SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling

Yanyan Li[1,2,#], Honghong Zhou[2,#], Xiaomin Chen[2,3,#], Yu Zheng[1,2], Quan Kang[2], Di Hao[2], Lili Zhang[2,3], Tingrui Song[2], Huaxia Luo[2], Yajing Hao[4], Yiwen Chen[5], Runsheng Chen[2,3,6,*], Peng Zhang[2,*], Shunmin He[1,2,*]

[1] College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[2] Key Laboratory of RNA Biology, Center for Big Data Research in Health, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

[3] University of Chinese Academy of Sciences, Beijing 100049, China

[4] Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093, USA

[5] Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[6] Guangdong Geneway Decoding Bio-Tech Co. Ltd, Foshan, 528316, China.

# Equal contribution.

* Corresponding authors.

Email: heshunmin@ibp.ac.cn (He S), zhangp@ibp.ac.cn (Zhang P), crs@ibp.ac.cn (Chen R)

Abstract

Small proteins specifically refer to proteins consisting of less than 100 amino acids translated from small open reading frames (sORFs), which were usually missed in previous genome annotation. The significance of small proteins has been revealed in current years, along with the discovery of their diverse functions. However, systematic annotation of small proteins is still insufficient. SmProt was specially developed to provide valuable information on small proteins for scientific community. Here we present the update of SmProt, which emphasizes reliability of translated sORFs, genetic variants in translated sORFs, disease-specific sORFs translation events or sequences, and significantly increased data volume. More components such as non-AUG translation initiation, function, and new sources are also included. SmProt incorporated 638,958 unique small proteins curated from 3,165,229 primary records, which were computationally predicted from 419 ribosome profiling (Ribo-seq) datasets and collected from the literature and other sources originating from 370 cell lines or tissues in 8 species (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Danio rerio*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Escherichia coli*). In addition, small protein families identified from human microbiomes were collected. All datasets in SmProt are free to access, and available for browse, search, and bulk downloads at http://bigdata.ibp.ac.cn/SmProt/.

KEYWORDS: Ribosome profiling; Small ORF; Upstream ORF; Variants; Disease

Introduction

Genome annotation is fundamental to life science. In recent years, it has been found that small open reading frames (sORFs) widely exist in genomes of many organisms including human [1] and human microbiomes [2], and some are able to be translated into small proteins [3–5]. Small proteins are proteins with less than 100 amino acids, which may derive from untranslated regions (UTRs) of mRNAs [6] or non-coding RNAs [7,8] including pri-miRNAs [9,10], lncRNAs [11], and circRNAs [12]. Small proteins were usually missed in previous coding sequence annotation, while their significance has been revealed in current years for diverse functions [13], such as embryonic development [14,15], cell apoptosis [16], muscle contraction [17], and antimicrobial activity [18]. Some are proved to play roles in many diseases [19,20] including tumors [9,11,12]. Despite the abundance of sORFs in genome, the number of well-studied small proteins is very limited. Annotation of numerous small proteins will contribute to studies on various physiology and pathology processes.

Identification of small proteins at proteomic level is challenging. Mass spectrum (MS) can provide direct evidence of small proteins, but it relies much on the coverage of existing libraries, which mainly focused on large proteins rather than small proteins. Protease cleavage sites were lacking in small proteins limited by length. Besides, small proteins are usually of low abundance, and tend to be filtered out during enrichment process [21]. Ribosome profiling (also named Ribosomal footprinting or Ribo-Seq) provides a more sensitive way for global detection of translation events based on the deep sequencing of ribosome-protected mRNA fragments (RPFs) [22,23], which allows for identifying the location of translated ORFs and translation initiation sites, the distribution of ribosomes on mRNA, and the speed of translating ribosomes [24]. Reference library for mass spectrometry can also be constructed with Ribo-Seq results. The regular Ribo-seq (rRibo-seq) utilizes cycloheximide (CHX) [25], a drug binding at the ribosome E-site [26], as a translation elongation inhibitor to freeze translating ribosomes. Translation is principally regulated at the initiation stage. TI-seq is a variation of rRibo-seq technique that use different translation inhibitors, usually lactomidomycin (LTM) [25] or harringtonine (HAR) [27], which can induce ribosomes stasis at translation initiation (TI) sites (TISs). TI-seq enables the global mapping of translation initiation sites, and is more accurate in prediction of non-AUG start codons. Many sORFs are proved to use non-classical AUG start codon [28], which is also an important mechanism for generating protein isoforms [29,30]. rRibo-seq data usually have clear triplet periodicity [26]. Different computational analysis strategies [31–38] have been developed to identify translated sequences using Ribo-seq data.

Emerging evidences showed that many upstream open reading frames (uORFs) act in cis to regulate the translation of downstream ORFs by leaky scanning [39], reinitiation [40], and ribosome stalling [41]. Recently, variants creating new upstream start codons or disrupting stop sites of existing uORFs (uORF-perturbing) were found under strong negative selection [42]. uORF-perturbing variants were demonstrated as an under-recognized functional class that contribute to human disease.

Since great importance has been attached to small proteins, in-depth investigations of small proteins across various species are in need. SmProt is dedicated to integrate knowledge of proteins shorter than 100 amino acids (hereinafter referred to as small proteins) translated from various sources, especially for ones from UTRs and non-coding RNAs. The annotation information and functional sections in the current release are much richer than those in the 1st release [43], and the data volume and reliability are greatly improved.

Data collection and processing

Data sources

rRibo-seq and TI-seq datasets derived from diverse tissues/cell lines were collected from GEO database [44] and European Nucleotide Archive [45]. The latest reference genomes and gene annotation were download from Ensembl [46] , GENCODE [47], and NCBI-Genome database. WGS Variants were collected from their respective websites. The construction pipeline of SmProt was summarized as follows (**Figure 1**).

Ribo-seq data processing

The fastq files of 547 Ribo-seq datasets were downloaded from GEO and European Nucleotide Archive database. Each dataset was checked manually to confirm the sequencing adapters. The adapters were removed using cutadapt 1.18 [48] and only reads with length 25–35 were kept. Then the sequences were mapped to the latest genome using STAR 2.5.2a [49] using EndToEnd mode with allowance of up to 2 mismatches.

Ribo-seq quality and P-site offsets were assessed by Ribo-TISH [34] quality module. For TI-seq data, more attention was put on TIS quality (-t). Manual checks were then carried out to verify offset values and eliminate datasets without obvious triplet periodicity. After the quality control, 419 Ribo-seq datasets (Supplementary Table S1) were kept.

Translated ORFs were predicted by Ribo-TISH predict module. Biological and technical duplication data under the same treatment in one dataset were merged. Minimum amino acid length of candidate ORF was set to 5. Considering both ATG and near-cognate start codons, rRibo-seq datasets using only CHX without matched TI-seq were analyzed twice. One is prediction of ORFs with canonical ATG start codon, the other is prediction of ORFs with near-cognate start codons with 1 base different from ATG (--alt). Preferring data evidence instead of prior assumption in our database, only the best frame test results from multiple candidate start codons in the same ORF were reported (--framebest). For datasets containing TI-seq data, alternative start codons were included (--alt), and different parameters were set for LTM-based TI-seq and HARR-based TI-seq (--harr).

sORFs with less than 100 amino acids were filtered from the above prediction results. To avoid confusion from classic proteins, those marked as *known* (means the translation initiation site is annotated in another transcript), *CDSFrameOverlap* (means the ORF overlaps with annotated CDS in another transcript in the same reading frame), and *Truncated* (means the ORF is part of annotated CDS in the same transcript) without translation initiation evidence (i.e. none significant results identified from paired TI-seq datasets) were further removed, considering these results may be supported by RPFs from other classic proteins longer than 100 AAs.

In-frame reads of sORFs were counted and normalized by library sequencing depth (in-frame total reads count) and sORF length, a similar method with RPKM (Reads Per Kilobase per Million mapped reads) in RNA-seq but using ribosome profiling data which represents the translation levels.

Finally, 3,060,793 records were kept. Results with the identical genome loci in one species were merged as the same small protein generating 577,206 unique IDs, while information derived from multiple datasets were kept, a similar integration method to piRBase [50].

Variants from ribosome profiling data

We performed germline variants detection on 96 human ribosome profiling datasets, referring to the workflow for processing RNA data for germline short variant discovery with GATK v4.1.8 [51–54]. Duplicate reads were identified using MarkDuplicates tool after alignment, then reads with N in Cigar were split using SplitNCigarReads tool. Base quality score recalibration was carried out based on true sites in training sets using BaseRecalibrator tool and applied using ApplyBQSR tool. Variants were called individually in each sample using the HaplotypeCaller tool. Variants with QualByDepth (QD) < 2 were removed using VariantFiltration tool. Germline single nucleotide variants (SNVs) were linked to small proteins in SmProt according to genomic positions.

Variants from WGS data

Variants from 1KGP3 [55], GAsP [56], TOPMed [57], gnomAD3 [42,58], and NyuWa [59] were collected. VCF files were lifted over from old genome version to GRCh38 using LiftoverVcf tool of GATK with allowance to recover swapped ref and alt alleles.

Variants in 5'UTRs were evaluated for their effects on translated uORFs in SmProt using VEP [60] with plugin UTRannotator [42,61], and classified by their functional consequences.

Disease-specific small proteins

Small proteins identified only from diseased cell lines/tissues but not from corresponding normal cell lines/tissues were predicted as disease-specific translation events: if there were matched data of normal and diseased groups in the same dataset, small proteins derived uniquely from diseased group were screened as disease-specific ones; if there's no matched control group in the same dataset, the same type of healthy tissue/cell line in other datasets were used as control. If there's no matched same tissue/cell line, all data from diverse normal tissues/cell lines were merged for comparisons (Supplementary Table S2), and small proteins identified only from the diseased cell lines/tissues were predicted as tissue-specific. Disease-specific or tissue-specific translation events require RiboPvalue in disease groups under 0.01 while similar proteins with different TISs at the same loci in control group not detected (RiboPvalue higher than 0.05).

Single nucleotide variants (SNVs) detected only in diseased variant sets but not in normal sets were predicted as disease-specific SNVs. SNVs in diseased cell lines/tissues derived from ribosome profiling data and located within the genomic region of small proteins were regarded as diseased variant sets. SNVs in corresponding normal cell lines/tissues (Supplementary Table S2) derived from ribosome profiling data were combined with all variants derived from multiple WGS projects, as control variant sets for comparison.

Function domain prediction

Besides function of small proteins collected from literature mining, we used InterProScan [62] to predict function domain of small proteins, which focuses on combination of protein family membership and the functional domains/sites, and has been extensively used by genome sequencing projects and the UniProt Knowledgebase

[63]. Default thresholds and additional parameters *-goterms -pa* were adopted for gene oncology and pathway annotations.

PhyloCSF calculation

Pre-calculated BigWig data of PhyloCSF [64] scores at each base across the whole genome were downloaded from the broad institute https://data.broadinstitute.org/compbio1/PhyloCSFtracks/, and the score for genomic region of each small protein was extracted with our script using pyBigWig (https://github.com/deeptools/pyBigWig).

Database implementation

Database website was organized with HTML (https://html.spec.whatwg.org/), JavaScript (https://www.javascript.com/), PHP (https://www.php.net/), and MYSQL (https://www.mysql.com/). UCSC Genome Browser (http://genome.ucsc.edu/) was used to visualize the small proteins and variants. NCBI BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used for sequence similarity searches.

Database content and usage

Overview

SmProt was constructed by pipeline described in Figure 1. Multiple ways were provided to search, browse, visualize, and study small proteins (**Figure 2**). Small proteins were found mainly from rRibo-seq and TI-seq data. All information for small proteins from different data sources and datasets were integrated. General information for small proteins was provided such as sequence, mass, location, blocks, tissue or cell line, predicted functions, conservation, and multiple IDs including small protein ID, Ensembl ID, and NONCODE [65] ID. Translation level (in frame counts and Ribo RPKM) of small proteins identified from each dataset and record was provided. Details for their related variants and diseases were also provided (**Figure 3**). SmProt now has 638,958 unique small proteins and 3,165,229 small protein records in total (**Table 1;** Supplementary Table S3).

Reliability of small proteins

SmProt emphasizes reliability of small proteins, which is guaranteed mainly by the significance of 3 nt periodicity in RPF P-site profile:

Firstly, we constructed new pipeline based on independently published toolkit Ribo-TISH [34], which allows for accurate detection of ORFs and TISs using rRibo-seq and TI-seq. Ribo-TISH uses rank sum test to detect 3 nt periodicity, and negative binomial test to detect translation initiation sites, which outperforms other established methods in prediction accuracy.

Secondly, in addition to the quality control based on Ribo-TISH quality module, manual checks were also carried out to ensure clear triplet periodicity and unambiguous offset of Ribo-seq data, which further eliminated noises.

Thirdly, we provided several evaluations as evidences: p-values of small proteins called from multiple ribosome profiling datasets indicating the confidence in different samples and conditions; PhyloCSF conservation of genomic regions reflecting coding potential; and peptide evidence derived from mass spectrum data. All evidences were exhibited in the small protein page. What’s more, a set with evidence of both translation events and protein fragments was provided on download page.

What’s more, information of small protein derived from multiple sources were integrated in small protein information page.

Variants related to small proteins

25,475 variants located on translated sORFs were provided, which were exhibited in the related small protein page. For uORF-perturbing variants are likely to impact translation of downstream proteins [42], variants from multiple WGS projects and ribosome profiling data were evaluated for their effects on translated uORFs in SmProt, which can be found at variants page.

Disease-specific small proteins

Disease-specific small proteins have potential to be candidates of molecular markers or targets for diagnosis and treatment. Disease-specific translation events as well as disease-specific SNVs of small proteins in 16 types of diseases were identified (see methods) (Supplementary Table S4). Besides, small proteins that have been verified experimentally in certain diseases were also documented through literature mining.

Human microbiomes small proteins

Over 4000 conserved small protein families identified from human microbiomes were collected [2]. A new section *HumanMicroBio* was created to integrate and display selected information of these small protein families.

Other sources

We use a set of keywords (Supplementary File S1) to search articles about small proteins in PubMed database. High-confidence small proteins in CCDS [66] and Swiss-Prot [67] were also integrated. Literature mining is processed in stages, and the newly published data from other sources is being released continuously after accomplishment of manual review and curation.

Function domain prediction

For successfully predicted functions of small proteins derived from ribosome profiling and literature mining, SmProt provided graph for visualization and prediction details including gene oncology (GO) and pathway annotation. Users can choose *predicted functions* on *Browse* page to filter the results with function domain prediction.

Inner BLAST

The abundant small proteins across multiple species allows for sequence similarity searches of both nucleotides and proteins. Users can search for sequences of interests using BLASTp and BLASTx (NCBI BLAST 2.2.24 release) online.

Visualization using UCSC Genome Browser

SmProt incorporated UCSC Genome Browser [68] for visualization of all the information including genomic loci of small proteins, variants from ribosome profiling data and multiple WGS projects related to small proteins, MS data, and gene annotation. The latest genome versions including hg38, mm10, rn6, dm6, ce11, sacCer3, and danRer11 were provided.

Comparison with other databases

SmProt currently includes 419 Ribo-seq datasets derived from 116 cell lines/tissues, compared to 60 datasets derived from 37 cell lines/tissues in the initial version. The number of small protein records identified from ribosome profiling in the current release is 60 times that of the 1st release (3 million to 0.05 million). The current release of SmProt combined a large amount of duplicate records in 1st release [43], and Ribo-seq analysis pipeline was optimized to ensure the reliability of our results. Variants in translated sORFs identified from Ribo-seq data as well as uORF-perturbing variants identified from WGS projects were provided. Disease-specific small proteins may provide new perspectives for clinical studies.

Currently, there are a few databases for small proteins such as ARA-PEPs [69], PsORF [70], and sORFs.org [71]. ARA-PEPs and PsORF only harbors small proteins in plants. sORFs.org developed simple inner TIS-calling algorithm not based on triplet periodicity, which should be the most important feature of Ribo-seq. SmProt emphasizes high confidence using our Ribo-TISH pipeline that is more accurate than previous methods. SmProt analyzed 419 Ribo-seq datasets, while there were only 78 in sORFs.org. SmProt pays special attention to function, variants, and related diseases of small proteins, and WGS data resource are also integrated, which other databases didn’t pay attention to.

Other proteomic databases such as UniProt, neXtProt [72], and OpenProt [73] are not specifically designed for small proteins. neXtProt only harbors proteins of human while SmProt harbors small proteins in 8 species. OpenProt also used ribosome profiling and mass spectrum to predict proteins including some small proteins longer than 30 amino acids, while SmProt analyzed much more ribosome profiling datasets

(419), which were about 5 times that in OpenProt (87), and provided small proteins longer than 5 amino acids.

Conclusion

In brief, SmProt integrated small proteins from large amount of ribosome profiling data, and provides more abundant details. We strongly believe that SmProt will provide valuable and accurate information on small proteins for scientific community. Moreover, it provides a new resource for users interested in function and mechanism study, and a reference for construction of mass spectrometry library of small proteins.

Data Availability

SmProt is publicly available at http://bigdata.ibp.ac.cn/SmProt/.

CRediT author statement

Yanyan Li: Conceptualization, Methodology, Investigation, Formal analysis, Data Curation, Writing - Original Draft, Software, Visualization. **Honghong Zhou:** Investigation, Data Curation, Funding acquisition. **Xiaomin Chen:** Investigation, Data Curation. **Yu Zheng:** Data Curation, Software, Visualization. **Quan Kang:** Software, Visualization. **Di Hao:** Data Curation, Software. **Lili Zhang:** Visualization. **Tingrui Song:** Visualization. **Huaxia Luo:** Writing - Review & Editing. **Yajing Hao:** Writing - Review & Editing. **Yiwen Chen:** Software. **Runsheng Chen:** Resources, Supervision, Funding acquisition. **Peng Zhang:** Conceptualization, Methodology, Investigation, Software, Writing - Review & Editing, Visualization, Project administration, Funding acquisition. **Shunmin He:** Conceptualization, Methodology, Resources, Investigation, Writing - Review & Editing, Supervision, Funding acquisition.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant Nos. 81902519, 31871294, 31701117, and 31970647); the National Key R&D Program of China (Grant Nos. 2016YFC0901702 and 2018YFA0106901); the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB38040300); the 13th Five-year Informatization Plan of Chinese Academy of Sciences (Grant No. XXH13505-05); special investigation on science and technology basic resources of the MOST, China (Grant No. 2019FY100102); the National Genomics Data Center. We thank Center for Big Data Research in Health (http://bigdata.ibp.ac.cn/), Institute of Biophysics, Chinese Academy of Sciences, for supporting data analysis and computing resource.

References

[1] Basrai MA, Hieter P, Boeke JD. Small open reading frames: beautiful needles in the haystack. Genome Res 1997;7:768–71.

[2] Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, et al. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. Cell 2019;178:1–15.

[3] Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J 2014;33:981–93.

[4] Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, et al. Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. Cell Rep 2014;7:1858–66.

[5] van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, et al. The translational landscape of the human heart. Cell 2019;178:242–60.e29.

[6] 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 U S A 2009;106:7507–12.

[7] Zhu S, Wang J, He Y, Meng N, Yan GR. Peptides/Proteins encoded by non-coding RNA: a novel resource bank for drug targets and biomarkers. Front Pharmacol 2018;9:1295.

[8] Li LJ, Leng RX, Fan YG, Pan HF, Ye DQ. Translation of noncoding RNAs: focus on lncRNAs, pri-miRNAs, and circRNAs. Exp Cell Res 2017;361:1–8.

[9] Fang J, Morsalin S, Rao V, Reddy ES. Decoding of non-coding DNA and non-coding RNA: pri-micro RNA-encoded novel peptides regulate migration of cancer cells. J Pharm Sci 2017;3:23–7.

[10] Razooky BS, Obermayer B, O'May JB, Tarakhovsky A. Viral infection identifies micropeptides differentially regulated in smORF-containing lncRNAs. Genes (Basel) 2017;8:206.

[11] Huang JZ, Chen M, Chen, Gao XC, Zhu S, Huang H, et al. A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. Mol Cell 2017;68:171–84.e6.

[12] Zhang M, Zhao K, Xu X, Yang Y, Yan S, Wei P, et al. A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. Nat Commun 2018;9:4475.

[13] Couso JP, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol 2017;18:575–89.

[14] Freyer L, Hsu CW, Nowotschin S, Pauli A, Ishida J, Kuba K, et al. Loss of Apela peptide in mice causes low penetrance embryonic lethality and defects in early mesodermal derivatives. Cell Rep 2017;20:2116–30.

[15] Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol 2007;5:e106.

[16] Guo B, Zhai D, Cabezas E, Welsh K, Nouraini S, Satterthwait AC, et al. Humanin peptide suppresses apoptosis by interfering with Bax activation. Nature 2003;423:456–61.

[17] Anderson DM, Anderson KM, Chang CL, Makarewich CA, Nelson BR, McAnally

JR, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell 2015;160:595–606.

[18] Knappe D, Goldbach T, Hatfield MP, Palermo NY, Weinert S, Strater N, et al. Proline-rich antimicrobial peptides optimized for binding to *Escherichia coli* chaperone DnaK. Protein Pept Lett 2016;23:1061–71.

[19] Yaran W, Yang L, Yiming X, Yiwei Z, Rui H, Kaibo W, et al. Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. Nature Genetics 2009;41:228–33.

[20] Cheng W, Wang S, Mestre AA, Fu C, Makarem A, Xian F, et al. C9ORF72 GGGGCC repeat-associated non-AUG translation is upregulated by stress through eIF2alpha phosphorylation. Nat Commun 2018;9:51.

[21] Hsu PY, Benfey PN. Small but mighty: functional peptides encoded by small ORFs in plants. Proteomics 2018;18:e1700038.

[22] Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 2009;324:218–23.

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

[24] Weiss RB, Atkins JF. Translation goes global. Science 2011;334:1509–10.

[25] Schneider-Poetsch T, Ju J, Eyler DE, Dang Y, Bhat S, Merrick WC, et al. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nat Chem Biol 2010;6:209–17.

[26] Calviello L, Ohler U. beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. Trends in Genetics 2017;33:728–44.

[27] Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc 2012;7:1534–50.

[28] Lee S, Liu B, Lee S, Huang S-X, Shen B, Qian S-B. Global mapping of translation

initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A 2012;109:E2424–E32.

[29] 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;273:491–6.

[30] Oyama M, Kozuka-Hata H, Suzuki Y, Semba K, Yamamoto T, Sugano S. Diversity of translation start sites may define increased complexity of the human short ORFeome. Mol Cell Proteomics 2007;6:1000–6.

[31] Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods 2016;13:165–70.

[32] Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, Mertins P, et al. A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. Mol Cell 2015;60:816–27.

[33] Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife 2015;4:e08890.

[34] Zhang P, He D, Xu Y, Hou J, Pan BF, Wang Y, et al. Genome-wide identification and differential analysis of translational initiation. Nat Commun 2017;8:1749.

[35] Malone B, Atanassov I, Aeschimann F, Li X, Grosshans H, Dieterich C. Bayesian prediction of RNA translation from ribosome profiling. Nucleic Acids Res 2017;45:2960–72.

[36] Raj A, Wang SH, Shim H, Harpak A, Li YI, Engelmann B, et al. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. Elife 2016;5.

[37] Chun SY, Rodriguez CM, Todd PK, Mills RE. SPECtre: a spectral coherence--based classifier of actively translated transcripts from ribosome profiling sequence data. BMC Bioinformatics 2016;17:482.

[38] Crappe J, Ndah E, Koch A, Steyaert S, Gawron D, De Keulenaer S, et al. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS

integration. Nucleic Acids Res 2015;43:e29.

[39] Wang XQ, Rothnagel JA. 5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. Nucleic Acids Res 2004;32:1382–91.

[40] Gunisova S, Valasek LS. Fail-safe mechanism of GCN4 translational control--uORF2 promotes reinitiation by analogous mechanism to uORF1 and thus secures its key role in GCN4 expression. Nucleic Acids Res 2014;42:5880–93.

[41] Ishimura R, Nagy G, Dotu I, Zhou H, Yang XL, Schimmel P, et al. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. Science 2014;345:455–9.

[42] Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, et al. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nat Commun 2020;11:2523.

[43] Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, et al. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Brief Bioinform 2018;19:636–43.

[44] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 2013;41:D991–5.

[45] Silvester N, Alako B, Amid C, Cerdeno-Tarraga A, Clarke L, Cleland I, et al. The European Nucleotide Archive in 2017. Nucleic Acids Res 2018;46:D36–D40.

[46] Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res 2018;46:D754–D61.

[47] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47:D766–D73.

[48] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 2011;17:3.

[49] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR:

ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.

[50] Wang J, Zhang P, Lu Y, Li Y, Zheng Y, Kan Y, et al. piRBase: a comprehensive database of piRNA sequences. Nucleic Acids Res 2019;47:D175–D80.

[51] Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 2018:201178.

[52] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43:1101–33.

[53] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8.

[54] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.

[55] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature 2015;526:75–81.

[56] Consortium GK. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature 2019;576:106–11.

[57] Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 2021;590:290–9.

[58] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581:434–43.

[59] Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, et al. NyuWa genome resource: deep whole genome sequencing based Chinese population variation profile and reference panel. bioRxiv 2020:2020.11.10.376574.

[60] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122.

[61] Zhang X, Wakeling M, Ware J, Whiffin N. Annotating high-impact 5'untranslated region variants with the UTRannotator. Bioinformatics 2020;14:btaa783.

[62] Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;30:1236–40.

[63] Consortium U. UniProt: a hub for protein information. Nucleic Acids Res 2015;43:D204–12.

[64] Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 2011;27:i275–82.

[65] He S, Liu C, Skogerbo G, Zhao H, Wang J, Liu T, et al. NONCODE v2.0: decoding the non-coding. Nucleic Acids Res 2008;36:D170–2.

[66] Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res 2018;46:D221–D8.

[67] Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–D15.

[68] Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 2019;47:D853–D8.

[69] Hazarika RR, De Coninck B, Yamamoto LR, Martin LR, Cammue BP, van Noort V. ARA-PEPs: a repository of putative sORF-encoded peptides in *Arabidopsis thaliana*. BMC Bioinformatics 2017;18:37.

[70] Chen Y, Li D, Fan W, Zheng X, Zhou Y, Ye H, et al. PsORF: a database of small ORFs in plants. Plant Biotechnol J 2020;18:2158–60.

[71] Olexiouk V, Van Criekinge W, Menschaert G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res 2018;46:D497–D502.

[72] Gaudet P, Michel PA, Zahn-Zabal M, Britan A, Cusin I, Domagalski M, et al. The neXtProt knowledgebase on human proteins: 2017 update. Nucleic Acids Res 2017;45:D177–D82.

[73] Brunet MA, Brunelle M, Lucier JF, Delcourt V, Levesque M, Grenier F, et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. Nucleic Acids Res 2019;47:D403–D10.

Figure legends

Figure 1 Construction pipeline of SmProt

Blue background: data sources. Yellow background: management processes. Red background: results. Abbreviations: WGS, whole genome sequencing; MS, mass spectrometry; TIS, translation initiation site; ORF, open reading frame; sORF, small open reading frame; uORF, upstream open reading frame.

Figure 2 Usage of SmProt

SmProt provided multiple ways to search, browse, visualize small proteins, related diseases, and variants. Abbreviations: WGS, whole genome sequencing; ORF, open reading frame.

Figure 3 Contents of SmProt

Detailed information for small proteins, including general annotation, information from ribosome profiling data, literature, other databases, mass spectrometry, function domain prediction, related diseases, related variants from WGS projects as well as corresponding effects, etc. Abbreviations: WGS, whole genome sequencing; TIS, translation initiation site.

Tables

Table 1 Statistics of unique small proteins in SmProt

Supplementary material

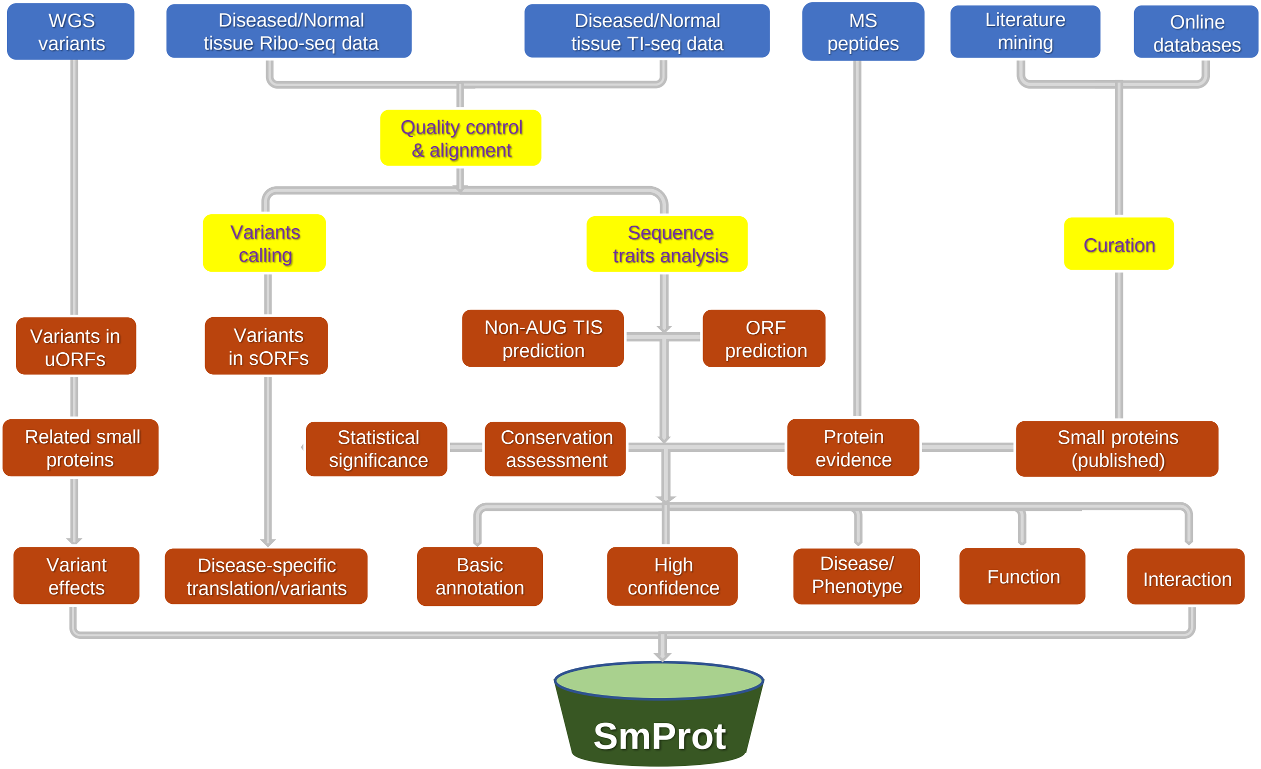
Supplementary Table S1: Information of ribosome profiling datasets analyzed in SmProt (.xlsx)

Supplementary Table S2: Dataset contrasts for generating disease-specific small proteins and variants (.xlsx)

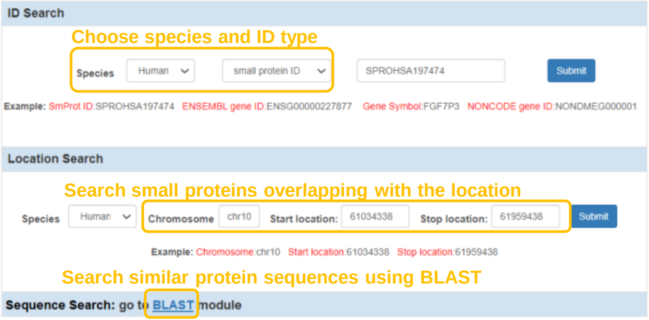
Supplementary Table S3: Statistics of small proteins primary records in SmProt (.xlsx)

Supplementary Table S4: Statistics of small proteins and variants specific in diverse diseases (.xlsx)

Supplementary File S1: Keywords for literature mining (.doc)


WGS variants
Diseased/Normal tissue Ribo-seq data
Diseased/Normal tissue TI-seq data
MS peptides
Literature mining
Online databases
Quality control & alignment
Variants calling
Sequence traits analysis
Curation
Variants in uORFs
Variants in sORFs
Non-AUG TIS prediction
ORF prediction
Related small proteins
Statistical significance
Conservation assessment
Protein evidence
Small proteins (published)
Variant effects
Disease-specific translation/variants
Basic annotation
High confidence
Disease/ Phenotype
Function
Interaction
SmProt

Search small proteins



ID Search

Choose species and ID type

Species: Human | small protein ID | SPROHSA197474 | Submit

Example: SmProt ID:SPROHSA197474 ENSEMBL gene ID:ENSG00000227877 Gene Symbol:FGF7P3 NONCODE gene ID:NONDMEG000001

Location Search

Search small proteins overlapping with the location

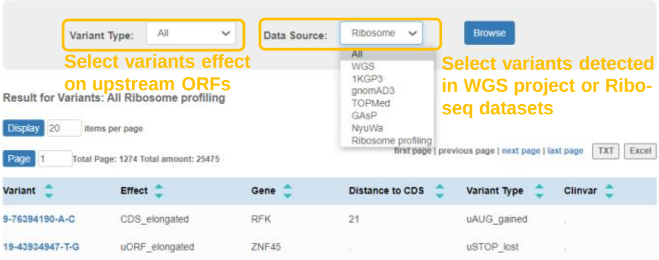
Species: Human | Chromosome: chr10 | Start location: 61034338 | Stop location: 61959438 | Submit

Example: Chromosome:chr10 Start location:61034338 Stop location:61959438

Search similar protein sequences using BLAST

Sequence Search: go to BLAST module

Browse variants related to small/upstream ORFs



Variant Type: All | Data Source: Ribosome | Browse

Select variants effect on upstream ORFs

Select variants detected in WGS project or Ribo-seq datasets

Result for Variants: All Ribosome profiling

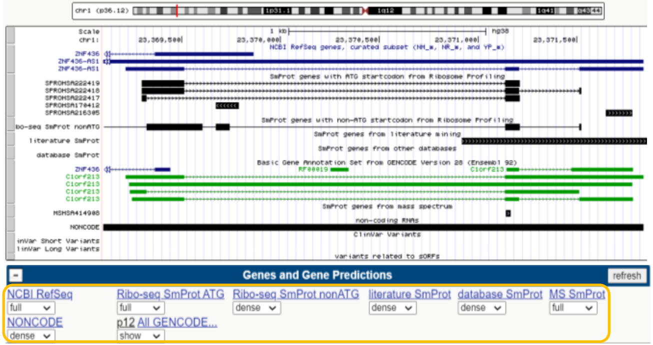
Display 20 items per page

Page 1 Total Page: 1274 Total amount: 25475

first page | previous page | next page | last page TXT Excel

Variant	Effect	Gene	Distance to CDS	Variant Type	Clinvar
9-76394190-A-C	CDS_elongated	RFK	21	uAUG_gained	.
19-43934947-T-G	uORF_elongated	ZNF45	.	uSTOP_lost	.

Visualize small proteins and variants in genome browser

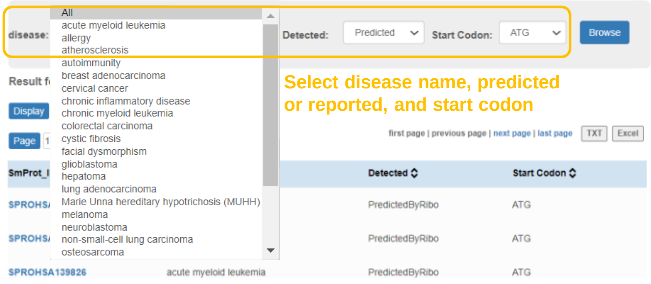


Genes and Gene Predictions refresh

NCBI RefSeq full	Ribo-seq SmProt ATG full	Ribo-seq SmProt nonATG dense	literature SmProt dense	database SmProt dense	MS SmProt full
NONCODE dense	p12 All GENCODE... show				

Users can manually change tracks to show or hide

Browse small proteins by related diseases



disease: All | Detected: Predicted | Start Codon: ATG | Browse

Select disease name, predicted or reported, and start codon

Result f

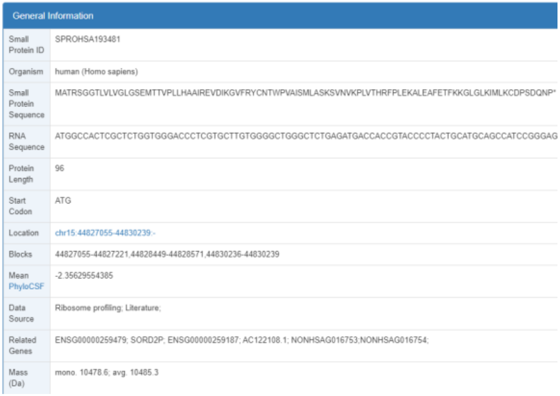
Display

Page 1

first page | previous page | next page | last page TXT Excel

SmProt_I		Detected	Start Codon
SPROHS/		PredictedByRibo	ATG
SPROHS/		PredictedByRibo	ATG
SPROHSA139826	acute myeloid leukemia	PredictedByRibo	ATG

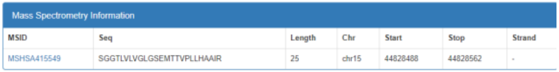
General annotation of small protein



General Information

Small Protein ID	SPROHSA193481
Organism	human (Homo sapiens)
Small Protein Sequence	MATRSGGTLVLVGLGSEMTTVPLLHAAIREVDIKGVFRYCNTWPVAISMLASKSVNVKPLVTHRFPLEKALEAFETFKKGLGLKIMLKCDPSDQNP*
RNA Sequence	ATGGCCACTCGCTCTGGTGGGACCCTCGTGCTTGTGGGGCTGGGCTCTGAGATGACCACCGTACCCCTACTGCATGCAGCCATCCGGGAG
Protein Length	96
Start Codon	ATG
Location	chr15:44827055-44830239:-
Blocks	44827055-44827221,44828449-44828571,44830236-44830239
Mean PhyloCSF	-2.35629554385
Data Source	Ribosome profiling; Literature;
Related Genes	ENSG00000259479; SORD2P; ENSG00000259187; AC122108.1; NONHSAG016753;NONHSAG016754;
Mass (Da)	mono. 10478.6; avg. 10485.3

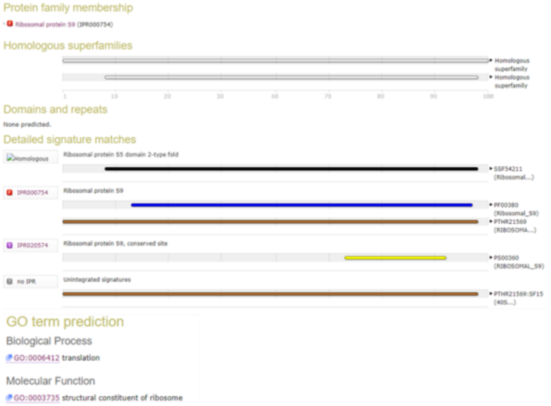
Information from other sources



Mass Spectrometry Information

MSID	Seq	Length	Chr	Start	Stop	Strand
MSHSA415549	SGGTLVLVGLGSEMTTVPLLHAAIR	25	chr15	44828488	44828562	-

Function domain prediction



Protein family membership

Ribosomal protein S9 (IPR000754)

Homologous superfamilies

Domains and repeats

None predicted.

Detailed signature matches

GO term prediction

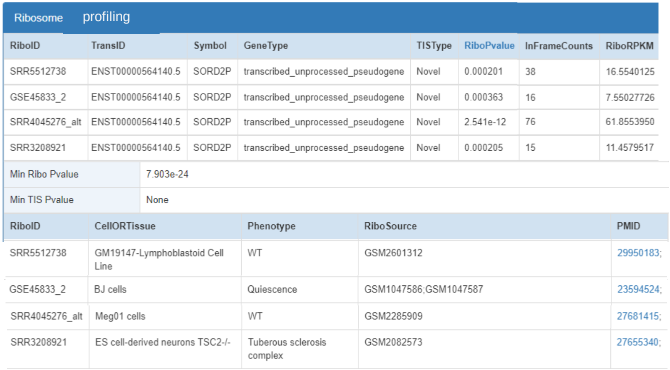
Biological Process

GO:0006412 translation

Molecular Function

GO:0003735 structural constituent of ribosome

Detailed information of the small protein in each dataset



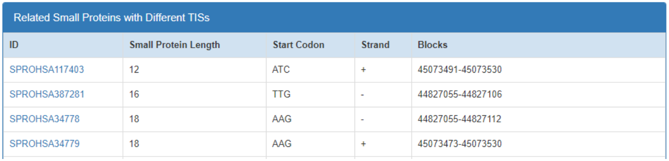
Ribosome profiling

RiboID	TransID	Symbol	GeneType	TISType	RiboPvalue	InFrameCounts	RiboRPKM
SRR5512738	ENST00000564140.5	SORD2P	transcribed_unprocessed_pseudogene	Novel	0.000201	38	16.5540125
GSE45833_2	ENST00000564140.5	SORD2P	transcribed_unprocessed_pseudogene	Novel	0.000363	16	7.55027726
SRR4045276_alt	ENST00000564140.5	SORD2P	transcribed_unprocessed_pseudogene	Novel	2.541e-12	76	61.8553950
SRR3208921	ENST00000564140.5	SORD2P	transcribed_unprocessed_pseudogene	Novel	0.000205	15	11.4579517

Min Ribo Pvalue	7.903e-24
Min TIS Pvalue	None

RiboID	CellORTissue	Phenotype	RiboSource	PMID
SRR5512738	GM19147-Lymphoblastoid Cell Line	WT	GSM2601312	29950183;
GSE45833_2	BJ cells	Quiescence	GSM1047586;GSM1047587	23594524;
SRR4045276_alt	Meg01 cells	WT	GSM2285909	27681415;
SRR3208921	ES cell-derived neurons TSC2-/-	Tuberous sclerosis complex	GSM2082573	27655340;

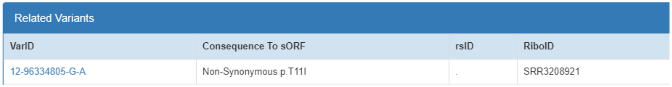
Related Small Proteins with Different TISs



Related Small Proteins with Different TISs

ID	Small Protein Length	Start Codon	Strand	Blocks
SPROHSA117403	12	ATC	+	45073491-45073530
SPROHSA387281	16	TTG	-	44827055-44827106
SPROHSA34778	18	AAG	-	44827055-44827112
SPROHSA34779	18	AAG	+	45073473-45073530

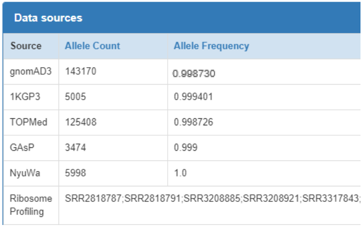
Variants on RNA sequence of the small protein



Related Variants

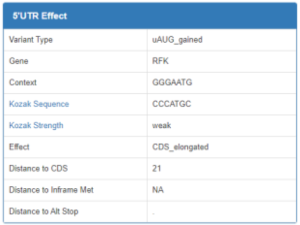
VarID	Consequence To sORF	rsID	RiboID
12-96334805-G-A	Non-Synonymous p.T11I	.	SRR3208921

Sources and effects of variants



Data sources

Source	Allele Count	Allele Frequency
gnomAD3	143170	0.998730
1KGP3	5005	0.999401
TOPMed	125408	0.998726
GAsP	3474	0.999
NyuWa	5998	1.0
Ribosome Profiling	SRR2818787;SRR2818791;SRR3208885;SRR3208921;SRR3317843;	



5'UTR Effect

Variant Type	uAUG_gained
Gene	RFK
Context	GGGAATG
Kozak Sequence	CCCATGC
Kozak Strength	weak
Effect	CDS_elongated
Distance to CDS	21
Distance to Inframe Met	NA
Distance to Alt Stop	.

Table 1 Statistics of unique small proteins in SmProt

Type	Start codon	Human	Mouse	Fruit fly	Rat	C.elegans	Yeast	E.coli	Zebrafish	All
Ribo-seq	ATG	70,931	48,909	5269	3560	4334	4535	1881	1924	141,343
	Near ATG	229,653	133,037	29,679	9910	9894	12,339	10,004	1347	435,863
Literature	All	38,157	8875	22,228	163	4	355	296	3612	73,690
Databases	All	786	797	100	271	120	336	955	64	3429
MS	All	768	51	66	38	0	3	0	1	927
All IDs	All	327,995	189,433	56,574	13,829	14,255	17,312	12,881	6679	638,958

Note: Not including human microbiomes small protein families. IDs mean unique entries with identical genome loci in one species. Abbreviations: Ribo-seq, ribosome profiling; MS, mass spectrometry.