

SmProt v2 Tutorial

2020.09.08

1、Help page: introduction of SmProt and each page

3.How to use SmProt database?

Search	ID Search: search through SmProt ID, NONCODE ID, ENSEMBL ID. Location Search: search concerned location of chromosome in specific species. Hits of small proteins will be reported if their locations are overlapped with the input location.
Browse	On Browse webpage, users can choose species (human, mouse, etc.), start codon (ATG, non-ATG), data source (ribosome profiling, mass spectrum, etc.), predicted function (yes/no, means whether have function domain prediction). Click Browse button and the filtered results with brief information will be listed below. Click on one SmProt_ID to jump to the page with detailed information.
Variants	On Variants webpage, variants related to small ORFs in 5'UTR called from WGS data of multiple projects are provided, as well as their effects on downstream gene expressions and translated uORF in SmProt. Users can choose data source (WGS project) and variant type (uAUG_gained, uSTOP_lost, etc., means effects of variants). Click on one variant to jump to the page with detailed information.
Diseases	On Diseases webpage, disease-specific translation events and variants in small proteins predicted from ribosome profiling data are provided (confidence: predicted specific), as well as disease-related small proteins reported in literature (confidence: reported related). Users choose species, then diseases list will be attached to the chosen species. Users can further choose confidence and start codon of small proteins.
Human Microbio	On HumanMicroBio webpage, users can choose body site (skin, gut, etc.) to see small proteins identified from microorganism samples from the body site. The brief results show total number, length and representative sequence of each family. Click on the Family ID to jump to the page with corresponding detailed information.
Inner BLAST	On Blast webpage, users can assess sequence similarity of small proteins in multiple species. All small proteins in SmProt v2.0 were added to the blast database. Program blastp means from protein to protein, blastx means from translated nucleotide to protein. Users can enter fasta format sequence directly or load fasta files from disk. The results can be generated with default parameters or specified parameters.
Genome Browser	Users can click Genome Button on Navigation Bar, or location link in General Information table in any small protein page, or genome browser link on Dataset table in any small protein page, to jump to Genome browser webpage to check small proteins on a genomic region. Users can manually change tracks to be shown or hidden.
Terminology Explanation	PhyloCSF: conservation of genomic region which reflects the coding potential. RiboPvalue: One tailed rank sum test p-value for regular riboseq frame bias inside ORF (frame test). TISPvalue: One tailed negative binomial test p-value for TISCount (TIS test). MS evidence: translation evidence from mass spectrum experiments. TISCount: Number of reads with P-site at TIS site. Kozak sequence: (GCC)GCCA/GCCATGG, emerges as the consensus sequence for initiation of translation in vertebrates. Kozak Strength: the likelihood of an AUG initiating translation. oORF: overlapping open reading frame (with downstream gene).

Beside help page, usage introduction can also be found in each page

Search

Search small proteins using multiple IDs or genomic location. Please use standard ID/symbol instead of unconventional synonyms to make sure the results to be found.

Location Search

Auto-filling text as example

Species

Human

Chromosome

chr10

Start location:

61034338

Stop location:

61959438

Submit

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

Search hints

ID Search: search through SmProt ID, NONCODE ID, ENSEMBL ID, and symbol annotated by ENSEMBL. Please use standard ID/symbol instead of unconventional synonyms to make sure the results to be found.

Location Search: search concerned location of chromosome in specific species. Hits of small proteins will be reported if their locations are overlapped with the input location.

2、Search page: Search small proteins

Search

Search small proteins using multiple IDs or genomic location. Please use standard ID/symbol instead of unconventional synonyms to make sure the results to be found.

ID Search

Choose species and ID type

Species

Human

small protein ID

ENSEMBL Gene

Gene Symbol

NONCODE Gene

Auto-filling text as example

SPROHSA197474

Submit

Example:

SmProt ID:SPROHSA197474

ENSEMBL Gene:ENSG00000187777

Gene Symbol:FGF7P3

NONCODE gene ID:NONDMEG000001

Location Search

Species

Human

Chromosome

chr1C

Start location:

61034338

Stop location:

61959438

Submit

Example:

Chromosome:chr10

Start location:61034338

Stop location:61959438

Search results of ID

Change items showed per page

Change page

Display

20

items per page

Page

1

Total Page: 1

Total amount: 1

first page

previous page

next page

last page

TXT

Excel

SmProt_ID	Organism	SmProt_length	Protein Sequence	Start Codon
SPROHSA197474		77	MAGVLKKTGLVGLAVCNTPHEEPDVKKLEDQLQG...	ATG

Users can further filter the search results:

download the results

Species

Human

StartCodon

All

Data Sources

All

predicted functions

All

Browse

Display

20

items per page

Page

1

Total Page: 1

Total amount: 13

first page

previous page

next page

last page

TXT

Excel

SmProt_ID	Organism	SmProt_length	Protein Sequence	Start Codon
SPROHSA84741		68	RDHKQQQVSVLVIFLLTGGLRARPAGSWGRRQGDV...	AGG
SPROHSA226181		80	MFVGTAADILEFTSETLEEQNVRNSPALVYAILVI...	ATG
SPROHSA283407		71	LEFTSETLEEQNVRNSPALVYAILVIWTWSMLQFP...	CTG

Search results of location

Click result ID to see detailed information

3、Browse page: browse small proteins

Small Proteins List

Customize conditions for screening small proteins of interest.
All results derived from ribosome profiling are totally new!
The organization structure of all data is brand new!

Species

Human

Start Codon

ATG

Data Sources

All

Mass Spectrum

All

predicted functions

All

Browse

Browse options

Species

- Human
- Mouse
- Zebrafish
- Yeast
- Fruitfly
- Escherichia coli
- Rat
- C.elegans

Data Source

- All
- Literature Mining
- Ribosome profiling
- Known Database

Start Codon

- All
- ATG
- non-ATG
- AAG
- ACG
- AGG
- ATA
- ATC
- ATT
- CTG
- GTG
- TTG

Mass Spectrum

- All
- Has MS evidence
- No MS evidence

Predicted function

- All
- Have predicted functions
- No predicted functions

Display

20

items per page

Page

1

Total Page: 5447 Total amount: 108931

first page | previous page | next page | last page

TXT

Excel

Click the triangle symbol to sort by ID, length, sequence...

SmProt_ID	Organism	SmProt_length	Protein Sequence	Start Codon
SPROHSA136182		65	MK K K K K K S L R S L Q F Q F L F H S V S Q T P T H H S L E N G K K ...	ATG
SPROHSA136183		88	M K K N N I P E P V V I E I V W S N V M S A V E W N K R E E I V A E Q ...	ATG

Click result ID to see detailed information

Detailed information of browse/search results in small protein page

General Information	
Small Protein ID	SPROHSA136197
Organism	human (Homo sapiens)
Small Protein Sequence	MKKMHYVDPDHSVKTYYTVPLKEAGPSLLKHSVSPGTSIFKPSLFSP*
RNA Sequence	ATGAAAAAATGCATTATGTGGACCCTGACCATGTAAAGACCTACACCGTGCCTTTAAAGGAAGCAGGGCCCTCCCTGCTGAAGCATTCAAGTGAGC...
Protein Length	45
Start Codon	ATG
Location	chr2:85852901-85861219:-
Blocks	85852901-85853000,85861180-85861219
Mean PhyloCSF	-8.62465220258

4、Browse Variants page: variants related to small proteins

Human 5'UTR Variants List

Now you can browse variants in 5'UTRs called from WGS data of multiple projects such as gnomAD2, gnomAD3, NyuWa, 1KGP, etc., and small proteins related to the variants, as well as vatiants in sORFs called from Ribo-seq data. The data are being updated continuously.

Select variants effect on upstream ORF Select variants detected in WGS project or ribo-seq datasets

Variant Type:

All

uAUG_gained

uSTOP_lost

Data Source

All

WGS

1KGP3

gnomAD3

TOPMed

GAsP

NyuWa

Ribosome profiling

Browse

Result for Variants: All All

Display 20 items per page

Page 1 Total Page: 2191 Total amount: 43815

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	.

Click result ID to see detailed information →

Specific Information of Variant: 9-76394190-A-C

General Information	
Variant ID	9-76394190-A-C
Genome Position	chr9:76394190
ClinVar	.

Related Small Proteins					
ID	Length	Start Codon	Strand	Blocks	Consequence
SPROHSA51610	35	ACG	-	76394167-76394275	Non-Synonymous p.F29C
SPROHSA67636	16	AGG	-	76394167-76394218	Non-Synonymous p.F10C

Data sources		
Source	Allele Count	Allele Frequency
gnomAD3	143170	9.98730e-01
1KGP3	5005	0.999401
TOPMed	125408	0.998726

5、Disease page: browse small proteins or variants on small proteins related to diseases

Disease-specific translation events and variants in small proteins predicted from ribosome profiling data are provided (confidence: predicted specific), as well as disease-related small proteins reported in literature (confidence: reported related)

Small Proteins List Related to Diverse Diseases

Now you can search small proteins related to some specific diseases as references.
Or you can browse **disease-specific SNVs.**

Click to switch between disease-related small protein and variants

Variants List in RNA Sequence of Small Proteins Related to Diverse Diseases

Now you can browse disease-specific SNVs.
Or you can browse **disease related small proteins.**

disease:

All

acute myeloid leukemia

allergy

atherosclerosis

autoimmunity

breast adenocarcinoma

cervical cancer

chronic inflammatory disease

chronic myeloid leukemia

colorectal carcinoma

cystic fibrosis

facial dysmorphism

glioblastoma

hepatoma

lung adenocarcinoma

Marie Unna hereditary hypotrichosis (MUHH)

melanoma

neuroblastoma

non-small-cell lung carcinoma

osteosarcoma

Parkinsons disease

Ph+ chronic myelogenous leukemia

Progressive rod-cone degeneration (PRCD)

prostate cancer

skin cancer

thyroid carcinoma

tuberous sclerosis complex

Detected:

All

Reported related

Predicted specific translation

Start Codon:

All

ATG

non-ATG

Browse

Result for

Display

20

Page

1

first page | previous page | next page | last page

SmProt_ID

Disease

Detected

Start Codon

SPROHSA136706

acute myeloid leukemia

PredictedByRibo

ATG

disease:

All

acute myeloid leukemia

breast adenocarcinoma

cervical cancer

chronic myeloid leukemia

colorectal carcinoma

cystic fibrosis

facial dysmorphism

hepatoma

lung adenocarcinoma

neuroblastoma

non-small-cell lung carcinoma

osteosarcoma

Parkinsons disease

Ph+ chronic myelogenous leukemia

tuberous sclerosis complex

Browse

Result for

Display

20

Page

1

first page

VarID

Disease

10-17233617-G-A

acute myeloid leukemia

10-73199842-T-C

acute myeloid leukemia

6、 HumanMicroBio page: browse small proteins of Human Microbiomes

Small Proteins of Human Microbiomes (Body Site: All)

Now you can browse small proteins related to Human Microbiomes identified by
Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. Cell. 2019 Aug 6.

Select to get small proteins of human microbiomes in corresponding body site

Body Site:

All

mouth

skin

vagina

gut

non-human environments

Browse

Display 20 items per page

Page 1 Total Page: 227 Total amount: 4539

[first page](#) | [previous page](#) | [next page](#)

Family ID ↕	NO. Members ↕	Length of peptide ↕	AA sequence of clusters representative ↕
0	25	50	MKTLVVYADFDWIEETELVGELGYESLRSSDCYCFTLNDGCFEKNTEICF
10	2455	50	MKRIKTIETRDLKKS VKTG GCGECQTSCQSACKTSCGVANQQCEKCMSEK

Click on family ID to see detailed information of small proteins in this family

Detail Information of small proteins related to Human Microbiomes for Family ID: 10

General Information	
Family ID	10
No. Members	2455
Small Protein Length	50
Representative AA Seq (detail)	MKRIKTIETRDLKKS VKTG GCGECQTSCQSACKTSCGVANQQCEKCMSEK
Representative DNA Seq (detail)	ATGAAGAGAATTAAAACTATCGAAACTCGTGACCTCAAGAAGAGCGTTAAGACAGGCGGTTGCGGCGAGTGCCAGA
Example of a Refseq homolog	WP_042682158.1
Associated HGT genes	HTH_MerR-trunc, Tn916-Xis, SR_IS607_transposase_like, INT_C_like_4, Y1_Tnp, PRK09871, INT_ICEBs1_C_like, RINT_StrepXerD_C_like, INT_RitA_C_like, SR_Res_par, INT_Lambda_C, RecA-like_NTPases, PHA02517, HTH_28, reT_den_put_tspse, INT_RitC_C_like, SR_ResInv, INT_C_like_5, int, MULE, INT_C_like_3, RecA, PRK09409, HTH_HinIntegrase_DNA, PinE, INT_Rci_Hp1_C, INT_C_like_6, INT_Intl_C, INT_C_like_1, PRK15417, Phage_integrase, integr
Number of non bacterial	1
Non bacterial classification_number of members that were classified to it	cellular organisms_3
RNA code p-value assigned to family	5.73E-11
number of species	192

7、Blast page: BLAST small proteins in SmProt similar with the input sequence

Blast in SmProt database

Use this to assess sequence similarity of small proteins in multiple species.

All new small proteins were added to the blast database, including human microbiomes.

Choose program to use and database to search:

Program **blastp** means from protein to protein, **blastx** means from translated nucleotide to protein

Program **blastp** Database **SmProt_All**

Enter sequence below in **FASTA** format **Example sequence**

>example sequence
MSGKSWVLISTTSPQSLEDEILGRLLKILFVLFVDLMSIMYVVI
TS

Input sequence or upload
sequence in text files

Or load it from disk **Browse**

Set subsequence: From To

Clear sequence **Search**

Click here to get BLAST results 

Sequences producing significant alignments:

SPROHSA409509
SPROMMU172247
SPROMMU95127
SPROHSA235457
SPROHSA149855
SPROMMU95126
SPROCEL7466

Results with similar sequences
to the input one, and respective
scores of similarity

Score (bits)	E Value
90	9e-19
90	9e-19
90	9e-19
79	1e-15
79	1e-15
57	5e-09
27	5.3

>SPROHSA409509

Length = 46

Score = 89.7 bits (221), Expect = 9e-19
Identities = 46/46 (100%), Positives = 46/46 (100%)

Query: 1 MSGKSWVLISTTSPQSLEDEILGRLLKILFVLFVDLMSIMYVWITS 46
MSGKSWVLISTTSPQSLEDEILGRLLKILFVLFVDLMSIMYVWITS
Sbjct: 1 MSGKSWVLISTTSPQSLEDEILGRLLKILFVLFVDLMSIMYVWITS 46

Detailed information of results

>SPROMMU172247

Length = 60

Score = 89.7 bits (221), Expect = 9e-19
Identities = 46/46 (100%), Positives = 46/46 (100%)

Query: 1 MSGKSWVLISTTSPQSLEDEILGRLLKILFVLFVDLMSIMYVWITS 46
MSGKSWVLISTTSPQSLEDEILGRLLKILFVLFVDLMSIMYVWITS
Sbjct: 15 MSGKSWVLISTTSPQSLEDEILGRLLKILFVLFVDLMSIMYVWITS 60

8、 Browse page: browse small proteins

[Home](#) [Search](#) [Browse](#) [Variants](#) [Diseases](#) [HumanMicroBio](#) [Blast](#) [Genome](#) [Submit](#) [Download](#) [Statistics](#) [Contact](#) [Help](#)

Click Genome Button on Navigation Bar, or location link in General Information table in any small protein page and variant page, to jump to Genome browser to check small proteins on a genomic region.

General Information	
Small Protein ID	SPROHSA136183
Organism	human (Homo sapiens)
Small Protein Sequence	MKKNNIPEPVVIEIVWSNVMSAVEWVKREEIVAEQAIKHLKQHSPLLAFTTQSQSELTLLKKIQEYCYDNIHFMKA FRKIVVLFIKL*
RNA Sequence	ATGAAAAAAAAACAACATCCCAGAACCGTTGTCATCGAAATAGTCTGGTCAAATGTAATGAGCGCTGTGGAATGGAACAAAAG
Protein Length	88
Start Codon	ATG
Location	chr3:172425991-172426258:-
Blocks	172425991-172426258

Genomes Genome Browser Tools Mirrors Downloads My Data View Help About Us

Human GRCh38/hg38
Mouse GRCm38/mm10
Other

Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

<<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

1,102,837-11,267,747 164,911 bp. enter position, gene symbol or search terms go

chr1 (p36.22) p31.1 1q12 q41 4344

Scale chr1: 11,150,000 50 kb hg38 11,200,000 11,250,000

NCBI RefSeq genes, curated subset (NM_*, NR_*, and YP_*)

MTOR MTOR-AS1 ANGPTL7

SmProt genes with ATG startcodon from Ribosome Profiling
SmProt genes with non-ATG startcodon from Ribosome Profiling

SPROHSA367659 SPROHSA311265 SPROHSA240061 SPROHSA209793 SPROHSA198623 SPROHSA229087 SPROHSA215605 SPROHSA389843 SPROHSA345682

database SmProt SPROHSA038613

NONCODE

SmProt genes from literature mining
SmProt genes from other databases
SmProt genes from mass spectrum

non-coding RNAs

Repeating Elements by Repeatmasker

SINE LINE LTR DNA Simple Low Complexity Satellite RNA Other Unknown

Users can manually change tracks to be shown or hidden.

Mapping and Sequencing refresh

Base Position Assembly Chromosome Band Gap Short Match

dense hide hide hide hide

Genes and Gene Predictions refresh

NCBI RefSeq Ribo-seq SmProt Ribo-seq SmProt literature SmProt database SmProt MS SmProt

pack pack pack pack dense pack