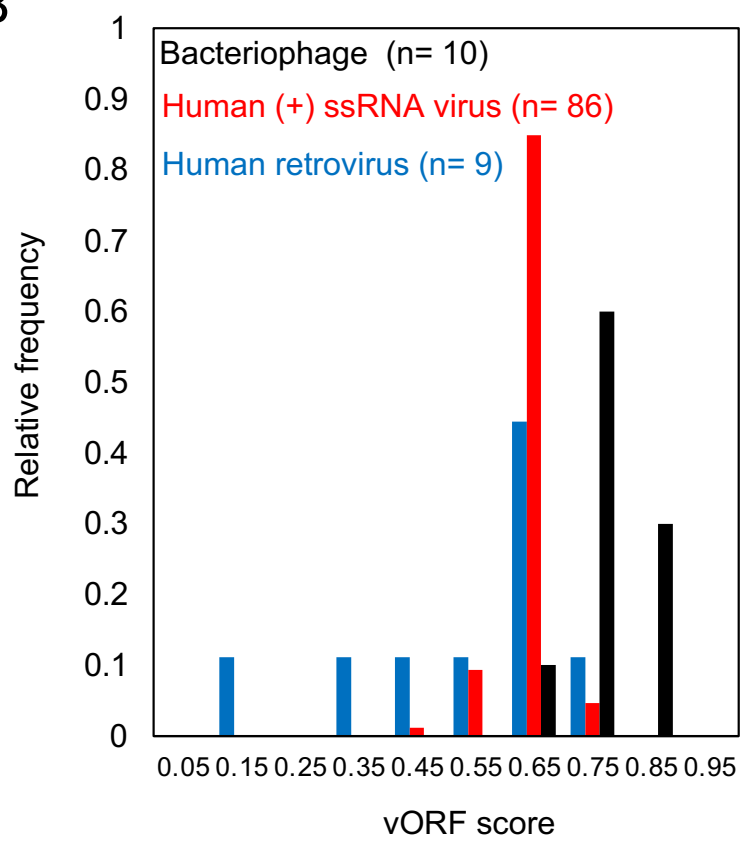
Figure 6

A



$$\sum l_{vORF} = l_{vORF1} + l_{vORF2} + \cdots + l_{vORFn}$$

$$\sum l_{sPTI} = l_{sPTI1} + l_{sPTI2} + \cdots + l_{sPTIn}$$

$$\text{vORF score} = \frac{\sum l_{vORF}}{\sum l_{vORF} + \sum l_{sPTI}}$$

B