Evolutionary formation of a human *de novo* open reading frame from a mouse non-coding DNA sequence via biased random mutations

Nicholas Delihas

https://orcid.org/0000-0002-1704-2587

Department of Microbiology and Immunology
Renaissance School of Medicine
Stony Brook University
Stony Brook, NY 11794 USA

Email: Nicholas.delihas@stonybrook.edu
Tel. # 001 631 286-9427

Keywords; Molecular evolution, *de novo* ORFs, biased mutations, evolutionary root, gene evolution, random mutations, purifying selection

Abstract

Two open reading frames (ORFs) of evolutionary interest stem from the human gene *SMIM45*. An investigation shows that one consists of an ultra-conserved 68 amino acid (aa) sequence that dates back to the amphibians, ~350 MYA. In contrast, the other, a 107 aa ORF, develops slowly during primate evolution. An early embryonic stage of development has been detected in the mouse genome that consists of a small sequence homologous to a segment of the human 107 aa ORF, the oligopeptide SGLE-VTVYGGGVQKGKT. This sequence appears to be the first recognizable segment of the ORF to appear evolutionarily. As the partial sequence is already present in the mouse, the evolutionary root of the human 107 aa is likely to be in a close ancestor of the mouse. The analysis presented shows that development of the 107 aa ORF proceeds via the biased selection of base pairs that are fixed during primate evolution. Random mutations also occur during development; proposed is the correction by purifying selection. The delineation of mutations occurring during development of the 107 aa ORF can provide a model for the evolutionary formation of protein genes.

Introduction

Gene *SMIM45* encodes a predicted 68 amino acid (aa) protein. This protein is termed a putative small integral membrane protein. Although there is a large body of computational information on *SMIM45*, a protein product has not been isolated and characterized experimentally, although there is evidence for transcription from the 68 aa locus (UniProt A0A590UK83 SMI45_HUMAN). *SMIM45* also encodes a 107 aa open reading frame (ORF), but here also, no protein product has been isolated from this locus. However, in an interesting study, An and coworkers (1) describe a protein that functions in human brain development and predict that it is the 107 aa open reading frame of *SMIM45*. Both the 68 aa and 107 aa ORFs stem from exon 2 of *SMIM45* (Ensembl; Gene: *SMIM45* (ENSG00000205704) and both represent direct reads from the DNA sequence without splicing. In the work presented here, the NIH Mammalian Gene Collection Program that identifies cDNA clones that contain complete ORFs for human and mouse genes (2) was crucial to the study.

In this paper, described is use of genomic DNA sequences and their translated open reading frames to analyze the evolutionary development of the *SMIM45* locus. Independent of experimental verification of proteins, the genomic DNA provides important information on ORFs and how particular DNA regions evolve over evolutionary time. The evolutionary development of the 107 aa ORF is described in this paper. It may serve as a model for *de novo* human protein gene formation, i.e., protein genes made from scratch.

De novo genes were originally defined as genes that are formed from a non-coding genomic region of an ancestral genome (3-6). However, the detection of an ancestral genomic DNA that has no coding capacity and forms the evolutionary root of a gene has not been reported. Finding the ancestral root is exceedingly difficult, as distinguishing root DNA base-pairs from random sequences is very problematic. However, it has been shown that a small segment of the yeast genomic locus, YBR196C-A is present in a yeast ancestral non-coding genomic region and evolved to form the putative gene *YBR196C-A* (7). Although not the evolutionary root, this appears to be an early developmental stage of the ORF. As there is no evidence that it is translated (7), YBR196C-A should tentatively be considered an ORF and not a gene.

In the work described here, the evolutionary development of the human *SMIM45* 107 aa ORF has been followed and found to stem from an ancestral genomic region of the mouse that is non-coding, but also found is a small nt sequence homologous to the human 107 aa mRNA. The translated aa sequence represents an early embryonic stage of development of the107aa. Together with the study on YBR196C-A, these are examples of finding early embryonic stages in the evolutionary development of ORFs.

Also reported is the mutational process that functions during molecular evolution. Proposed is the formation of the 107 aa ORF by biased random mutations, i.e., the biased selection of a bp from a mutation where the bp is fixed during evolution, however, random mutations also occur during polypeptide growth. Presented is an

analogy of the biased random mutational process to a biased random walk that occurs during chemotaxis.

Results

In the work described here, it was possible to follow the evolution of DNA sequences that lead to the formation of the human *SMIM45* 68 aa and the 107 aa open reading frames because of the extreme evolutionary conservation of the 68 aa sequence, which was used as a guidepost, as well as the genomic synteny displayed by this chromosomal region. *SMIM45* shows evolutionary synteny with two neighbor genes, *CENPM* and *SEPTIN3*, with *SMIM45* situated in between (Figure 1a).

There are two adjacent DNA sequences encoded in exon 2 of the gene *SMIM45*. These consist of open reading frames that are separated by a small intervening 323 bp sequence in human chromosome 22Q11 (Figure 1b); a translation of the intervening sequence shows that it is devoid of methionine and open reading frames.

To determine origins of *SMIM45* and how its sequence evolved, an analysis was performed on the evolutionary formation of the gene through its two ORFs, which constitute the translated sequences from the two segments of *SMIM45* exon 2 DNA. The 68 aa open reading frame displays a remarkable conservation over evolutionary time (Figure 2). There is only one amino acid residue difference between the mouse (*Mus musculus*) and human sequence; the aa sequence from the tarsier [Philippine tarsier (*Carlito syrichta*)], a lower primate, shows 100% identity with the human sequence. The Philippine tarsier primate evolutionarily appeared ~40 million years ago (MYA) and the mouse ~90 MYA. This degree of ultra-conservation during primate evolution points to an important function of this aa open reading frame. However, we have also been able trace the 68 aa sequence as far back as the amphibians (*Bombina bombina*, the fire-bellied toad), to an evolutionary age ~350 MYA. Figure 2 bottom shows an alignment of the toad sequence with that of the human 68aa and displays a 74% identity.

In sharp contrast to the 68 aa ORF, the 107 aa ORF evolved slowly over ~100 million years of evolutionary time. The evidence presented suggests that a sequence representing an early embryonic stage of the 107 aa sequence development is in a mouse genomic region that has sequence similarity to the human107 aa nt sequence. An alignment of a segment of the mouse *SMIM45* gene sequence, a segment of exon 2, and the human mRNA that encodes the 107 aa open reading frame shows a 57.5% nucleotide sequence identity (Figure 3), with one section, nt positions 7753-7810 of the mouse SMIM45 gene sequence displaying a 79.7% identity. The alignment locates the mouse genomic region that is homologous to the mRNA of the human 107aa.

A translation of the nucleotide (nt) sequence of human exon 2 displays the 68 aa and the 107 aa open reading frames; these are color highlighted in yellow (68 aa) and tan (107 aa) in Figure 4a. However, a translation of the mouse genomic nt sequence

homologous to the human 107 aa nt sequence encodes no proteins and has no open reading frames with the exception of two methionine start codons followed by small peptides (highlighted in light red, Figure 4b); these have no identity to the 107 aa ORF, and from Blast searches, do not form part of proteins in any organisms. A lack of a significant open reading frame indicates that the mouse homologous sequence is non-coding. However, found in the mouse genomic region is a nt sequence that translates to the aa sequence SGLE-VTVYGGGVQKGKT (Figure 4b. highlighted in turquoise), the dash in the sequence represents a stop codon. Figure 4c shows an alignment of segments of nt sequences from the mouse *SMIM45* gene with that of the human *SMIM*45 exon 2, including nt sequences that translate to the human aa sequence SGLELVRVCGGGMQRDKT and the mouse homologous aa sequence, SGLE-VTVYGGGVQKGKT. The nt sequence that encodes the aa segment in humans is 54 nt in length; there is an 81% identity between the human and the mouse homologous nt sequences. The probability of a random 54 nt sequence that has 81% identity with the human counterpart by chance, is exceedingly small, $1/(4^{54})xC(54,45)x3^{9}$. [C(54,45) is the binomial coefficient counting the ways to pick 45 positions out of 54 positions. The denominator is all sequences of 4 symbols of length 54, the 3^{9} represents ways to build sequences of non-matching nucleotides]. Thus, the mouse aa sequence, SGLE-VTVYGGGVQKGKT appears to represent a recognizable beginning of the formation of the human 107 aa open reading frame and an early embryonic stage in 107 aa open reading frame development.

The human sequence SGLELVRVCGGGMQRDKT is found to be highly conserved between primates (Figure 5). The lemur, a primitive and early evolutionary primate, formed an almost a complete sequence corresponding to the human sequence and shows 83% identity. However, the evolutionary pattern shows that sequences from several species deviate in similarity, the tree shrew, tarsier, and gorilla. This appears to be due to random mutations involving insertions, deletions, and point mutations in the genomic sequences that encode the mRNAs, e.g., the gorilla mRNA has three regions of deletions in the 3' half of the sequence. These mutations appear to be of doubtful significance as they are not present in progenies. Thus, this may constitute purifying selection. The random mutations present in the conserved sequence of the gorilla vs the lemur are consistent with the findings showing the gorilla harbors the largest mutation rate amongst the primates and the lemurs the least (8).

The pattern of development of the 107 aa ORF with evolutionary time was analyzed. This readily provided a picture of the growth of the 107 aa sequence. With the mouse sequence (Figure 6), alignment of the translated 5'3' Frame 3 aa sequence (from Figure 4b) displays the SGLE-VTVYGGGVQKGKT sequence but no other contiguous stretch of aa residues with an identity to the human 107 aa. There are individual amino acid residues that show similarities in the N-terminal half of the mouse sequence, e.g., $_{3}$CP, $_{11}$F but it is not possible to determine if these are root bps or show identity by chance.

With further evolutionary development of the 107 aa, the Philippine tarsier sequence shows that it has progressed from that of the mouse and displays a methionine start aa

residue with four contiguous aa residues of identity that are close to the start codon (Figure 6, 5'3' Frame 1). There is also the beginning of development of aa identity in the middle of the sequence and at the C-terminal end (5'3' Frame 3). The presence of several segments of identity in different regions of the tarsier sequence suggests that the 107 aa sequence begins by formation of identity in different genomic regions, thus implying sequence growth from different regions of the genome that may have originated from different and separate putative seed bps. Of interest, the human aa sequence, $_{73}$GGGMQRDKT$_{81}$, which is largely present in the mouse genome, is absent in the tarsier homolog. This is due to point mutations and the deletion of the mouse nt sequence GGTGTGCA in the tarsier (Figure 6, bottom). These represent random mutations specific to the tarsier, are not found in progeny, but are likely to be eliminated by purifying selection.

In analyzing a genomic segment of the mouse and lemur nt sequences that are homologous to that of the human mRNA sequence,13 mutations are evolutionarily fixed in the lemur/human, compared to 5 random mutations (highlighted in purple, Figure 6, bottom "Biased and random mutations"); these biased mutations may be fixed by cellular regulatory mechanisms. The lemur predates the tarsier by approximately 5 MY, but both are primitive primates and members of the prosimians. The rationale for choosing this particular genomic segment is that it shows a high similarity in nt sequence between species. Analysis in other regions of the lemur sequence where there is more sequence variability is difficult as some bp identities may be by chance. The alignment shows both fixed (biased) and random mutations that supports the concept of biased random mutations occurring during 107 aa ORF development.

The rhesus sequence shows a larger stretch of similarity at the amino terminal end compared to the tarsier (Figure 7). However, similar to the tarsier, there is significant development of sequence identity in other sections; the 5'3'Frame 2 shows sequence identities in the middle and carboxyl terminal half. Prominent is the large stretch of sequence identity around the Rhesus sequence, SGLELVRVCGGGVQKGKT. Perhaps this is because it is the first developed sequence to appear in the mouse and thus this provided an opportunity to grow faster. The elongation of tandem aa residues appears to be by biased nearest neighbor additions.

The orangutan sequence displays further development at the N-terminal half and does not display frame shift mutations (Figure 7). The aa sequence identity is 79.4%. However, a 13 bp insertion in the orangutan mRNA, which is not found in the rhesus, chimpanzee or human mRNAs, results in addition of the aa sequence $_{30}$SMAPAL$_{35}$ (Figure 7). This is due to a random mutation but, as with other random mutations, it is likely eliminated by purifying selection.

The chimpanzee sequence shows an almost complete human 107 aa ORF except for the lack of ten C-terminal amino acid residues present in the human counterpart (Figure 8). Of interest, this is due to a four bp tandem repeat AGAG that is missing in the

chimpanzee mRNA (Figure 8, bottom) and results in a frame shift, however, the 5'3' Frame 3 shows the presence of the ten C-terminal aa sequence. With the exception of three point mutations, the absence of the small tandem nt sequence repeat, AGAG distinguishes the chimpanzee 107 aa homolog from the human 107 aa. The tandem repeat may be due to polymerase stuttering that occurs at repeat sequences. Polymerase stuttering constitutes a random process. Thus, the chimpanzee and human aa sequences differ primarily because of a random mutation.

Discussion

By comparing genomic sequences and their translated aa sequences, the course of development of two ORFs during the evolution of primates was established.
What is striking is the sharp contrast in sequence conservation vs sequence development in the two *SMIM45* exon 2 DNA ORF sequences that are separated by a short intervening sequence. The DNA sequence of the mouse 68 aa ORF shows 18 bp that differ from the tarsier and human sequences, but there is only one bp substitution that results in a change in an aa residue; the other 17 mutations involve the wobble position. This implies an ancient regulatory element that protects against mutations that alter the ORF. Why the 68 aa open ORF is ultra-conserved over ~80 MY during primate evolution is an interesting question. It does have one protein conserved domain, PLN02960, but there is no experimental information on function. This degree of conservation is not found in exons of several protein genes, for example, exon 11 of the human *USP18* protein gene shows an identity of 64% with the mouse *USP18* exon 11 (9). The *USP18* gene functions in interferon regulation and is regulated by its 3'UTR exon 11 (10). There are examples of functions of ultra-conserved non-coding genomic sequences that consist of enhancer sequences (11, 12), but there is little information on ultra-conserved protein genes. The protection against bp changes may possibly indicate the importance of each aa residue of the polypeptide chain in maintaining polypeptide tertiary structure essential for function.

The analysis presented here also shows that the mouse sequence, $_{48}$SGLE-VTVYGGGVQKGKT$_{64}$ is the earliest recognizable segment of the human 107 aa sequence to appear in mammals and represents an early embryonic stage of development. But the evolutionary root is likely to be in an ancestor to the mouse, where perhaps individual base pairs or small oligo base pairs form the very first stage of development. An analysis of the *Gallus gallus* (chicken) *SMIM45* sequence (annotated as *LOC100859314*) yields the 68 aa sequence but shows random identity to the107 aa nt sequence in the putative genomic region. As the evolutionary age of the chicken is ~325 MYA, the evolutionary root is likely in a species that predates the chicken.

A model of the evolutionary development of the 107 aa sequence can be visualized as an initial formation of a mutation that results in a starter seed bp or bps that are to be evolutionarily fixed. The next stage is nearest neighbor sequence growth that occurs in several different genomic regions, which leads to the full human 107 aa sequence when

nearest neighbor positions fill in. There is a bias in seed base pair formation and in nearest neighbor sequence growth. Random mutations occur during the process of aa chain growth that result in frame shifts during the 107 aa development, which may be eliminated by purifying selection. Regulatory elements may dictate the initial seed bps, the specific nearest neighbor formation, and the protection of the growing sequence from disruptive mutations.

Mutational bias across the genome has been reported (13, 14) whereby certain genes crucial to survival of the organism display lower mutation rates relative to other genes (14). However, these results have been questioned (15). Nevertheless, the mutational bias described in this paper concerns mutations in an evolutionarily evolving macromolecule and not in an already formed gene, and most likely does not involve the survival of the species.

The evolutionary formation of a class of microproteins during primate evolution has been reported (16). This comprehensive study shows *de novo* gene formation in humans that appears to involve one base pair substitution resulting in the formation of the Met start codon. However, early embryonic stages were not determined. In another interesting study, the partial evolutionary progression of a number of protein genes in primates was also investigated (17). One example is the *MYEOV* (myeloma overexpressed) gene that shows one base pair change in the first codon creating a Met initiation at the N-terminal end, but also a frame shift that distinguish the chimpanzee sequence from that of the human sequence. Here also, data have not been presented for determining a possible early embryonic stage of this gene. However, the *MYEOV* mRNAs from various species show a very significant development of the *MYEOV* mRNA sequence in the rhesus macaque, suggesting a rapid sequence development from the primitive lower primates.

Biased random mutations can be considered analogous to a biased random walk, a concept used to describe the migration of bacterial cells, leucocytes and other cells in space during chemotaxis (18-22). In chemotaxis, the bias is movement towards a "goal", e.g., a high concentration of nutrient, but there are also random movements away from the nutrient. Biased mutations as presented here lead to the formation of the 107aa open reading frame, the "goal", but with random mutations that represent drift; these can be considered analogous to small movements in random directions in chemotaxis. In both cases, the randomness is corrected. The progression with time of ORF development is the variable analogous to movement in space in chemotaxis. If enough random mutations can be pinpointed for statistical significance, a mathematical model of a biased mutational process, analogous to a mathematical model of a biased random walk may be helpful in better understanding mutations associated with open reading frame development during evolution. A mathematical model for chemotaxis has been presented before to better access biased random walks (22).

But the missing link is the lack of definitive evidence that the human 107 aa open reading frame encodes a protein. An *et al* (1) however may have discovered a function

for the 107aa ORF and the gene may be expressed in certain cells and transiently. Further experimental work on protein isolations would verify this. But in support of the existence of a protein product are the biased mutations that occur during the human 107 aa ORF growth. As there is stringent regulation of mutations in the DNA sequence during evolution, as well as the correction of random mutations that spans ~45 MY, it is highly unlikely that that the 107 aa ORF is not translated and have a function.

Methods

Protocol for finding the evolutionary early embryonic stage of the 107 aa ORF.

1. The genomic region of the mouse *SMIM45* containing the 68 aa ORF displays synteny with the 107 ORF and serves as a guidepost for locating nt sequences homologous to the human 107 aa. By aligning the mouse *SMIM45* gene sequence with the sequences of human *SMIM* 45 exon 2 and the human 107 aa mRNA, the mouse *SMIM45* region that has homology to the human 107 aa mRNA was determined.
2. A segment of the mouse nt sequence that produces a stretch of high similarity to the 107 aa mRNA was used to translate the sequence by the Expasy translation tool. The resulting short aa sequences shown in the three 5'3' Frames provided the early embryonic developmental sequence.
3. Homologous 107 aa mRNAs from primate species were translated and the aa sequences compared to the human sequence.
4. For species that do not have the *SMIM45* gene annotated (e.g., the tarsier), the *SMIM45* sequence was determined by analyzing the DNA sequence between two genes that show synteny, *SEPTIN3* and *CENPM* and then nt sequences were aligned with the 68 aa sequence. This gave an accurate determination of the homologous 107 aa nt sequence in these species.

Species analyzed for 68 aa and 107 aa ORFs

Bombina bombina, the fire-bellied toad), *Gallus gallus* (chicken), *Mus musculus* (house mouse),*Tupaia chinensis* (Chinese tree shrew), *Lemur catta* (Ring-tailed lemur), *Carlito syrichta* (Philippine tarsier), *Macaca mulatta* (Rhesus monkey), *Macaca thibetana thibetana*, *Papio anubis* (olive baboon), *Pongo abelii* (Sumatran orangutan), *Gorilla gorilla gorilla* (western lowland gorilla), *Pan troglodytes* (chimpanzee), and *Homo sapiens* (human).

Source of gene sequences and species evolutionary age

SMIM45 gene sequences were obtained from the website: home gene NCBI, https:// (23).

The exon 2 sequence is from *Ensembl* (24).

The *SMIM45* protein data were from UniProtKB/Swiss-Prot (25) and the NLM/NCBI (23).

The evolutionary ages of primates and other species were from (26, 27).

Nucleotide and amino acid sequence alignment and translation tools

Clustal Omega was employed to align three or more nt or aa sequences. Both EMBOSS Needle and Clustal Omega were used for alignment of two sequences. In regions where the two sequences are highly similar, no significant differences were found. The data presented in Figure 6a,b,c show alignments by Clustal Omega. This was chosen over alignments by EMBOSS Needle as the by Clustal Omega format is far superior and readily allows for identification of aa positions of identity.

The ExPASy Translational tool (28) was used for the translation of nucleotide sequences.

The Sequence Manipulation Suite (29) was used to obtain reverse complement sequences and sequence clean up.

Acknowedgements

I thank Dr. Alan Tucker, Department of Applied Mathematics and Statistics, Stony Brook University for statistical analyses.

There was no funding for this work.

The author declares no conflict of interest.

Figure captions

Figure 1a. L Chromosomal locations of *SMIM45* in the human and mouse genomes. The synteny with genes *CENPM* and *SEPTIN3* is shown. Diagrams are from the NLM/NCBI website that describes genes from different species (Home-Gene-NCBI, https://www.ncbi.nlm.nih.gov/gene). **1b**. A schematic of exon 2 of *SMIM45* showing the positions of the 68 aa, intervening sequence and the 107 aa sequence.

Figure 2. Alignment of mouse 68aa open reading frame and the human 68 aa open reading frame (top), Philippine tarsier (*Carlito syrichta*) (middle) and fire-bellied toad (*Bombina bombina)* (bottom). The translated human and mouse *SMIM45* nt sequences are identical with the exception of one aa residue difference (position 57).

Figure 3. Alignment of a segment of the mouse *Smim45* gene sequence with the human sequence that encodes the 107 aa open reading frame, together with a segment of the nt sequence of exon 2 of the human *SMIM45* gene and a segment of the human *SMIM45* gene sequence (as provided by *Ensembl*). The mouse *Smim45* gene sequence is 8374 bp as annotated by the NCBI. The alignment was performed with the entire mouse *Smim45* but only a segment of the sequence is shown in Figure 3.

Figure 4. a. The translation of the nt sequence from human *SMIM45* exon 2. The 5'3' Frame 3 is shown**.** The yellow highlighted sequence is the 68 aa open reading frame and the tan highlighted is the 107 aa open reading frame. **4b.** The translation of the mouse genomic region that is predicted to be homologous to the human 107 aa nucleotide sequence of SMIM45 exon 2. The three 5'3' reading frames are shown. The 5'3' Frame 3 aa sequence displays the highest similarity to the human 107aa sequence and contains the SGLE-VTVYGGGVQKGKT sequence (highlighted in turquoise blue). The highlighted sequences in red are small open reading frames. The 5' 3' Frame 3 was used for the alignment of the mouse sequence to the human 107 aa sequence shown in Figure 6. **4c.** An alignment of nt sequences from the human and mouse SMIM45 gene together with the nt sequences that translate to the human aa sequence SGLELVRVCGGGMQRDKT and the mouse homologous aa sequence, SGLE-VTVYGGGVQKGKT.

Figure 5. The evolutionary progression of the SGLELVRVCGGGMQRDKT sequence. Shown the aa sequences, the approximate evolutionary age of the species, and the % identity relative to the human species.

Figure 6. Development of the *SMIM45* 107aa ORF in the trarsier. The source of the mouse aa sequences is from reading frames of translated genomic nt sequences homologous to the 107aa mRNA sequence. The dash in the sequence $_{48}$SGLE-VTVYGGGVQKGKT$_{64}$ refers to a stop codon. The alignment of the human 107 aa sequence with the mouse 5'3' Frame 3 aa (obtained from a translation of the mouse nt sequence t homologous to the 107 aa mRNA), and, alignment of 5'3' Frames 1 and 3 of the translated tarsier sequence. **Figure 6. bottom**, "Deletion in tarsier 107naa sequence". Nt sequence alignment show the deletion of nine bps in the tarsier nt 107 aa sequence, and, **Figure 6. Bottom**, "Biased and random mutations" shows mutations between the mouse, lemur and human sequences.

Figure 7. Amino acid alignments of the rhesus homologous 107 aa sequence, the 5'3' Frames 1 and 2 with that of the human 107 aa. Bottom, the homologous 107 aa orangutan sequence with the human 107 aa.

Figure 8. Alignment of the evolving 107 aa sequence in the chimpanzee with that of the human 107 aa. Displayed are 5'3' Frames 1 and 3 of the chimpanzee homologous

sequence. Bottom. An Alignment of segments of the chimpanzee *SMIM45* sequence, human *SMIM45* exon 2, the human *SMIM45* sequence from Ensembl, and the human 107 aa mRNA sequence. The alignment shows the absence of the 4 bp AGAG tandem repeat in the chimpanzee sequence.

References

1. An Ni A. et al. De novo genes with an lncRNA origin encode unique human brain developmental functionality. Nat Ecol Evol.2023 (2):264-278. doi: 10.1038/s41559-022-01925-6. PMID: 36593289

2. Strausberg RL et al; Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Mammalian Gene Collection Program Team. Proc Natl Acad Sci U S A. 2002 Dec 24;99(26):16899-903. doi: 10.1073/pnas.242603899. .PMID: 12477932

3 Long M. et al. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 2003 Nov;4(11):865-75. doi: 10.1038/nrg1204.PMID: 14634634

4. Van Oss SB. et.al De novo gene birth. PLoS Genet. 2019 May 23;15(5):e1008160. doi: 10.1371/journal.pgen.1008160. .PMID: 31120894). eCollection 2019 May.

5. Schlötterer C. Genes from scratch--the evolutionary fate of de novo genes. Trends Genet. 2015 Apr;31(4):215-9. doi: 10.1016/j.tig.2015.02.007. PMID: 25773713

6 Broeils LA. et al. Evolution and implications of de novo genes in humans. Nat Ecol Evol. 2023. PMID: 36928843 Review.

7.Vakirlis et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences..Nat Commun. 2020. doi: 10.1038/s41467-020-14500-z PMID: 32034123

8. Kuderna LFK. et al. A global catalog of whole-genome diversity from 233 primate species. Science. 2023. 906-913. doi: 10.1126/science.abn7829. Epub 2023 Jun 1. PMID: 37262161

9. Delihas N. Genesis of Non-Coding RNA Genes in Human Chromosome 22-A Sequence Connection with Protein Genes Separated by Evolutionary Time. Noncoding RNA. 2020;6(3):36. doi: 10.3390/ncrna6030036.PMID: 32899105

10 Rubino E, et al, Human Ubiquitin-Specific Peptidase 18 Is Regulated by microRNAs via the 3'Untranslated Region, A Sequence Duplicated in Long Intergenic Non-coding RNA Genes Residing in chr22q11.21. Front Genet. 2021 Feb 3;11:627007. doi: 10.3389/fgene.2020.627007. eCollection 2020. PMID: 33633774

11 Dickel DE. et al. Ultraconserved Enhancers Are Required for Normal Development. Cell. 2018. 172(3):491-499.e15. doi: 10.1016/j.cell.2017.12.017. PMID: 29358049

12 Hecker N, and Hiller M. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. Gigascience. 2020 Jan 1;9(1):giz159. doi: 10.1093/gigascience/giz159.PMID: 31899510

13. Martincorena I, Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature. 2012. 485(7396):95-8. doi: 10.1038/nature10995. PMID: 22522932

14 Monroe JG,. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature. 2022. 602(7895):101-105. doi: 10.1038/s41586-021-04269-6. Epub 2022 Jan 12.PMID: 35022609

15 Liu H. and Zhang J.. Is the Mutation Rate Lower in Genomic Regions of Stronger Selective Constraints? Mol Biol Evol. 2022. 39(8):msac169. doi: 10.1093/molbev/msac169. PMID: 3590724

16 Vakirlis N, De novo birth of functional microproteins in the human lineage. Cell Rep. 2022. 41(12):111808. doi: 10.1016/j.celrep.2022.111808. PMID: 36543139 Free PMC article.

17 Broeils LA, et al. Evolution and implications of de novo genes in humans. . Nat Ecol Evol. 2023. 7(6):804-815. doi: 10.1038/s41559-023-02014-y.

18. Macnab RM. and Koshland DE Jr. The gradient-sensing mechanism in bacterial chemotaxis. Proc Natl Acad Sci U S A. 1972. 69(9):2509-12. doi: 10.1073/pnas.69.9.2509. PMID: 4560688

19. Koshland DE Jr.Bacterial chemotaxis in relation to neurobiology. Annu Rev Neurosci. 1980. PMID: 6774654

20. Codling EA, et al. Random walk models in biology. J R Soc Interface. 2008. PMID: 18426776.

21. Sourjik V, et al Responding to chemical gradients: bacterial chemotaxis. Curr Opin Cell Biol. 2012. PMID: 22169400

22. Liu R, Hirn M, and Krishnan A. Accurately modeling biased random walks on weighted networks using node2vec. Bioinformatics. 2023. 39(1):btad047. doi: 10.1093/bioinformatics/btad047. PMID: 36688699

23. Sayers EW. et al. GenBank., Nucleic Acids Res. 2022. 50(D1):D161-D164. doi: 10.1093/nar/gkab1135.PMID: 34850943

24. Cunningham, F Ensembl 2022. *Nucleic Acids Res.* 2022, vol. 50(1):D988-D995 PubMed PMID: 34791404. doi:10.1093/nar/gkab1049

25 UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023. 51(D1):D523-D531. doi:10.1093/nar/gkac1052.PMID: 36408920

26. Zhang YE, Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol. 2010. 8(10):e1000494. doi: 10.1371/journal.pbio.1000494.PMID: 20957185

27. Fan Y. et al. Genome of the Chinese tree shrew. Nat Commun. 2013;4:1426. doi: 10.1038/ncomms2416.PMID: 23385571

28. Gasteiger E, et al. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 2003. 31(13):3784-8. doi:10.1093/nar/gkg563.PMID: 12824418

29. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. .Biotechniques. 2000. 28(6):1102, 1104. doi: 10.2144/00286ir01.PMID

Figure 1

a

SMIM45 small integral membrane protein 45 *Homo sapiens* (human)

Chromosome 22 - NC_000022.11

Smim45 small integral membrane protein 45 [*Mus musculus* (house mouse)

Chromosome 15 - NC_000081.7

b

SMIM45 Exon 2

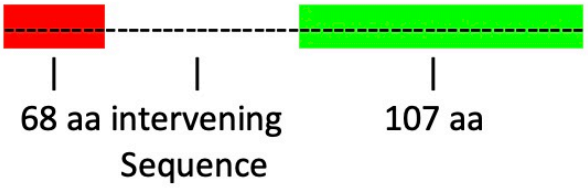

Figure 2

Mouse

```
CLUSTAL O(1.2.4) multiple sequence alignment


mouse.68aa      MPHFLDWFVPVYLVISVLILVGFGACIYYFEPGLQEAHKWRMQRPLVDRDLRKTLMMRDN 60
human.68aa      MPHFLDWFVPVYLVISVLILVGFGACIYYFEPGLQEAHKWRMQRPLVDRDLRKTLMVRDN 60
                ************************************************************:***

mouse.68aa      LAFGGPEV  68
human.68aa      LAFGGPEV  68
                ********


#  Percent Identity  Matrix - created by Clustal2.1

     1: mouse.68aa   98.53
     2: human.68aa  100.00
```

==========

Tarsier

```
CLUSTAL O(1.2.4) multiple sequence alignment


Tarsier.68aa      MPHFLDWFVPVYLVISVLILVGFGACIYYFEPGLQEAHKWRMQRPLVDRDLRKTLMVRDN    60
human.68aa        MPHFLDWFVPVYLVISVLILVGFGACIYYFEPGLQEAHKWRMQRPLVDRDLRKTLMVRDN    60
                  ************************************************************

Tarsier.68aa      LAFGGPEV     68
human.68aa        LAFGGPEV     68
                  ********


#  Percent Identity  Matrix - created by Clustal2.1

     1: Tarsier.68aa  100.00
     2: human.68aa    100.00
```

==========

Toad

```
CLUSTAL O(1.2.4) multiple sequence alignment


toad.68aa.        MTHFLDWFVPCYLILSVLILVGFGACIYYFEPGLKEAHKWRAQRPLLDSDIQKSLLMRDN     60
human.68.aa       MPHFLDWFVPVYLVISVLILVGFGACIYYFEPGLQEAHKWRMQRPLVDRDLRKTLMVRDN     60
                  * ******** **::*******************:****** ****:* *::*:*::***

toad.68aa.        MAYSLPDV 68
human.68.aa       LAFGGPEV 68
                  :*:. *:*


#  Percent Identity  Matrix - created by Clustal2.1

     1: toad.68aa.     73.53
     2: 68.aa.human   100.00
```

Figure 3

```
SMIM45.mouse              cctttgtgctgtccttccctctaccactctacctggagtcaggtctgactgtcccactga    7619.
human.seq.encode.107aa    ---------------------------------------------atgtctatggctgcct    16
exon2.SMIM45.human        CCTTTGCGCTGCGCAGCACTCTGCCCTC--CACTGGACTCAGGCATGTCTATGGCTGCCT    561
SMIM45gene..Ensembl       CCTTTGCGCTGCGCAGCACTCTGCCCTC--CACTGGACTCAGGCATGTCTATGGCTGCCT    6683
                                                                        ** ** *  *

SMIM45.mouse              attctgagttgc------------------tttctgttcccatctctgttgc-acagctg    7660
human.seq.encode.107aa    gtcctgaggctccggagccctcatttcttcgtgaagtccccagctcccctgcctccactc    76
exon2.SMIM45.human        GTCCTGAGGCTCCGGAGCCCTCATTTCTTCGTGAAGTCCCCAGCTCCCCTGCCTCCACTC    621
SMIM45gene..Ensembl       GTCCTGAGGCTCCGGAGCCCTCATTTCTTCGTGAAGTCCCCAGCTCCCCTGCCTCCACTC    6743
                           * *****   *                  *   ** **** ***   ***  *  **

SMIM45.mouse              agctggacaggcctcaa-----------------aactgtcccaagaca--------aga    7695
human.seq.encode.107aa    aatggcaccggccctgcaactttaggcaggtcgaagccaacccaaggaaagaacctaaga    136
exon2.SMIM45.human        AATGGCACCGGCCCTGCAACTTTAGGCAGGTCGAAGCCAACCCAAGGAAAGAACCTAAGA    681
SMIM45gene..Ensembl       AATGGCACCGGCCCTGCAACTTTAGGCAGGTCGAAGCCAACCCAAGGAAAGAACCTAAGA    6803
                          *   * ** ****                     * *   ******  *       ***

SMIM45.mouse              atctaattcagaaggatcctggcttggtatagatcagccccgctcacctctgagaggtca    7755
human.seq.encode.107aa    acctcgtttggagggatgtcagcttgggccag----cccagccgcaccccgcggggctca    192
exon2.SMIM45.human        ACCTCGTTTGGAGGGATGTCAGCTTGGGCCAG----ACCAGCCGCACCCCGCGGGGCTCA    737
SMIM45gene..Ensembl       ACCTCGTTTGGAGGGATGTCAGCTTGGGCCAG----ACCAGCCGCACCCCGCGGGGCTCA    6859
                          * **  **  ** ****    ******   **     **   * **** *   * * ***

SMIM45.mouse              ggcttagaatgagtgacagtgtatggtgggggtgtgcagaagggtaagactttggggagg    7815
human.seq.encode.107aa    ggcttggaactggtgagggtgtgtggtgggggtatgcagagggataagaccgtggtagag    252
exon2.SMIM45.human        GGCTTGGAACTGGTGAGGGTGTGTGGTGGGGGTATGCAGAGGGATAAGACCGTGGTAGAG    797
SMIM45gene..Ensembl       GGCTTGGAACTGGTGAGGGTGTGTGGTGGGGGTATGCAGAGGGATAAGACCGTGGTAGAG    6919
                          ***** ***   ****  **** ********** ****** ** ******  ***    *

SMIM45.mouse              gtgtgcctgtatgtgtgcacacatgtgagggggagggagggaagggggaggaagagag    7875
human.seq.encode.107aa    gagaggttggtgaggagagagagagagagagagagag------------------agtc    294
exon2.SMIM45.human        GAGAGGGTTGGTGAGGAGAGAGAGAGAGAGAGAGAG------------------AGTC    839
SMIM45gene..Ensembl       GAGAGGGTTGGTGAGGAGAGAGAGAGAGAGAGAGAG------------------AGTC    6961
                          * * *  *   ** *   * * *     * * *** *                 **

SMIM45.mouse              ggagggaaggagggatagaaggagaagaagaggaggaggaggaggaagagaaggagaaag    7935
human.seq.encode.107aa    tgggggagc----------gggcaagcatggggagatga--------------------    324
exon2.SMIM45.human        TGGGGGGAGC----------GGGCAAGCATGGGGAGATGAGATGTGTA------------    877
SMIM45gene..Ensembl       TGGGGGGAGC----------GGGCAAGCATGGGGAGATGAGATGTGTA------------    6999
                           * *** **          **  *** *  **  ** **
```

```
Percent Identity  Matrix - created by Clustal2.1

     1: SMIM45.mouse                57.50
     2: human.seq.encode.107aa     100.00
     3: exon2.SMIM45.human          99.69
     4: SMIM45gene..Ensembl         99.69
```

Figure 4

a.

5'3' Frame 3

KAAAMPHFLDWFVPVYLVISVLILVGFGACIYYFEPGLQEAHKWRMQRPLVDRDLRKTLMVRDNLAFGGPEV-
ADLQRG-AGGTGRPPPARPARPARRPPRQREAGLLALHAIPLQPGPSRQRPQVRWQWRWGT-
VPGWPRACSPCQADSPRAHPPKLRGILRWAPVPLRCAALCPPLDSGMSMAACPEAPEPSFLREVPSSPASTQWHR
PCNFRQVEANPRKEPKNLVWRDVSLGQTSRTPRGSGLELVRVCGGGMQRDKTVVEERVGEERERERERVWGERAS
M G R - D V Y M - E R V W G P Q A G Q E V V E T G - T P W A V -
GLSIVGPNPLPLLEFPSPSPSPRLWQQAGAPALPPCPVPHFWGQTQPP-ARPCQTPSPP-
NPPPVWRLCCPTLDTDWQPASPRPPSLGLPS-FILNNKDFMMISA

b.

5'3' Frame 1

LTVPLNSELLSVPISVAQLSWTGLKTVPRQESNSEGSWLGIDQPRSPLRGQA-NE-
QCMVGVCRRVRLWGGCACMCAHM-GGREGRGEEEREGRRDRRRRRGGG

5'3' Frame 2

-LSH-ILSCFLFPSLLHS-AGQASKLSQDKNLIQKDPGLV-ISPAHL-EVRLRMSDSVWWGCAEG-
DFGEGVPVCVHTCEGGGREGGRKRGREGGIEGEEEEE

5'3' Frame 3

DCPTEF-VAFCSHLCCTAELDRPQNCPKTRI-FRRILAWYRSAPLTSERSGLE-
VTVYGGGVQKGKTLGRVCLYVCTHVRGEGGKGGGREGGKEG-KEKKRRR

c.

```
exon2.SMIM45.human          GTCAGCTTGGGCCAGACCAGCCGC----ACCCCGCGGGGCTCAGGCTTGG
human.seq.tansl.to.64SGLELVRVC  ----------------------------------------TCAGGCTTGG
SMIM45.mouse                CCTGGCTTGGTATAGATCAGCCCCGCTCACCTCTGAGAGGTCAGGCTTAG
mouse.seq.transl.to.SGLE-VTVYG  ----------------------------------------TCAGGCTTAG
                                                                        ******** *

exon2.SMIM45.human          AACTGGTGAGGGTGTGTGGTGGGGGTATGCAGAGGGATAAGACCGTGGTA
human.seq.tansl.to.64SGLELVRVC  AACTGGTGAGGGTGTGTGGTGGGGGTATGCAGAGGGATAAGACC------
SMIM45.mouse                AATGAGTGACAGTGTATGGTGGGGGTGTGCAGAAGGGTAAGACTTTGGGG
mouse.seq.transl.to.SGLE-VTVYG  AATGAGTGACAGTGTATGGTGGGGGTGTGCAGAAGGGTAAGACT------
                            **   ****  **** ********** ****** ** ******
```

Figure 5

		% identity	evolutionary age (approximate)
SGLE-VTVYGGGVQKGKT	**mouse.**	70.59.	90 MYA

N terminal----------C terminal

GLELLRAYVGGVQRGCT	tree shrew.	58.82.	80 MYA
SGLELVRVCGGGAQRGET	lemur.	83.33.	45 MYA
SGLEIMRLCSRDRADGS	tarsier.	35.29	40 MYA
GLELVRVCGGGVQKGKT.	rhesus	83.33.	27 MYA
SGLELVRVCGGGMQRDKT	**Tibetan macaque.**	100.00	27 MYA
SGLELVRVCGGGMQRDKT	baboon.	100.00	24 MYA
SGLELVRVCGGGMQRGKT	orangutan	100.00.	10 MYA
SGLELVRVWWGYAEG-DR	gorilla	52.9	8 MYA
SGLELVRVCGGGMQRDKT	chimpanzee	100.00.	6 MYA

N terminal----------C terminal

SGLELVRVCGGGMQRDKT	**human**	100.00	0 MYA -

.

Figure 6

Mouse (evolutionary appearance ~90 MYA)

```
CLUSTAL O(1.2.4) multiple sequence alignment

human.107aa                     MSMAACPEAPEPSFLREVPSSPASTQWHRPCNFRQVEANPRKEPKNLVWRDVSLGQPSRT    60
mouse.homolog.to.107aa.         ---SDCPTEF-VAFCSHLCC---TAELDRPQNCPKTR---I-FRRILAWYR----SAPLT    45
                                   : **     :*  .: .   ::: .** *  :..       : *.*      .   *

human.107aa                     PRGSGLELVRVCGGGMQRDKTVVEERVGEERERERERVWGERASMGR  107
mouse.homolog.to.107aa.         SERSGLE-VTVYGGGVQKGKTLGRVCLY------------------  72
                                 . **** * * ***:*:.**: .  :
```

==========

Tarsier (evolutionary appearance ~40 MYA

5'3'FRAME 1

```
CLUSTAL O(1.2.4) multiple sequence alignment

human.107aa                     MSMAACPEAPEPSFLREVPSSPASTQWHR--PC--------NFRQVEANPRKEPKNLVWR  50
tarsier.homolog.to.107aa.5'3'Frame1  MAIPACPEPLDVPFPPPTEFPGTLSSRLHPVGRCQS-PKERAEFGGMSG--WSRPAPPLWA  57
                                *:: ****  :  *      .  *:: *    *  *     :*  :..   ..*   :*

human.107aa                     ----DVS--LGQPSRTPRGSGLELVRVCGGGMQRDKTVVEERVGEERERERERVWGERAS  104
tarsier.homolog.to.107aa.5'3'Frame1  QAWK--GCAAGTVQTVP-GR---------GGAGRRESVRES--GES-----GQA-GG---  94
                                      .   *  . .* *          **  * ::* *.  **.      :. *

human.107aa                     MGR  107
tarsier.homolog.to.107aa.5'3'Frame1  ---  94
```

==========

5'3FRAME 3

```
CLUSTAL O(1.2.4) multiple sequence alignment

human.107aa                     -----MSMAACPEAPEPS-FLREVPSSPASTQWHRPCNFRQVEANPRKEPKNLVWRDVSL  54
tarsier.homolog.to.107aa.5'3'Frame3  GNPGLS-APGCSISPH-IPRYTQLPSPPS------------GQVSKLT-GKGRVWRDVRL  45
                                         .*  :*.      ::** *:             :..  .  *. ***** *

human.107aa                     GQPSRTPRGSGLELVRVCGGGMQRDKTVVEERVGEERERERERVWGERASMGR    107
tarsier.homolog.to.107aa.5'3'Frame3  VPPSPSPLGSGLEIMRLCSRDRADGS--IGKG-RGREKRECERVWGEWTGVGR    95
                                  ** :* *****::*:*. .   ..  : :    ..:** ****** :.:**
```

==========

Deletion in tarsier 107aa mRNA sequence

```
mouse.homolog.to.107aa.mRNA                 cacctctgagaggtcaggcttagaatgagtgacagtgtatggtgggggtgtgcagaaggg 222
tarsier.intergenic.SEPTIN3-CENPM            ctcccctctgggctcaggcttggaaataatgaggttgtgcagcagg--------gaccgt 11235
human.SMIM45.Ensembl                        CACCCCGCGGGGCTCAGGCTTGGAACTGGTGAGGGTGTGTGGTGGGGGTATGCAGAGGGA 6903
human.exon2.SMIM45.                         CACCCCGCGGGGCTCAGGCTTGGAACTGGTGAGGGTGTGTGGTGGGGGTATGCAGAGGGA 781
human.107aa.mRNA                            cacccccgcgggctcaggcttggaactggtgagggtgtgtggtgggggtatgcagaggga 236
                                            * ** *   * * ******** ***    ***   ***   *  **        **  *
```

==========

Biased and random mutations

```
mouse.homolog.to.107a.mRNA.7578-7906   ATAGATCAGCCCCGCTCACCTCTGAGAGGTCAGGCTTAGAATGAGTGACAGTGTATGGTG 206
mRNA.lemur.homolog.7596-7937           CTAGCCCAG----CCCCGCCCCTCTGGGCTCAGGCTTGGAACTGGTGAGGGTATGTGGTG 233
human.107aa.mRNA                       CCAGCCCAG----CCGCACCCCGCGGGGCTCAGGCTTGGAACTGGTGAGGGTGTGTGGTG 220
                                         **  ***     * * ** *   * * ******** ***   ****  ** * *****
```

==========

Figure 7

Rhesus (evolutionary appearance ~25 MYA)

5'3'FRAME 1

```
CLUSTAL O(1.2.4) multiple sequence alignment


human.107aa                         MSMAACPEAPEPSFLREVPSSPASTQ----WHRPCNF-RQVEANPRKEPKNLVWRDVSLG  55
Rhesus.Frame1.homolog.of.107aa      MATAACPEPPEPSFLREILFPHSAPPPPLSGTGPAASGGRRQRQPKERT-EPRLEGCQLG  59
                                    *: ***** ********:    ::         *.    : : :*::.  :   .. .**

human.107aa                         QTSRTPRG----SGLELVRVCGGGMQRDKTVVEERVGEERERERERERVWGERASMGR 107
Rhesus.Frame1.homolog.of.107aa      PAQPHPAGLRFGTGEGVWWGCAEG-DRGRGLGR---------------GERE---- 95
                                     :.  * *    :*  :   *. * :*.: : .               ***
```

5'3'FRAME 2

```
CLUSTAL O(1.2.4) multiple sequence alignment


human.107aa                         -MSMAACPEAPEPSFLREVPSSPASTQW-----HRPCNFRQVEANPRKEPKNLVWRDVSL  54
Rhesus.Frame2.homolog.of.107aa      WLRL-PVLSLPSPHFFVKSCFPTQLPRLHSVAPALQLQAAGVKGNPRKEPKNLVWRDVSL  59
                                     : :    . *.* *: :        :          :   *:.****************

human.107aa                         GQTSRTPRGSGLELVRVCGGGMQRDKTVVEERVGEERERERERERVWGERASMGR   107
Rhesus.Frame2.homolog.of.107aa      GQPSRTPQGSGLELVRVCGGGVQKGKTEVEGWG-----GEREKVWGEGESLGR   107
                                    ** ****:*************:*:.** **         ***:****  *:**
```

==========

Orangutan (evolutionary appearance ~10 MYA)

5'3'FRAME 1

```
CLUSTAL O(1.2.4) multiple sequence alignment


SMIM45.aa.human.107aa           MSMAACPEAPEPSFLREVPSSPASTQWHR------PCNFRQVEANPRKEPKNLVWRDVSL    54
partial.107aa.seq.orangutan     MSMAACPEAPEPSFLREIPGFPTQLPCLHSMAPALQL-AAGVEASPRKEPKNLVWRDVSL    59
                                *****************:*. *:.    :            ***.***************

SMIM45.aa.human.107aa           GQTSRTPRGSGLELVRVCGGGMQRDKTVVEERVGEERERERERERVWGERASMGR 107
partial.107aa.seq.orangutan     GQPSRTPQGSGLELVRVCGGGMQRGKTVVGERVGEE----RERVWGERASTGR 108
                                ** ****:****************.**** ******    ********** **
```

Figure 8

Chimpanzee (evolutionary appearance ~6 MYA)

5'3'FRAME 1

```
CLUSTAL O(1.2.4) multiple sequence alignment

human.107aa                          MSMAACPEAPEPSFLREVPSSPASTQWHRPCNFRQVEANPRKEPKNLVWRDVSLGQPSRT   60
chimp.homolog.to.107aa.5'3'Frame3    MSMAACPEAPEPSFLREVPSSPASTQWHQPCSFRQVEANPRKEPKNLVWRDVSLGQPSRT   60
                                     ****************************:**.****************************

human.107aa                          PRGSGLELVRVCGGGMQRDKTVVEERVGEERERERER-----VWGERASMGR  107
chimp.homolog.to.107aa.5'3'Frame3    PRSSGLELVRVCGGGMQRDKTVVEERVGEERERERESGGSGQAWGD------  106
                                     **.********************************      .**:
```

5'3'FRAME 3

```
CLUSTAL O(1.2.4) multiple sequence alignment

human.107aa                          MSMAACPEAPEPSFLRE-VPSSPASTQWHRPCNFRQVEANPRKEPKNLVWRDVSLGQPSR   59
chimp.himolog.to.107aa.5'3'Frame2    -NVYGCLS-GSRALISS-SPQLPCLHSMAPALQL-AGRSQPKERT-EPRLEGCQLGPAQP   55
                                      .: .* .  . ::: .* *. *.  .     ::   .::*::.  :   .. .**  .

human.107aa                          TP----RGSGLELVRVCGGGMQRDKTVVEERVGEERERERERVWGERASMGR  107
chimp.himolog.to.107aa.5'3'Frame2    HPAELRLGTGEGVWWGYAEG-DR-----GRGEGWGGERERERVWGERASMGR  101
                                      *     *:*  :    . * :*      .  *   ****************
```

==========

Alignment of chimpanzee nt sequence with the human 107aa mRNA

```
SMIM45.chimp                                          gggtatgcagagggataagaccgtggtagaggagagggttggggaggagagagag----a 6409
exon2.SMIM45.human                                    GGGTATGCAGAGGGATAAGACCGTGGTAGAGGAGAGGGTTGGTGAGGAGAGAGAGAGAGA 826
chromosome:GRCh38:22:41951550:41959539:1.SMIM456.Ensembl GGGTATGCAGAGGGATAAGACCGTGGTAGAGGAGAGGGTTGGTGAGGAGAGAGAGAGAGA 6948
human.107aa.mRNA                                      gggtatgcagagggataagaccgtggtagaggagagggttggtgaggagagagagagaga 281
                                                      ***************************************** ************    *

SMIM45.chimp                                          gagagagagagtctgggggggagcgggcaagcatggggagatgagatgtgtatatgtgaga 6469
exon2.SMIM45.human                                    GAGAGAGAGAGTCTGGGGGGAGCGGGCAAGCATGGGGAGATGAGATGTGTATATGTGAGA 886
chromosome:GRCh38:22:41951550:41959539:1.SMIM456.Ensembl GAGAGAGAGAGTCTGGGGGGAGCGGGCAAGCATGGGGAGATGAGATGTGTATATGTGAGA 7008
human.107aa.mRNA                                      gagagagagagtctgggggggagcgggcaagcatggggagatga
```