Exploring the selective constraint on the sizes of insertions and deletions in 5’ untranslated regions in mammals

Chun-Hsi Chen¹, Ben-Yang Liao¹* and Feng-Chi Chen¹,²,³*

Abstract

Background: Small insertions and deletions (“indels” with size ≦ 100 bp) whose lengths are not multiples of three (non-3n) are strongly constrained and depleted in protein-coding sequences. Such a constraint has never been reported in noncoding genomic regions. In 5’ untranslated regions (5’UTRs) in mammalian genomes, upstream start codons (uAUGs) and upstream open reading frames (uORFs) can regulate protein translation. The presence of non-3n indels in uORFs can potentially disrupt the functions of these regulatory elements. We thus hypothesize that natural selection disfavors non-3n indels in 5’UTRs when these regulatory elements are present.

Results: We design the Indel Selection Index to measure the selective constraint on non-3n indels in 5’UTRs. The index controls for the genomic compositions of the analyzed 5’UTRs and measures the probability of non-3n indel depletion downstream of uAUGs. By comparing the experimentally supported transcripts of human-mouse orthologous genes, we demonstrate that non-3n indels downstream of two types of uAUGs (alternative translation initiation sites and the uAUGs of coding sequence-overlapping uORFs) are underrepresented. The results hold well regardless of differences in alignment tool, gene structures between human and mouse, or the criteria in selecting alternatively spliced isoforms used for the analysis.

Conclusions: To our knowledge, this is the first study to demonstrate selective constraints on non-3n indels in 5’UTRs. Such constraints may be associated with the regulatory functions of uAUGs/uORFs in translational regulation or the generation of protein isoforms. Our study thus brings a new perspective to the evolution of 5’UTRs in mammals.
within 5'UTR; (ii) CDS-overlapping uORFs (“overlapping uORFs”) - uORFs with their start codons located in 5'UTR and their stop codons located within the CDS; and (iii) alternatively translated uORFs (“alternative uORFs”) - uORFs that start with an alternative initiation site of translation (“AIS”, a potential translation initiation site located in 5'UTR), and share the same reading frame and the same stop codon with the main CDS (Figure 1A). Note that the protein products that include alternative uORFs are currently unidentified. Therefore, for extensively studied species such as human and mouse, AISs are less likely to be misannotated canonical translation start sites. The scanning model for translation [15] posits that a translation process begins with the binding of 40S ribosomal complexes onto the 5’-cap structure of an mRNA. The ribosomal complexes then slide from 5’-cap to 3’ end base by base until they encounter the first AUG triplet of the mRNA and turn on the translation process. According to this model, different types of uORF have distinct effects on the translation of the main CDS: (1) strict uORFs compete for ribosomes, leading to reduced protein production of the main CDS [16]; (2) overlapping uORFs cause translation re-initiation downstream of the main start codon, resulting in N-truncation of the peptide coded by the main CDS [10,17]; they may also cause strong translational inhibition by facilitating ribosome skipping of the main start codon [10,18,19]; and (3) alternative uORFs result in the generation of N-extended peptides [17,20,21]. Notably, non-3n indels downstream of uAUGs can lead to interchanges between uORF types (Figure 1B), while 3n indels usually can not. Such interchanges may result in either changes of protein abundance (quantitative changes) or gain/loss of protein isoforms (qualitative changes) (Figure 1B). Considering that the level of protein translation is tightly constrained during evolution [22], most changes between uORF types are supposedly deleterious. Therefore, we hypothesize that in 5'UTRs, non-3n indels that occur downstream of uAUGs are subject to purifying selection.

To test our hypothesis, we develop the Indel Selection Index (ISI, see Methods) to examine whether the occurrence of non-3n indels between the first uAUGs and the translation initiation site (“TIS”, also known as the main start codon) of the main CDS are selectively disfavored during the evolution of mammalian genes after primate-rodent divergence. Since non-3n indels may affect protein expression more seriously when they occur downstream of the uAUGs of alternative uORF and overlapping uORFs (Figure 1B), purifying selection on non-3n indels is expected to be particularly stringent in these regions, while relatively relaxed in the case of strict uORFs. Our results clearly support this hypothesis. This study thus offers a new perspective to the evolution of 5'UTRs, in that the sizes of indels can be subject to selective constraint in these genomic regions conditional on the presence of certain regulatory elements.

Results

We compared human and mouse orthologous genes to examine whether non-3n indels are selectively constrained downstream of uAUGs. While the range of 5'UTRs varies in alternatively spliced mRNA downstream of uAUGs (or a reference point), as compared to the 5'UTR region upstream of the uAUG (or reference point). In other words, a small ISI indicates a depletion of non-3n indels downstream of 5'UTR. This study thus offers a new perspective to the evolution of 5'UTRs, in that the sizes of indels can be subject to selective constraint in these genomic regions conditional on the presence of certain regulatory elements.

According to our criteria, more than 45% human genes and 41% mouse genes have at least one uORF (Table 1). These percentages are similar to those reported in previous studies [16,24]. Next, we divided the transcripts of each dataset into four subgroups for comparisons: (1) transcripts without uAUGs (G0); (2) transcripts with only alternative uORF(s) (Ga); (3) transcripts with only strict uORF(s) (Gs; their uAUGs are designated as “SuAUGs”); and (4) transcripts with only overlapping uORF(s) (Gv; their uAUGs are designated as “VuAUGs”). Transcripts with multiple types of uORFs were excluded from this study for simplicity.
Figure 1 Classification of upstream open reading frames (uORFs) and potential effects of non-3n indels when they occur between uAUGs and TISs. (A) The open circle represents the 5'-cap structure of the transcript. The solid- and dashed-line open boxes represent the main coding sequences and the uORFs, respectively. The open and solid inverted triangles, respectively, indicate the locations of the uAUGs and the translation initiation sites (TIS). The open and solid triangles indicate locations of the stop codons of the uORFs and the main coding sequences. (B) Symbols “○”, “×”, and “?” represent that the protein isoforms or protein expression “is affected”, “is not affected”, and “uncertain”, respectively.
of a certain 5'UTR position. The use of ISI thus controls for the properties potentially specific to 5'UTRs. We first analyzed the ISI distribution of the G0 transcripts, where the ratios of non-3n to 3n indels are expected to be approximately equal between the “upstream” and “downstream” regions to a given position of the 5’UTRs. We assigned a reference point to each of the transcript and shifted the point from 10% to 90% of the 5’UTR lengths (with intervals of 10%) from the cap to obtain the ISI values. ISI values of the G0 transcripts vary with the reference point position (Additional file 1). For both human and mouse, the median ISI values remain approximately equal between the 30~70% of the 5’UTR lengths from the 5’ cap, but drop toward both ends of 5’UTR. Therefore, controlling the ratio of upstream/downstream length is necessary when analyzing the ISI values. We then compared the ISI values of uAUG-containing transcripts (Gs, Ga, and Gv) with those of G0 transcripts with corresponding upstream/downstream ratios. Note that the uAUG-containing transcripts may have multiple uAUGs of the same type. In such cases, we used the first uAUG from the cap as the reference point.

We found that Gs transcripts consistently show significantly lower ISIs than the corresponding G0 transcripts in all of the six comparisons (P < 10E-8, all of the P values for the ISI value comparisons were estimated by using the Mann-Whitney U test, Figure 2). Meanwhile, the Gs transcripts have significantly lower-than-expected ISIs (P < 0.05) in five of the six comparisons. The only exception lies in mouse transcripts with pure 5’UTRs (P = 0.216), although the median ISI of the Gs transcripts is lower than that of G0 for this dataset. The lack of statistical significance might have resulted from the relatively small sample sizes (only 33 indel-containing Gs transcripts are available for both human and mouse). In comparison, in only two of the six comparisons (randomly selected 5’UTRs in both human and mouse) do Gs transcripts have significantly lower ISIs than expected (P < 0.05). We noticed that the sizes of indels might change with alignment tools [25], which could potentially affect our results. Therefore, we used Pecan [26], another alignment tool known for its accuracy, to re-align the human-mouse orthologous sequences and obtained similar results (Additional file 2).

Notably, the locations of reference points of the corresponding G0 transcripts actually differ significantly among Gs, Ga, and Gv. In view of the variations of ISI values with different reference point locations (Additional file 1), the comparisons among the three transcript groups appear unfair. Particularly, all of the Gs transcripts are compared against the Gv transcripts with their reference points located at ~25% from the cap, whereas the percentages for Ga and Gv fall between 54% ~69% (Figure 2). To address this issue, we divided the Gv transcripts into three equal-sized groups according to the relative positions of their uAUGs (we did not perform the analysis for Ga because of its small sample size). As shown in Additional file 3 the first Gv subgroups (Gv_1) were compared against G0 transcripts with reference points located at 14%~16% from the cap for human, and 24%~26% for mouse. These Gv transcripts have uAUGs located closer to the cap than their Gv counterpart, and they still have ISI values significantly lower than the corresponding G0 transcripts. Similar results are also observed for Gv_2 (Additional file 3). Gv_3 transcripts show a similar trend, although the differences in ISI values are statistically insignificant, possibly due to reduced sample sizes.

Apparently, non-3n indels are significantly more depleted when occurring downstream of AISs and VuAUGs than SuAUGs. However, since Gs, Ga, and Gv have different sample sizes, we are interested in comparing how the ISIs of Gs, Ga, and Gv deviate from the expected values when controlling for the difference in sample size. We performed a bootstrap simulation (with 1,000 re-samplings with replacement) for each of the six comparisons by reducing the sample sizes of Gs and Gv transcripts to be the same as that of the G0 transcripts. As shown in Figure 3, Gs-G0 and Gv-G0 comparisons have smaller P-

---

**Table 1 Transcripts of human-mouse orthologous genes analyzed in this study**

| Type               | Class | Randomly-selected 5’UTR | Longest 5’UTR | Pure 5’UTR |
|--------------------|-------|-------------------------|---------------|------------|
|                    |       | Human | Mouse | Human | Mouse | Human | Mouse |
| Without uORF       | G0    | 3,265 (54.0%) | 3,560 (58.9%) | 2,701 (46.6%) | 3,153 (54.4%) | 3,144 (55.2%) | 3,368 (59.1%) |
|                    | Gs    | 73 (1.2%)   | 61 (1.0%)   | 99 (1.7%)   | 76 (1.3%)   | 38 (0.7%)   | 40 (0.7%)   |
| Single uORF type   | G0    | 1,558 (25.8%) | 1,456 (24.1%) | 1,564 (27.0%) | 1,487 (25.6%) | 1,638 (28.7%) | 1,523 (26.7%) |
|                    | Gs    | 401 (6.6%)  | 385 (6.4%)  | 356 (6.1%)  | 333 (5.7%)  | 380 (6.7%)  | 345 (6.1%)  |
| Multiple types of uORF | G0 | 749 (12.4%) | 584 (9.7%) | 1,080 (18.6%) | 751 (12.9%) | 500 (8.8%) | 424 (7.4%) |
| Total              |       | 6,046 | 6,046 | 5,800 | 5,800 | 5,700 | 5,700 |

* G0, Ga, Gs, and Gv indicate transcripts without uAUGs, with AISs, with SuAUGs, and with VuAUGs, respectively.

**For each gene, only one transcript is selected (a randomly selected transcript, or the one that has the longest 5’UTR or pure 5’UTR).**
values (more significant differences) than the G\textsubscript{c}-G\textsubscript{0} comparisons ($P < 4.0\text{E}-12$ by the Kolmogorov-Smirnov Test).

Therefore, when the factor of sample size is controlled, the overall result that indels are more depleted downstream of AISs and VuAUGs holds well. Notably, even in the G\textsubscript{c}-G\textsubscript{0} comparison in the case of pure 5'UTRs (which gives the most conservative estimation), more than 6% of the $P$ values are smaller than 0.05 for both human and mouse. This observation suggests that non-3n indels may be subject to weak purifying selection pressure downstream of SuAUGs.

### Discussion

#### Possible reasons for non-3n indel depletion in 5'UTRs

We have demonstrated that non-3n indels are subject to purifying selection in mammalian 5'UTRs conditional on the presence of uORFs. We show that both alternative and overlapping uORFs contribute to decreased

---

**Figure 2** Distributions of ISI values of (from left to right) (Ga VS. G\textsubscript{0}), (Gs VS. G\textsubscript{0}), and (Gv VS. G\textsubscript{0}) for (A) human and (B) mouse. The numbers in the parentheses following G\textsubscript{0} indicate the median distances of the uAUGs from 5' cap in terms of percentage of 5'UTR length in the non-G\textsubscript{0} transcripts. These proportions of length are referenced to determine which G\textsubscript{0} distributions to use in the comparisons. The $P$ values of pair-wise differences (calculated by using the Mann-Whitney $U$ test) are shown at the top. The symbols "*", "**", and "***" represent $0.01 \leq P < 0.05$, $0.001 \leq P < 0.01$, and $P < 0.001$, respectively.

---

| Randomly selected 5'UTR | Longest 5'UTR | Pure 5'UTR |
|--------------------------|---------------|------------|
|                          |               |            |
| $P = 0.002$              | $P = 0.044$   | $P = 3.17E-20$ |
| $G_a$                    | $G_a$         | $G_a$      |
| $G_a(5%)$                | $G_{025}(5%)$ | $G_{025}(5%)$ |
| $G_a(25%)$               | $G_{025}(25%)$ | $G_{025}(25%)$ |
| $G_a(60%)$               | $G_{025}(60%)$ | $G_{025}(60%)$ |

**Figure 2**

To view this figure, please refer to the original document.
non-3n indels downstream of their uAUGs, and that strict uORFs have only minor effects in this regard. These results hold well when technical issues in transcript isoform selection, difference in alignment tool, and differences in transcript structures between human and mouse are controlled.

The suppression of non-3n indels downstream of AISs and VuAUGs implies the functional importance of these two uAUG types. Two possible reasons may explain this observation. The first is the functional importance of uORF-associated protein products. Translation of overlapping uORFs always generates radically different protein products from the same transcript. The second reason is that the suppression of non-3n indels downstream of AISs and VuAUGs could be a result of translational regulation.
peptides from those translated from the main CDS because of the difference in reading frame [27]. Strict uORFs can also produce functional proteins when translated [28,29]. Therefore, additional constraints on non-3n indels unrelated to peptide coding may have separated strict uORFs from the other two types of uORFs. Another explanation is that overlapping uORFs can give rise to in-frame N-truncated peptides, which may have different molecular functions from the original, full-length peptides [17]. Such N-truncated peptides may change in length or simply disappear if non-3n indels occur downstream of VuAUGs. On the other hand, alternative uORFs can lead to the generation of N-extended peptides, which could have different functions from original peptides (e.g. the human regulators of G-protein signalling (RGS2)) [30]. The functional disruption of such N-extended or N-truncated peptides by non-3n indels may be detrimental to the organism in general, and thus could have been removed by natural selection. A recent study provides evidence of the importance of alternative translation start sites by showing that start codons downstream of TISs are evolutionarily conserved [31]. It is suggested that alternative start sites may work as “backup” translational initiation sites and thus may increase the efficiency of translation [31]. The same comment likely also applies to the AISs analyzed in this study.

The second possible explanation for the uAUG-related selection pressure on non-3n indels is the evolutionary conservation of protein abundance. As shown in Figure 1B, non-3n indels may cause interchanges between the three types of uORFs (Figure 1B). In the case of alternative uORFs, downstream non-3n indels can lead to the generation of either strict or overlapping uORFs, causing strong inhibition of protein production of the main CDS in both cases. Such drastic changes in protein abundance are likely deleterious. By contrast, when a non-3n indel occurs downstream of an SuAUG, the affected strict uORF may either become an alternative or overlapping uORF, or remain a strict uORF (but with a different length) (Figure 1B). We suggest that the latter case is more likely, for changing a strict uORF to an alternative or overlapping uORF requires that the reading frame starting from the uAUG remain open until it reaches the TIS. Furthermore, in the case of alternative uORF, the reading frame must be the same as that used by the main CDS. These requirements are difficult to fulfil considering that SuAUGs are relatively distant from the TIS (about 70–75% of the 5′UTR length). As such, non-3n indels that occur downstream of SuAUGs may not lead to changes in uORF type in most cases. Such indels may have no significant effects on changing the protein abundance of the downstream CDS, and thus may have small fitness effects. Lastly, in the case of overlapping uORFs, the occurrence of non-3n indels has a higher possibility of changing them into alternative uORFs than in the case of strict uORFs, because VuAUGs are typically closer to the TIS (about 40% of the 5′UTR length). Overlapping uORFs can result in nearly complete inhibition of protein translation or generation of N-truncated protein [10,18,19]. Furthermore, overlapping uORFs can serve important regulatory roles [10]. Therefore, non-3n indels in overlapping uORFs may be selectively disfavoured. A non-3n indel may also change an overlapping uORF to a strict uORF, or simply change the length of the original uORF (without changing its type). In these cases, non-3n indels may not have significant effects in changing protein abundance, and thus may be tolerated by selection.

One unexpected observation from our results is that four of the six datasets demonstrate lower P values in the Gv transcripts than in the Ga transcripts (Figure 3), indicating that overlapping uORFs may have contributed stronger constraints on non-3n indels in 5′UTRs than alternative uORFs. That said, the real cause of this Gv–Ga difference remains unclear. Notably, it has been recently reported that 3′UTRs actually have a larger effect on protein abundance than 5′UTRs, which appear to account for ~1% of the variations in protein abundance [32]. How can we observe any selection pressure on non-3n indels in 5′UTRs if these non-coding regions have such a “small” effect on protein abundance? There are three possible explanations. Firstly, as we mentioned above, the non-3n indels in 5′UTRs may affect both the abundance and the peptide sequence of the affected gene. The “qualitative” change may also be subject to selection pressure. Secondly, even though on average 5′UTRs account for only a small proportion of the variations in protein abundance genome-widely, in individual genes the variations can be very large, as was demonstrated by a recent study [16]. The cases where 5′UTRs have very small effects on protein abundance may actually add to the noise in our analysis. However, we have found clear signals of selection pressure on non-3n indels despite these potential noises, which in fact reflects the strength of the “real” signals. Finally, the uORFs per se may be biologically functional, in terms of either their peptide products or their regulatory roles. The disruption of functional uORFs is thus likely subject to selective constraint.

**Limitations of the study**

The current analysis contains a few limitations. Firstly, determining which transcript of a gene to analyze is controversial. This study used three different criteria for transcript selection. In the case of the transcripts with the longest 5′UTRs, one uORF may partly overlap with
in coding sequences. In addition, the classification of the uORFs (alternative, strict, or overlapping) in these 5’UTRs is sometimes ambiguous (see Additional file 3 for an example). Analyzing the transcripts with pure 5’UTRs avoids this problem. However, this practice will lead to significantly decreased numbers of uORFs and severely reduced sample sizes, which in turn may result in decreased statistic power and potential sampling biases. These two criteria for transcript selection represent two extremes. The results derived using randomly selected transcripts fall in-between. Nevertheless, this study obtained consistent results across all datasets, indicating that depletion of non-3n indels is unlikely to result solely from constraints in the main coding sequences.

Secondly, since our study is based on pairwise sequence alignments, we cannot distinguish between insertions and deletions, nor can we infer the lineage specificity of the identified indels. We do not know exactly in which lineage the indels have affected the structures of the 5’UTRs (i.e. the types of uORFs). This is important because the 5’UTRs of human and mouse transcripts may have different lengths and uORFs. When a non-3n indel occurs to a lineage whose transcript does not contain any uORFs or 5’UTR exons, this indel adds to the noise in our analysis. We cannot eliminate such noises without using multiple-species sequence alignments (which, however, will dramatically decrease the sample size and render the analysis infeasible). To overcome this problem, this study performed analyses using the transcript structures of human and mouse separately. The results from both species turn out to be consistent with each other. Therefore, in spite of the above limitation, our results seem to have revealed a biological fact.

Conclusions
To the best of our knowledge, this is the first study to demonstrate the selective constraint on non-3n indels in 5’UTRs. This constraint may result from the requirement to preserve either the translational regulatory elements (uORFs) in 5’UTRs or the functions of the peptides whose translation is associated with uORFs. Our results thus demonstrate the impacts of indels in the evolution of 5’UTRs in mammalian genomes and reassure the functional importance of uORFs from a different angle.

Methods
Data collection
In this study, we analyze the human and mouse transcripts because their genomes have been fully sequenced and extensively curated [33,34]. In addition, the transcriptomes of these two species have been well characterized. The annotations for 5’UTRs are thus fairly accurate for the two species. The sequences of experimentally verified transcripts with known protein products of one-to-one human-mouse orthologous genes, based on the Ensembl release 54 http://www.ensembl.org, were retrieved through BioMart [35]. Non-protein-coding genes and protein-coding genes whose transcripts did not contain 5’UTRs were excluded. Potentially misannotated transcripts (whose locations of TISs were inconsistent with that observed in the DNA sequences) were also excluded. To avoid repetitive counts of the same indels, only one transcript was selected for each gene by three different criteria: (1) a randomly selected transcript, (2) the transcript with the longest 5’UTR, and (3) the transcript with a “pure” 5’UTR. A pure 5’UTR is one that does not overlap with any coding sequences in other splicing isoforms (See Additional file 4 for an example).

Sequence alignments and identification of indels and uORFs
Indels were identified based on the human-mouse pairwise genomic sequence alignments downloaded from the University of California, Santa Cruz (UCSC) Genome Browser http://genome.ucsc.edu/ [36]. The versions of the human and mouse genomes (hg18 and mm9, respectively) correspond to Ensembl release 54.

To ensure that human-mouse orthologous sequences were compared in our study, this work only retained the alignments that include one-to-one human-mouse orthologous genes annotated by Ensembl. In addition, to avoid mis-assignment of gaps (indels), the alignable exonic sequences in one species are required to overlap with > 80% of the annotated exonic sequences of the other species. The alignments must cover the entirety of the annotated 5’UTRs. Consequently, this study obtained ~6,000 human and mouse genes for subsequent analyses (Table 1). To examine whether different alignment tools affect the overall results, the Pecan alignment program [26] was used with default parameters to re-align the retrieved human-mouse orthologous sequences.

The 5’UTRs of the retrieved transcripts were then scanned for the existence of uAUGs. Around half of the analyzed human and mouse transcripts were found to have at least one uAUG (Additional file 5). These proportions are similar to those observed in previous studies [16,24]. Here, a uORF is defined as a putative open reading frame in 5’UTR starting with a uAUG and composed of at least nine nucleotides, including the stop codon. Human and mouse orthologous genes may have different 5’UTR structures and different numbers and
types of uORFs. Therefore, we performed our analyses according to the human and mouse gene annotations separately.

Measurement of selection pressure on indel lengths – the indel selection index

To evaluate the evolutionary constraints on indel lengths in different parts of a 5'UTR, we defined

$$R = \log_2 \left( \frac{(N_{n3n,d} + 0.5)/(N_{3n,d} + 0.5)}{(N_{n3n,u} + 0.5)/(N_{3n,u} + 0.5)} \right),$$  \hspace{1cm} (1)$$

where $N$ represents the number of indels, the subscripts “$d$” and “$u$” represent non-3n and 3n indels. The subscripts “$n3n$” represent non-3n and 3n indels. The ratio of non-3n to 3n indels upstream of uAUGs serves as the “background” to measure the depletion (or enrichment) of non-3n indels downstream of uAUGs.

Theoretically, if selection has no preference on indel sizes across 5'UTRs, $R$ should be equal to zero. However, when non-3n indels downstream of uAUGs are suppressed by purifying selection, $R$ should be smaller than 0. The statistical power of $R$ is expected to increase with the number of indels. Yet in our dataset, the lengths of more than 90% of the 5'UTRs are shorter than 1 kb (and thus the numbers of indels are small). The relatively small sample sizes may lead to unexpected biases. To address this issue, we developed the “Indel Selection Index” (ISI) as in our previous paper. The ISI is defined as

$$ISI = Pr \left( R_{\text{shuffled}} < R \right),$$  \hspace{1cm} (2)$$

which represents the proportion of the randomized $R$-values ($R_{\text{shuffled}}$) that is smaller than the observed $R$-value. The distribution of $R_{\text{shuffled}}$ was generated by randomly shuffling the locations of the indels 1,000 times for each transcript, while retaining the lengths and numbers of indels of the 5'UTR. An ISI value smaller than the neutral expectation indicates depletion of non-3n indels downstream of a uAUG. The neutral distribution of ISIs was derived from the transcripts without uAUGs (see Additional file 1), with the upstream/downstream differentiation point moving from the cap to TIS by an increment of 1% of 5'UTR length. Note that we use ISI rather than comparing the non-3n to 3n indel ratios between 5'UTR and intergenic/intronic regions to control for the potential biological differences between 5'UTR and the other noncoding regions.
References

1. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwala P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlov K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent M, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Cannicci P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002, 420(6915):520-562.

2. Consortium TCSA. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005, 437(7055):69-87.

3. Mills RE, Lutting CT, Larks JS, Beauchamp A, Tsai C, Pittard WS, Devine SJ. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 2006, 16(9):1182-1190.

4. Chen FC, Chen CJ, Li WH, Chuang TJ. Human-specific insertions and deletions inferred from mammalian genome sequences. Genome Res 2007, 17(1):16-22.

5. Chen CH, Chuang TJ, Liao BY, Chen FC. Scanning for the signatures of positive selection for human-specific insertions and deletions. Genome Biol Evol 2009, 1(1).

6. Lunter G, Ponting CP, Hein J. Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol 2006, 2(1):e5.

7. de la Chaux N, Messor PW, Armit PF. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. BMC Evol Biol 2007, 7(1):191.

8. Wilkie GS, Dickson KS, Gray NK. Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. Trends Biochem Sci 2003, 28(4):182-188.

9. Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell 2009, 136(4):731-745.

10. Jackson RJ, Heelen CJU, Pestova TV. The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol 2010, 11(2):113-127.

11. Kozak M. Effects of intercistronic length on the efficiency of reinitiation by eukaryotic ribosomes. Mol Cell Biol 1987, 7(10):3438-3445.

12. van der Velden AW, Thomas AA. The role of the 5' untranslated region of an mRNA in translation regulation during development. Int J Biochem Cell Biol 1999, 31(8):177-190.

13. Mignone F, Gissi C, Liuni S, Pesole G. Translated regions of mRNAs. Genome Biol 2002, 3(3):REVIEWS0004.

14. Kochetov AV. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. Bioessays, 2008, 30(7):663-671.

15. Kozak M. The scanning model for translation: an update. J Cell Biol 1989, 108(2):229-241.

16. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci USA 2009, 106(18):7507-7512.

17. Kochetov AV, Sarai A, Rogozin IB, Shumny VK, Kolchanov NA. The role of alternative translation start sites in the generation of human protein diversity. Mol Genet Genomics 2005, 273(6):491-496.

18. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 2005, 361:13-17.

19. Kozak M. Pushing the limits of the scanning mechanism for initiation of translation. Gene 2002, 299(1-2):1-14.

20. Sarrazin S, Stark J, Conen C, Deubnikovski A, Melet F, Morl F. Negative and translation termination-dependent positive control of Fli-1 protein synthesis by conserved overlapping 5' upstream open reading frames in Fli-1 mRNA. Mol Cell Biol 2000, 20(9):2959-2969.

21. Song KY, Choi HS, Hwang CK, Kim CS, Law PY, Wei LN, Loh HH. Differential use of an in-frame translation initiation codon regulates human mu opioid receptor (OPRM1). Cell Mol Life Sci 2009, 66(17):2933-2942.

22. Schirmpf SP, Weiss M, Reiter L, Ahmad CH, Jovanovic M, Malmstrom J, Brunner E, Mohanty S, Lercher MJ, Hunziker PE, Aebersold R, von Mering C, Hengartner MO. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. PLoS Biol 2009, 7(3):e1000056.

23. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Schef F, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O’Keefe S, Haas S, Vingron M, Lehrach H, Yaso ML. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 2008, 321(5891):956-960.

24. Reich AM, Ogurtsov AV, Rogozin IB, Shabalina SA, Koonin EV. Evolution of alternative and constitutive regions of mammalian 5'UTRs. BMC Genomics 2009, 10:162.

25. Chen X, Tompa M. Comparative assessment of methods for aligning multiple genome sequences. Nat Biotechnol 2010, 28(6):567-572.

26. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E, Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res 2008, 18(11):1814-1828.

27. Klemke R, Kehlenbach RH, Huttner WB. Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. EMBO J 2001, 20(14):3849-3860.

28. Cvijovic M, Dalsev B, Bilandzic E, Kemp GJ, Sonnerhagen P. Identification of putative regulatory upstream ORFs in the yeast genome using heuristics and evolutionary conservation. BMC Bioinformatics 2007, 8:205.

29. Tautz D. Polycistronic peptide coding genes in eukaryotes—how widespread are they? Brief Funct Genomic Proteomic 2009, 8(1):68-74.

30. Gu S, Anton A, Salim S, Blumberg KJ, DesSauver CW; Hexsiner SP. Alternative translation initiation of human regulators of G-protein signaling-2 yields a set of functionally distinct proteins. Mol Pharmacol 2008, 73(1):1-11.

31. Bazykin GA, Kochetov AV. Alternative translation start sites are conserved in eukaryotic genomes. Nucleic Acids Res 2010, 39(2):567-577.

32. Vogel C, Abreu Rde S, Ro D, Le SY, Shapiro BA, Burns SC, Sandhu D, Bouzit DR, Mascotte EM, Penahila LO. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol 2010, 6:400.

33. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, Hlavina W, Kapustin Y, Mercier P, Maglott D, Britte Z, Marques AC, Graveti T, Zhou S, Teague B, Potamouski K, Churas C, Place M, Herschibl J, Runinshem R, Forrest D, Amos-Landgraf J, Schwartz DC, Cheng Z, Lindblad-Toh K, Echeler EE et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 2009, 7(5):e1000112.

34. Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature 2004, 431(7011):931-945.

35. Karapetyan A, Keefe D, Smidley D, London D, Spooner W, Mellopp C, Hammond M, Rocca-Serra P, Cox T, Birney E; Ensmart: a generic system for fast and flexible access to biological data. Genome Res 2004, 14(1):160-169.

36. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res 2002, 12(6):996-1006.

Submit your next manuscript to BioMed Central and take full advantage of:

• Convenient online submission
• Thorough peer review
• No space constraints or color figure charges
• Immediate publication on acceptance
• Inclusion in PubMed, CAS, Scopus and Google Scholar
• Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit

doi:10.1186/1471-2148-11-192
Cite this article as: Chen et al.: Exploring the selective constraint on the sizes of insertions and deletions in 5' untranslated regions in mammals. BMC Evolutionary Biology 2011:1192.