



# Canadian Bioinformatics Workshops

[www.bioinformatics.ca](http://www.bioinformatics.ca)

This page is available in the following languages:

Afrikaans Български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto  
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)  
Euskara Suomeksi français français (CA) Galego ລາວ hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu  
Nederlands Norsk Sesotho sa Leboa polski Português română slovenščina jezik srpski (latinica) Sotho svenska  
中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:

 to Share — to copy, distribute and transmit the work

 to Remix — to adapt the work





Under the following conditions:

 Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

 Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

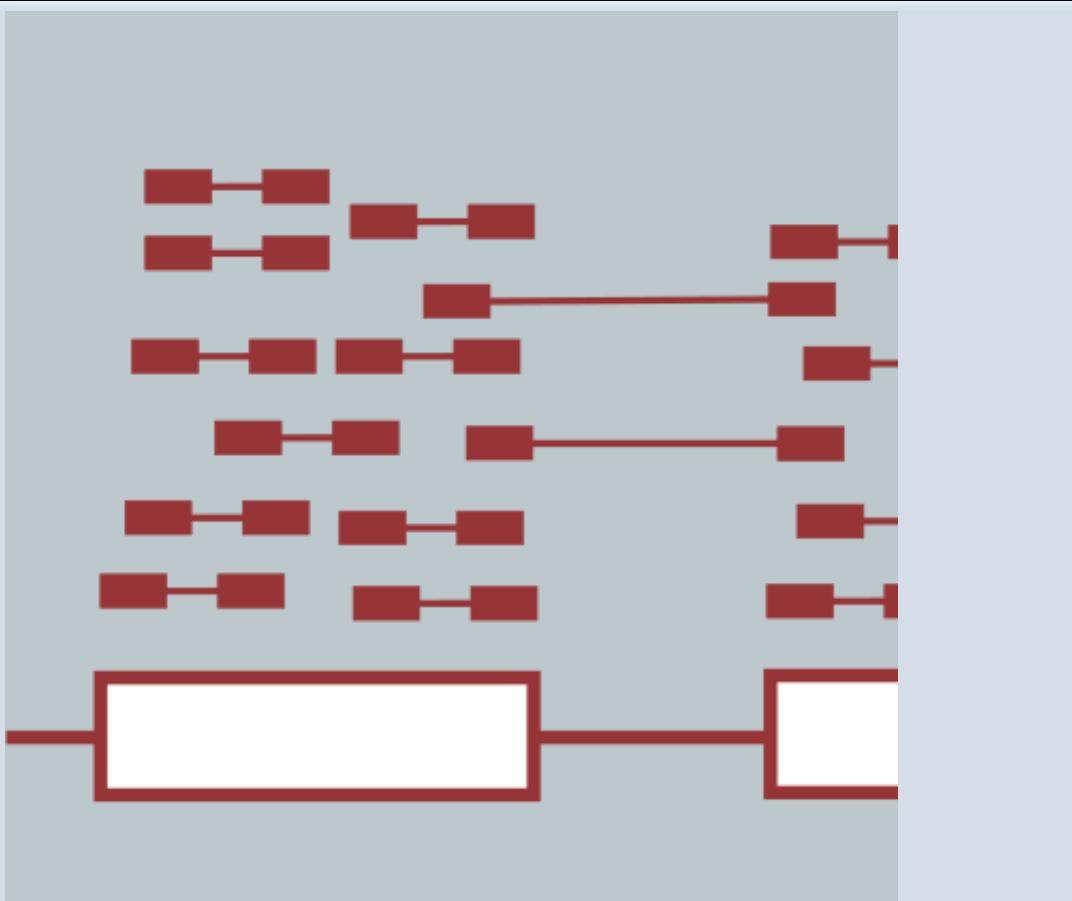
Your fair dealing and other rights are in no way affected by the above.  
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:  
[English](#) [French](#)

# Functional Annotation and Analysis of Transcripts

Brian Haas

Informatics for RNA-Seq Analysis

May 28-30, 2018



# Learning Objectives of Module

- Explore methods to glean biological function from transcript sequences.
- Differentiate between homology-based and sequence composition-based functional inference.

# Transcript Functional Annotation

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGCTGGGCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCGTGGCCT  
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGAACGTGTC  
TCTTCTGCAGGTCCCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGAA  
TCTCCG  
AAAGAC  
GGCTTC  
TGACCT  
GAAAAAC  
TTGTCA  
TCGAC  
TCCCA  
CCTGG  
CCTAA  
TGCTG  
CAGCC  
TTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTGCCAACCGCCCAGACCCCAACACGCCATGGAAGAGAGACCGTGCAGGGCCA  
TGACCCATGTCATCAACCAGGGATGGCCATGTACTGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTGGCAGTTCAACCTGATCCGCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCTCTGGCGTGCAGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCAGCAGGCCAAGCTGAAGGAACGTG

Can we gather hints of biological function  
from sequence?

# Methods used to predict function from sequence

- Sequence homology

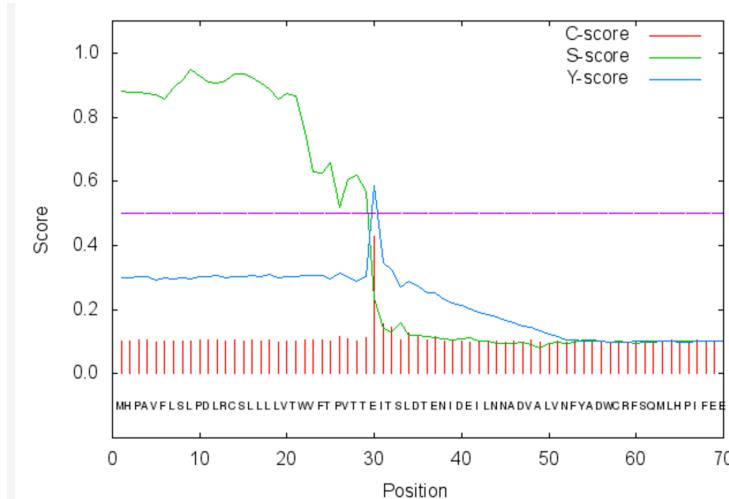
Searching protein database for sequence similarity

Query THVHRPYNEHKSLSGTARYMSINTHLGREQSRRDDLESMGHVFMYFLRGSLPW--QGLKA  
T P + K GT Y S + HLG RR DLE +G L LPW Q L A  
Database Match TGDFKP-DPKKMHNGTIEYTSRDAHLG-VPTRRADLEILGYNLIEWLGAELPWVTQKLLA

- Sequence composition

Predict functions of sequence using machine learning methods for pattern recognition.

- Neural Networks
- Hidden Markov Models



# Use BLAST to search for sequence similarity to known proteins



The screenshot shows the NCBI BLAST homepage. At the top, there's a navigation bar with links for NIH, U.S. National Library of Medicine, NCBI National Center for Biotechnology Information, and Sign in to NCBI. Below the navigation bar, the word "BLAST" is prominently displayed with a registered trademark symbol. To the right of "BLAST" are links for Home, Recent Results, Saved Strategies, and Help. The main content area features a large heading "Basic Local Alignment Search Tool". Below this, a text block explains what BLAST does: "BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance." A "Learn more" link is provided. On the right side, there's a "NEWS" section with a teal header containing the text "Magic-BLAST 1.2.0 released" and a brief description: "A new version of the BLAST RNA-seq mapping tool is now available." It also includes the date "Mon, 27 Feb 2017 14:00:00 EST" and a "More BLAST news..." link. The bottom of the page features three large boxes for "Nucleotide BLAST", "tblastn", and "Protein BLAST", each with a corresponding diagram and text.

BLAST®

Home

Recent Results

Saved Strategies

Help

NCBI National Center for Biotechnology Information

Sign in to NCBI

## Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

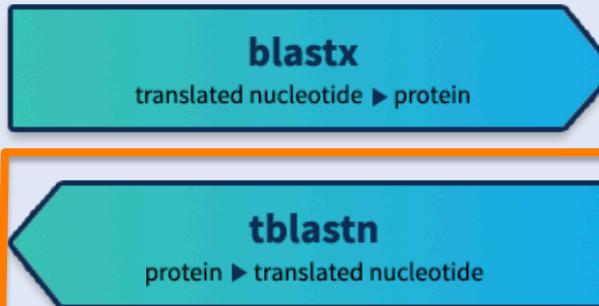
### Magic-BLAST 1.2.0 released

A new version of the BLAST RNA-seq mapping tool is now available.

Mon, 27 Feb 2017 14:00:00 EST

[More BLAST news...](#)

## Web BLAST



# The Swiss-Prot database is a valuable source of proteins with known functions

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

**UniProtKB**  
UniProt Knowledgebase  
Swiss-Prot (557,275) Manually annotated and reviewed.  
TrEMBL (114,759,640) Automatically

**UniRef** Sequence clusters

**UniParc** Sequence archive

**Proteomes**

**Supporting data**

Literature citations

Cross-ref. databases

Taxonomy

Diseases

Subcellular locations

Keywords

(as of May, 2018)

**News**

Forthcoming changes Planned changes for UniProt

UniProt release 2018\_04 The Matrix (enzymes) Reloaded | Cross-references to GlyConnect

UniProt release 2018\_03 Ama-(not a)-toxin: a cap on death | Cross-references to VGNC

News archive

## Getting started



## UniProt data

Download latest release  
Get the UniProt data

Statistics

## Protein spotlight



Giving In To Time  
May 2018

Time runs its treacherous fingers along everything.

# Example of a Swiss-Prot Record

www.uniprot.org/uniprot/Q9H479

UniProtKB Advanced Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

Basket

## UniProtKB - Q9H479 (FN3K\_HUMAN)

Display

Entry Publications Feature viewer Feature table

None

Function Names & Taxonomy Subcell. location Pathol./Biotech PTM / Processing Expression Interaction Structure Family & Domains Sequence Cross-references

Protein Fructosamine-3-kinase

Gene FN3K

Organism Homo sapiens (Human)

Status Reviewed - Annotation score: ●●●●○ - Experimental evidence at protein level<sup>i</sup>

### Function<sup>i</sup>

May initiate a process leading to the deglycation of fructoselysine and of glycated proteins. May play a role in the phosphorylation of 1-deoxy-1-morpholinofructose (DMF), fructoselysine, fructoseglycine, fructose and glycated lysozyme.

#### GO - Molecular function<sup>i</sup>

- fructosamine-3-kinase activity Source: UniProtKB
- kinase activity Source: Reactome

Complete GO annotation...

#### GO - Biological process<sup>i</sup>

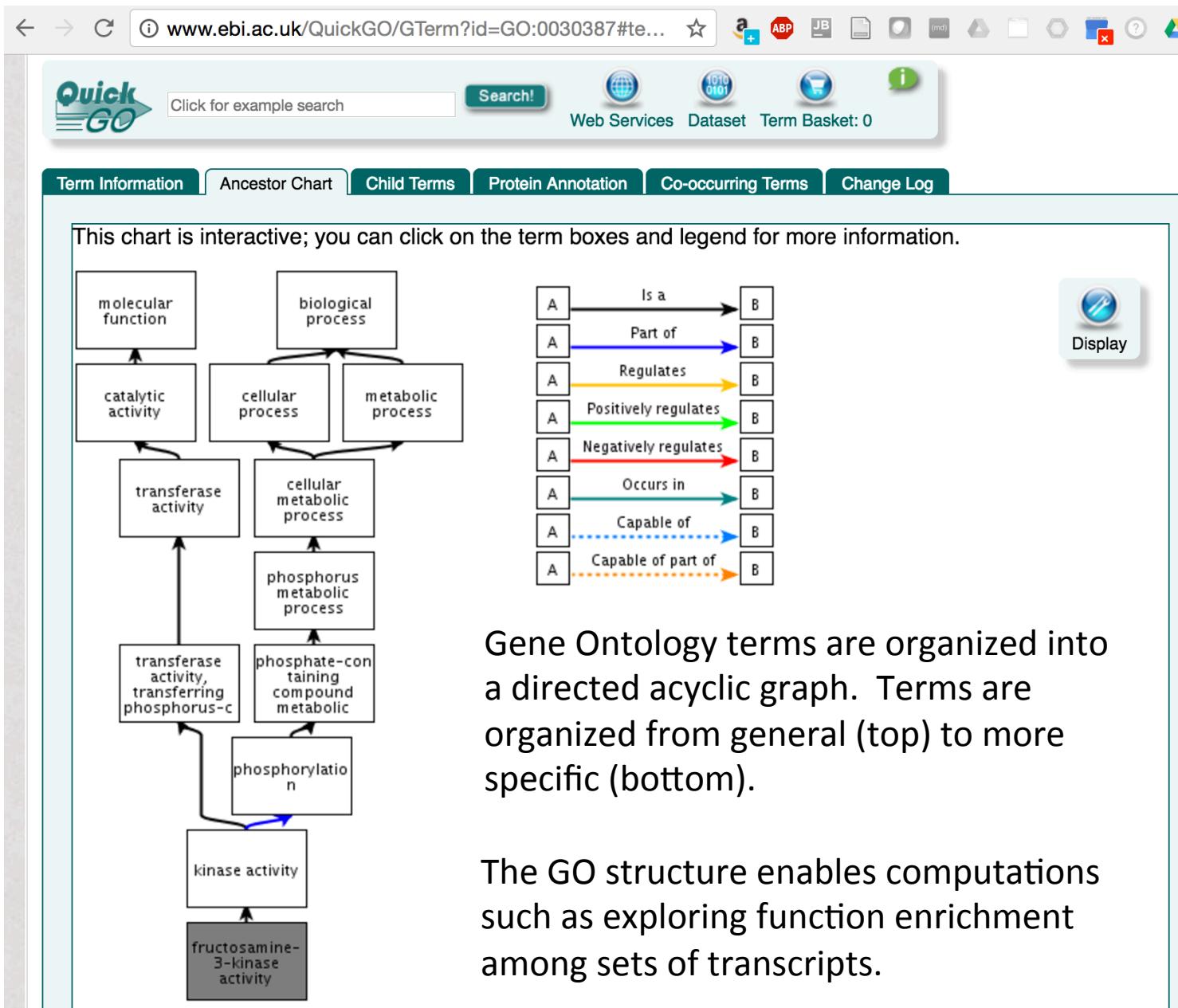
- epithelial cell differentiation Source: UniProtKB
- fructosamine metabolic process Source: GO\_Central
- fructoselysine metabolic process Source: UniProtKB
- post-translational protein modification Source: Reactome

Complete GO annotation...

### Gene Ontology (GO):

Structured vocabulary for defining molecular functions, biological processes, and cellular components.

# Gene Ontology: a structured relational vocabulary for describing biological functions



# Gene ontology functional enrichment

	(+) Differentially Expressed	(-) Not Differentially Expressed	Totals
+ Gene Ontology	50	200	250
- Gene Ontology	1950	17800	19750
Totals	2000	18000	20000

	drawn	not drawn	total
<b>green marbles</b>	$k$	$K - k$	$K$
<b>red marbles</b>	$n - k$	$N + k - n - K$	$N - K$
<b>total</b>	$n$	$N - n$	$N$

The probability of drawing exactly  $k$  green marbles can be calculated by the formula

$$P(X = k) = f(k; N, K, n) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}.$$

# No significant sequence similarity... What else?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGCTGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCGTGGCCT  
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGAACGTGTC  
TCTTCTGCAGGTCCCAGCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCA  
AAAGACAGCTCAGTTACAGGAATCTGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTGATACGGCAGGACTACGCTGCTG  
AAAAAGCTGAAGTGGTATTAGGGAACATCATTAAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAAGATCTTCTGGGTGGAAAAGCGGAGACTGAGAGAGGGCTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTGCCAACCGCCCAGACCCAACACGCCATGGAAGAGAGACCGTGCAGGGCCA  
TGACCCATGTCATCAACCAGGGATGGCCATGTACTGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTGGCAGTTCAACCTGATCCGCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCTCTGGCGTGCAGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCAGCAGGCCAAGCTGAAGGAAGTGC

# Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGCTGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCGTGGCCT  
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGAACGTGTC  
TCTTCTGCAGGTCCCAGCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCA  
AAAGACAGCTCAGTTACAGGAATCTGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTGATACGGCAGGACTACGCTGCTG  
AAAAAGCTGAAGTGGTATTAGGGAACATCATTAAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAAGATCTTCTGGGTGGAAAAGCGGAGACTGAGAGAGGGCTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTGCCAACCGCCCAGACCCAACACGCCATGGAAGAGACCGTGCAGGGCCA  
TGACCCATGTCATCAACCAGGGATGGCCATGTACTGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTGGCAGTTCAACCTGATCCGCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCTCTGGCGTGCAGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCAGCAGGCCAAGCTGAAGGAACCTGC

# Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGCTGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGCCCTGGTTAGTCTCTGAGTGTGCA  
GTTGCTGCAC**ATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCGTGGCCT**  
**TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGAACGTGTC**  
TCTTCTGCAGGTCCCAGCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCA  
AAAGACAGCTCCAGTTTACAGGAATCTGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTGATACGGCGGAGGTCTACGCTGCTG  
AAAAAGCTGAAGTGGTATTAGGGAACATCATTAAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAAGATCTTCTGGGTGGAAAAGCGGAGACTGAGAGAGGGCTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTGCCAACCGCCCAGACCCAACACGCCATGGAAGAGACCGTGCAGGGCCA  
TGACCCATGTCATCAACCAGGGATGGCCATGTACTGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTGGCAGTTCAACCTGATCCGCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCTCTGGCGTGCAGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCAGCAGGCCAAGCTGAAGGAACCTGC

# Find all ORFs using ORFfinder

Secure <https://www.ncbi.nlm.nih.gov/orffinder/>

NCBI Resources How To Sign in to NCBI

ORFfinder PubMed Search

## Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for [Linux x64](#).

**Examples** (click to set values, then click Submit button) :

- NC\_011604 *Salmonella enterica* plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM\_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

**Enter Query Sequence**

Enter accession number, gi, or nucleotide sequence in FASTA format:

```
GGAGCTGGAGGCCCGCAGGCAACTACACCGTCCACGTACCCAGAGGGCTGGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCCTAGGGCCCTGGTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGCCCTGGCCTTGCTGCACCAACTCCTGTTGGGCCGTGGCCT  
TGGAGGCATGCAGTTACGAGACAGTCAGCCATCCACCCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGAATCAACCCACGGGCTCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTACAGGAATCTGGCAAATCTGCCCTCGGGCTCCTGCCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTGGCCTACGATAATGGCATCAACCTGTTGATACGGGGAGGTACGCTGCTG
```

From: To:



# ORFfinder finds all open reading frames and provides translations

The screenshot shows the NCBI ORFfinder interface. At the top, there's a browser header with 'Secure' and the URL 'https://www.ncbi.nlm.nih.gov/orffinder/'. Below it is a blue navigation bar with 'NCBI Resources' and 'How To'. On the left, 'ORFfinder' is selected. A search bar has 'PubMed' dropdown and a 'Search' button. The main content area is titled 'Open Reading Frame Viewer'. It displays a sequence 'ORFFinder\_4.25.202734829' with a length of 1.8K (1.8Kbp). The sequence is shown as a green bar with various orange arrows indicating the direction of each predicted ORF. Labels for the ORFs include ORF4, ORF5, ORF1, ORF11, ORF3, ORF9, ORF2, ORF10, ORF7, ORF8, and ORF12. A search bar at the top of the viewer says 'Find: ATG'. Below the viewer, a message box says 'ORFs can appear in random sequence – so further analysis is required'. At the bottom, there's a link 'Predict coding vs. non-coding ORFs: http://TransDecoder.github.io'.

Predict coding vs. non-coding ORFs: <http://TransDecoder.github.io>

Add six-frame translation track

ORF5 (367 aa)

Display ORF as...

Mark

Mark subset...

Marked: 0

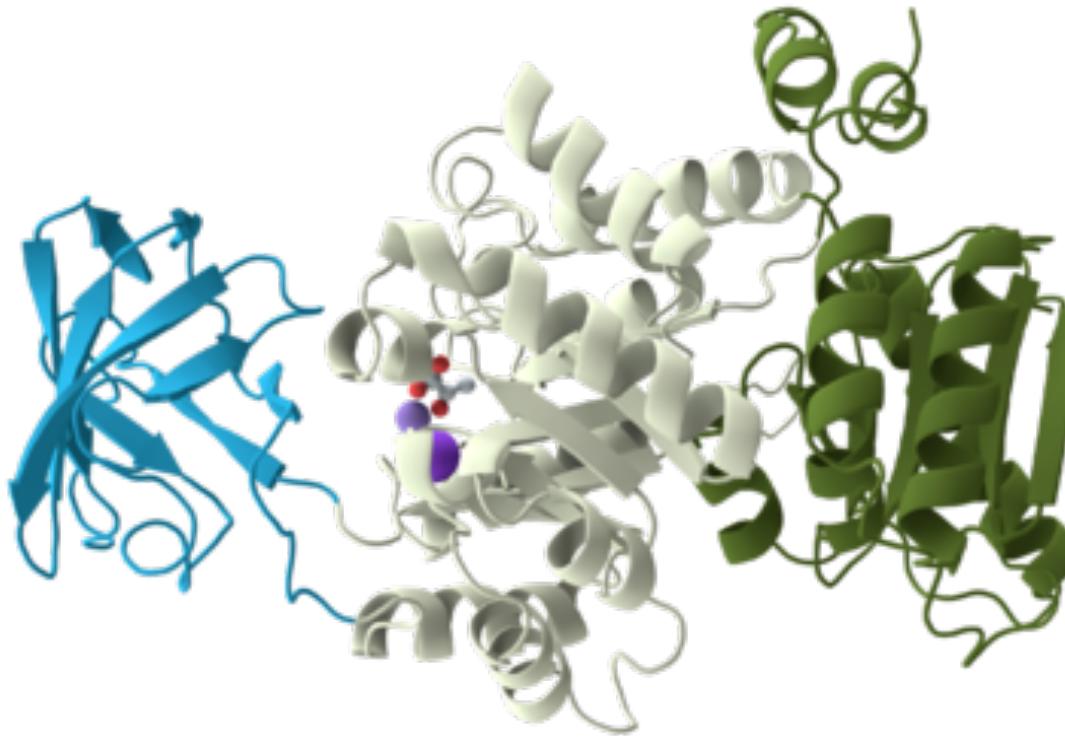
Download marked set

as Protein FA

>1c1|ORF5  
MYPESTGSPARLSLRQTGSPGMIFYSTRYGSPKRQLQFYR  
NLGKSGRLRVSLCLGLTWVTFGGQITDEMAEHMLTAYDNG  
INLFDTAEEVYAAKGAEVVVLGNIIKKKGWRRSSLVITTKIF  
WGGKAETERGLSRKHIIIEGLKASLERLQLEYVVDVFANRP  
DPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYS  
VARQFNLIIPPICEQAEYHMFQREKVEVQLPELFHKIGVGA  
MTWSPLACGIVSGKYDSGIPPPSRSALKGYQWLKDYLSE  
EGRRQQAKLKELOQAIAPERLGCTLPQLAIACLNEGVSSV  
LLGASNQELMENIGAIQVLPKLSSSVHEIDSIILGNKPY  
SKKDYRS

Label	Strand	Frame	Start	Stop	Length (nt)
ORF5	+	3	324	1427	1104   36
ORF3	+	1	1264	1758	495   16
ORF7	-	1	492	103	390   12
ORF11	-	3	910	590	321   10
ORF9	-	3	1384	1130	255   8
ORF12	-	3	325	86	240   7

# Can we recognize functional domains in putative coding regions?



Hints at substrate binding or catalytic activity

DNA, RNA, calcium,  
phosphate, etc.

Glycoslase, methylase, kinase, nuclease,  
lipase, protease, etc.

**Search the Pfam library of HMMs to identify potential functional domains**

EMBL-EBI 

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam 31.0 (March 2017, 16712 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

**QUICK LINKS**

[SEQUENCE SEARCH](#)

[VIEW A PFAM ENTRY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

[KEYWORD SEARCH](#)

[JUMP TO](#)

**ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES**

Paste your protein sequence here to find matching Pfam entries.

**Go Example**

```
METGGRARTGTGTPQPAAPGVWRARPAGGGGGGASSWLLDGNSSWLLCYGFLY  
LALYAQSQSCKPCERTGSCFSGRCVNSTCLCDPGWVGDCQCQHCQGRFKLT  
EPSGYLTDPINPKYKTKTWLIEGYPNAVLRLRFNHATECSWHDHMVY  
DGDSIYAPLIAVLSGLIVPEIRGNETPVPEVTTSGYALLHFFSDAAYNL  
GFNIFYSINSCPNNCNSGHGKCTTSVSPSVQYCECDKYWKGEACDIPYCK  
ANCQSPDHGKGEKLCVNDSWQGPDCSLNVPSTESYWILPNVKPFS  
PSVGRASHKAVLHGKFMWVIGGYTFNYSSFQMVLNYLESSIWNVGTPSR  
GPLQRQYGHSLALYQENIFMYYGRIETNDGNVTDELWVFNIHSQSWSKTP  
TVLGHQGQQYAVEGHSAHIMELDSRDVMIIIFGYSAIYGYTSSIQEYHIS  
SNTLVLPETKGAIVQGGYQHTSVDEITKSIYVHGGYKALPGNKYGLVDD  
LYKYEVTKTWTILKESGFARYLHSASLINGAMLIFFGGNTHNDTLSNGA  
KCFSAFLAYDIACDEWKLPKPNLHRDVNRFGHSAVINGSMYIFGGFS  
SVLNLILVYKPPNCKAFRAFRDEELCKNAGPKIKCWNKNHNCEWESGNTNN  
ILRACKPPKTAASDDRCYRADCASCTANTNGCQWCDDKKCISANSNCNM  
SVKNTYTKCHVRNEQICNKLTSCSKSCLNLCQWDQRQQECQALPAHLCGE  
GWSHIGDACLRVSNSRENNYDNAKLYCENLSNLASLTTSKVEFVLDEIQ  
KYTQKVSPWVGLRKINISYWGDEMSPFTNTLQWLPGEPNDSGFCAYL  
KTEAAVGLKANPCTSMTANGLVCEKPVSPNQARPCKKPCSLRTSCSNCT  
SNGMECMWCSSKTRCVDSNAYIISFPYGGCLEWQTATCSPQNCSGLRTCG  
QCLEQPGCGWCDNPSNTGRGHICEGSSRGPMKLIGMHSEMVLDTNLCPK  
EKNYEWFSFIQCPACQCNGHSTCINNNVCEQCKNLTTGKQCCQDCMPGYYGD  
PTNGQQCTACTCSGHANICHLHTGKCFCTTKGKGDQCLCDSERRYVGN  
PLRGTCYSSLIDYQFTFSLLQEDDRHTAINFIANPEQSNKNLDISINA  
SNNFNLLNTWSVGSTAGTISGEETSIVSKNNIKEYRDSFSFYKEFNRFSNP  
NITFVVVVSFNSWPKIQIAFSQHNTIMDLVQFFFSCFLSLLLVAAV  
VWKIKQTCAWSRREQLLERQQMASRPFAVDVALEVGAEQTEFLRGPL  
EGAKPKIAIEPCAGNRRAAVLTVFLCLPRGSSGAPPQGSGLIASALIDI  
SQQKASDSDKTKTSGVRNRKHLSTRQGTCV
```

This search will use an E-value of 1.0. You can set your own search parameters and perform a range of other searches [here](#).

# Example Pfam report illustrating modular domain architecture

← → ⌂ pfam.xfam.org/search/sequence

EMBL-EBI 

**HOME** | **SEARCH** | **BROWSE** | **FTP** | **HELP** | **ABOUT**

**Pfam**  
keyword search **Go**

## Sequence search results

[Show](#) the detailed description of this results page.

We found **9** Pfam-A matches to your search sequence (**all** significant)



[Show](#) the search options and sequence that you submitted.

[Return](#) to the search form to look for Pfam domains on a new sequence.

### Significant Pfam-A Matches

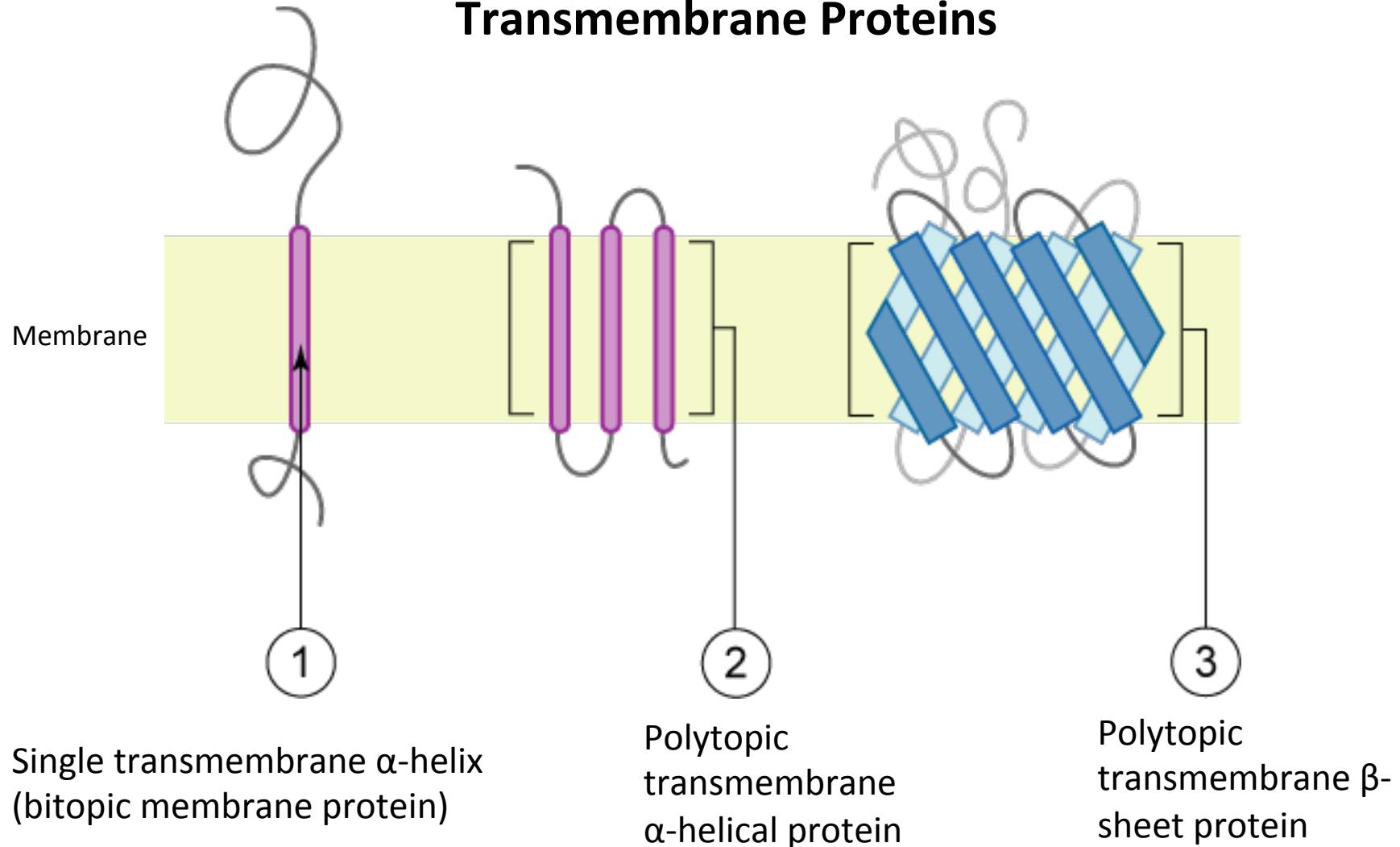
[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	<a href="#">Show/hide alignment</a>
				Start	End	Start	End	From	To					
<a href="#">CUB</a>	CUB domain	Domain	<a href="#">CL0164</a>	93	206	93	206	1	110	110	42.2	7.7e-11	n/a	<a href="#">Show</a>
<a href="#">EGF_2</a>	EGF-like domain	Domain	<a href="#">CL0001</a>	249	280	249	280	1	32	32	22.5	0.0001	n/a	<a href="#">Show</a>
<a href="#">Kelch_5</a>	Kelch motif	Repeat	<a href="#">CL0186</a>	351	393	352	392	2	41	42	33.7	2.2e-08	n/a	<a href="#">Show</a>
<a href="#">Kelch_4</a>	Galactose oxidase, central domain	Repeat	<a href="#">CL0186</a>	466	518	468	514	3	44	49	20.6	0.0003	n/a	<a href="#">Show</a>
<a href="#">Kelch_1</a>	Kelch motif	Repeat	<a href="#">CL0186</a>	520	574	520	573	1	45	46	20.0	0.00033	n/a	<a href="#">Show</a>
<a href="#">Kelch_5</a>	Kelch motif	Repeat	<a href="#">CL0186</a>	579	614	581	613	5	40	42	25.3	9.7e-06	n/a	<a href="#">Show</a>
<a href="#">Lectin_C</a>	Lectin C-type domain	Domain	<a href="#">CL0056</a>	765	874	766	874	2	108	108	70.2	2e-19	n/a	<a href="#">Show</a>
<a href="#">PSI</a>	Plexin repeat	Family	<a href="#">CL0630</a>	889	939	890	938	2	50	51	27.8	2.5e-06	n/a	<a href="#">Show</a>
<a href="#">PSI</a>	Plexin repeat	Family	<a href="#">CL0630</a>	942	1012	942	1012	1	51	51	50.0	2.9e-13	n/a	<a href="#">Show</a>

Comments or questions on the site? Send a mail to [pfam-help@ebi.ac.uk](mailto:pfam-help@ebi.ac.uk).

**European Molecular Biology Laboratory**

# Transmembrane Proteins



# Using TMHMM to identify putative transmembrane proteins

www.cbs.dtu.dk/services/TMHMM/

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

**CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS**

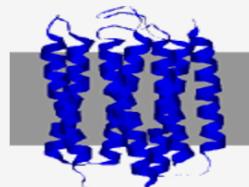
EVENTS NEWS RESEARCH GROUPS CBS PREDICTION SERVERS CBS DATA SETS PUBLICATIONS EDUCATION

STAFF CONTACT ABOUT CBS INTERNAL CBS BIOINFORMATICS TOOLS CBS COURSES OTHER BIOINFORMATICS LINKS

CBS > CBS Prediction Servers >> TMHMM

**TMHMM Server v. 2.0**

**Prediction of transmembrane helices in proteins**



**Instructions**

**SUBMISSION**

Submission of a local file in **FASTA** format (HTML 3.0 or higher)

No file chosen

OR by pasting sequence(s) in **FASTA** format:

```
MEILCEDNTSLSSIPNSLMQVDGDSGLYRNDFNNSRDANSSDASNWTDGENRTNLSEGVLPPTCLSIHLQEKNWSALLTAVVIIAGNIVMAVSLEKKLQNATNYFLMSLAIDMLLGFLVMPVSMILTYGYRWPLPSKLCAVWIYLDVLFSTASIMHLCaisLDRYVAIQNPPIHHSRFNSRTKAFLKIIAVWTISVGVSMPVIPVFLQDDSKVFQGSCLADDNFVLIGSFVAFFIPLTMVITYFLTIKSLQKEATLCVSDLSTRAKLASFSFL
```

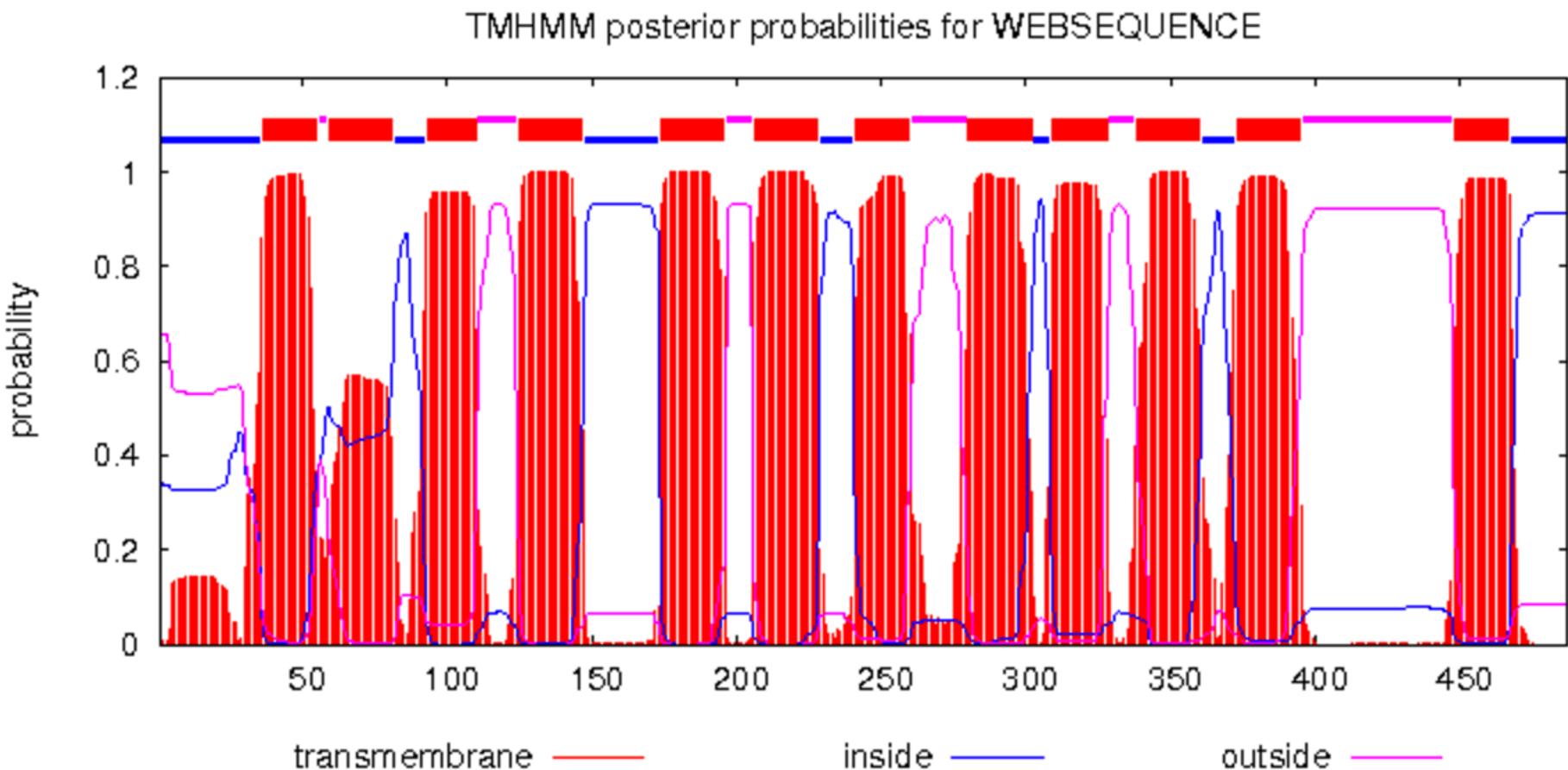
**Output format:**

Extensive, with graphics  
 Extensive, no graphics  
 One line per protein

**Other options:**

Use old model (version 1)

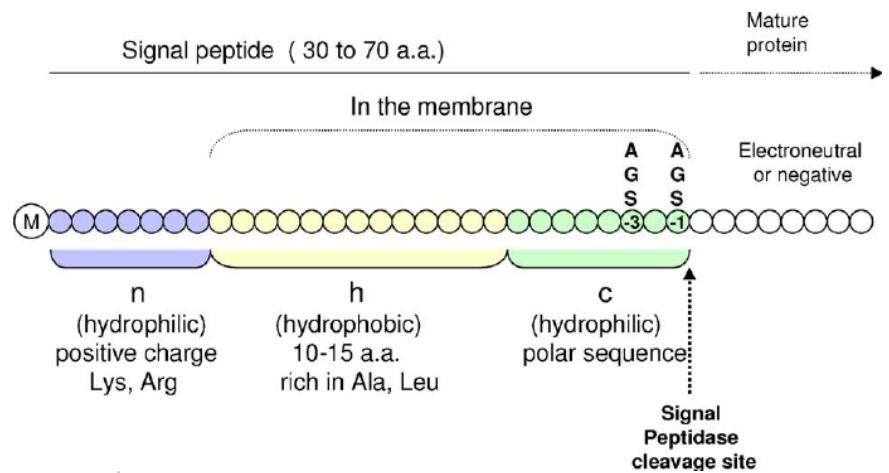
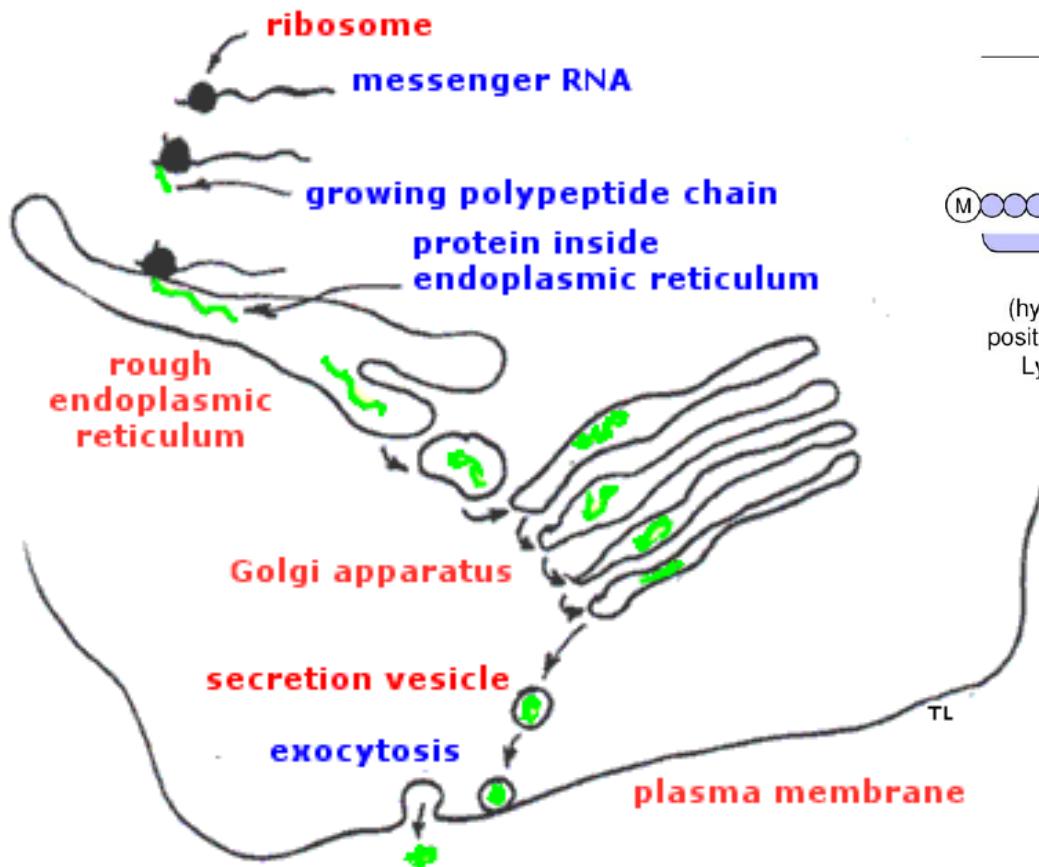
# Trans-membrane Domains via TmHMM



Topology=i36-55o59-81i93-110o125-147i174-196o206-228i241-260o280-302i309-328o338-360i373-395o448-467i

<http://www.cbs.dtu.dk/services/TMHMM/>

# Predicting Secreted Proteins



(from: Vaccine 23(15):1770-8)

(from: <https://courses.washington.edu/conj/cell/secretion.htm>)

# SignalP: Prediction of N-terminal signal peptides

## (predict secreted proteins)

www.cbs.dtu.dk/services/SignalP/

The navigation menu is a horizontal bar with colored boxes for different sections: CENTERFOR BIOLOGICAL SEQUENCES (yellow), EVENTS (yellow), NEWS (green), RESEARCH GROUPS (orange), CBS PREDICTION SERVERS (pink), CBS DATA SETS (purple), PUBLICATIONS (blue), and EDUCATION (dark purple). Below the main menu, there are sub-sections: STAFF (green), CONTACT (green), ABOUT CBS (orange), INTERNAL (red), CBS BIOINFORMATICS TOOLS (purple), CBS COURSES (blue), and OTHER BIOINFORMATICS LINKS (dark purple). A search bar and a logo for the Center for Biological Sequence Analysis are also present.

CBS > CBS Prediction Servers > SignalP

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

## SignalP 4.1 Server

SignalP 4.1 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks.

View the [version history](#) of this server. All the previous versions are available online, for comparison and reference.

**NEW:** The portable version of SignalP 4.1, previously only available for Mac (Darwin), Linux, and IRIX, is now also available for Windows systems. Academic users: select the "CYGWIN" option at the [download page](#). [Cygwin](#) or [MobaXterm](#) is required to install SignalP under Windows. For details, read the [installation instructions](#).



### SUBMISSION

Paste a single amino acid sequence or several sequences in [FASTA](#) format into the field below:

```
MHPAVFLSLPDLRCSLLLLLTVFTPVTTETSLDTENIDEILNNADVALVNFYADWCRFSQMLHPIFEASDVIKEEFPNENQVVFARVDCDQHSDIAQRYRISKYPTLKLFRNGMM  
KREYRGQRSVKALADYIRQQKSQDPIQEIRDALAEITLDRSKRNIIGYFEQKDSNDNYRVFERVANILHDDCAFLSAFGDVSCKPERYSGDNIIYKPPGHSAPDMVYLGAMTNFDVTYNIQ  
DKCVPVLVREITFENGELTEEGLPFLILFHMKEDTESLEIFQNVARQLISEKGTTNFLADCDKFRHPLLHIQKTPADCPVIAIDSFRHMYVFGDFKDVLIPGKLKQFVFDLHSGKLHREF  
HHGPDPDTAPGEQAQDVASSPPESSFQKLAPSEYRTLLRDRDEL
```

Submit a file in [FASTA](#) format directly from your local disk:

Choose File | No file chosen

#### Organism group ([explain](#))

- Eukaryotes
- Gram-negative bacteria
- Gram-positive bacteria

#### D-cutoff values ([explain](#))

- Default (optimized for correlation)
- Sensitive (reproduce SignalP 3.0's sensitivity)
- User defined:
  - 0.4 D-cutoff for SignalP-noTM networks
  - 0.5 D-cutoff for SignalP-TM networks

#### Graphics output ([explain](#))

- No graphics
- PNG (inline)
- PNG (inline) and EPS (as links)

#### Output format ([explain](#))

- Standard
- Short (no graphics)
- Long
- All - SignalP-noTM and SignalP-TM output (no graphics)

#### Method ([explain](#))

- Input sequences may include TM regions
- Input sequences do not include TM regions

#### Positional limits ([explain](#))

- Minimal predicted signal peptide length. *Default: 10*
- N-terminal truncation of input sequence (0 means no truncation).  
*Default: Truncate sequence to a length of 70 aa*

# Example SignalP predicted signal peptide

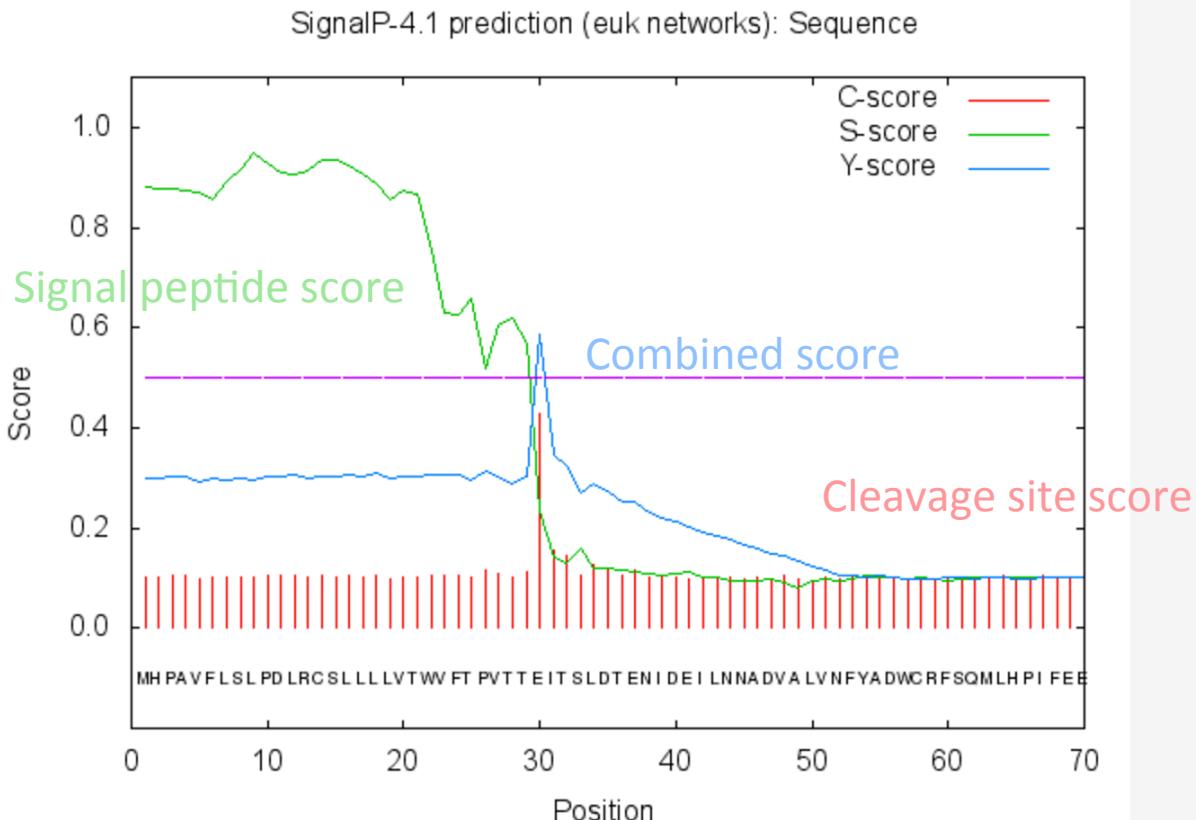
← → ⌂ ⓘ www.cbs.dtu.dk/cgi-bin/webface2.fcgi?jobid=58FFF29C00005F854B357EEA&w... ☆



## SignalP 4.1 Server - prediction results

Technical University of Denmark

```
# SignalP-4.1 euk predictions  
>Sequence
```



# Transcriptome-scale functional annotation using Trinotate



## Trinotate: Transcriptome Functional Annotation and Analysis

# Trinotate



TransDecoder



eggNOG  
version 3.0



Pfam



TMHMM

SignalP



RNA-Seq → Trinity → Transcripts/Proteins → Functional Data → Discovery

There's no substitute for experimentally validating protein functions



## Transcriptome Assembly is Just the End of the Beginning...

NATURE PROTOCOLS | PROTOCOL

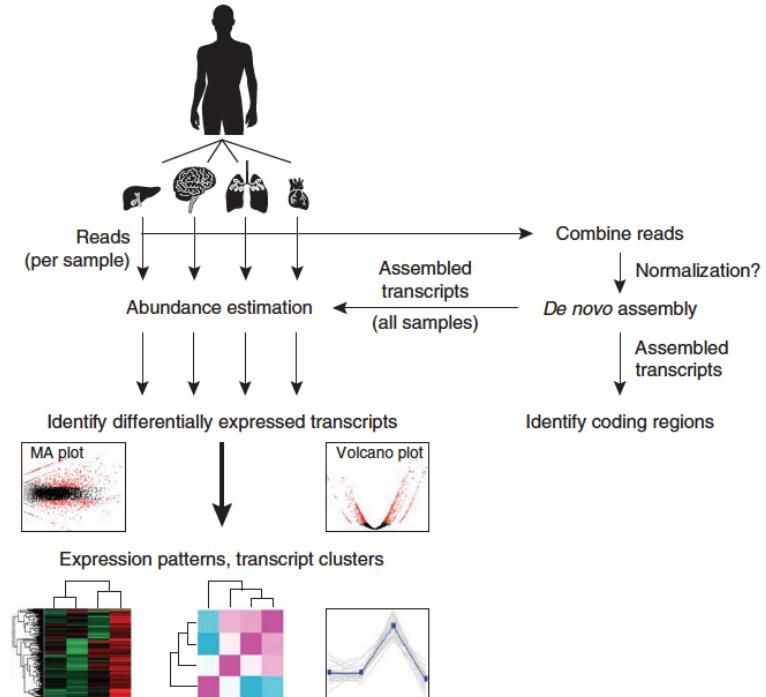
*De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

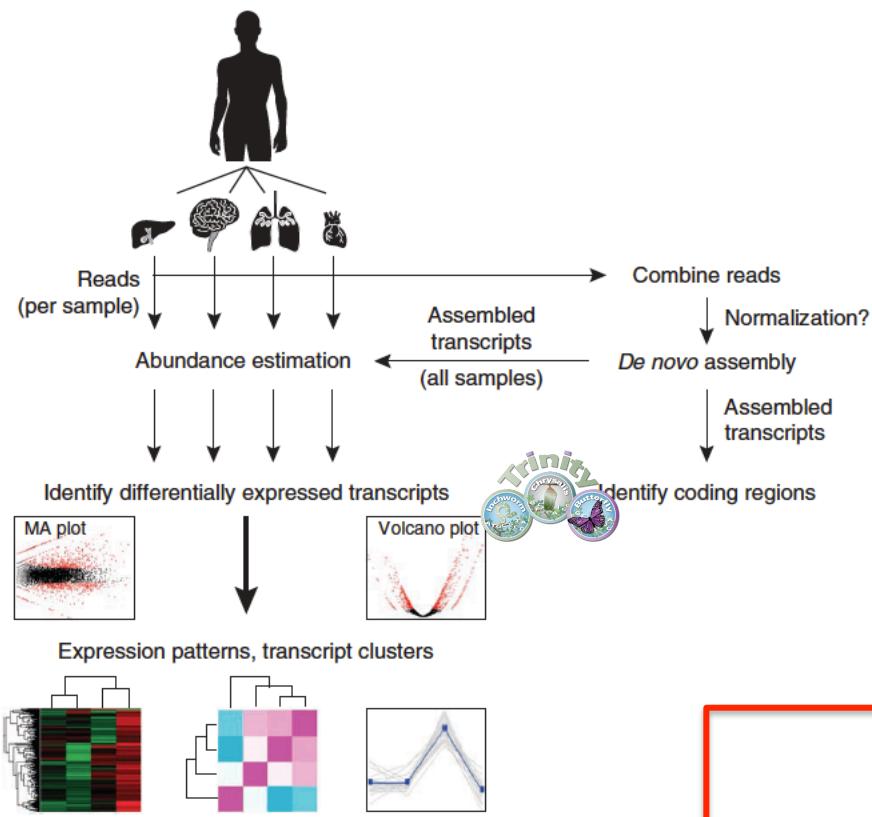
[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

*Nature Protocols* 8, 1494–1512 (2013) | doi:10.1038/nprot.2013.084

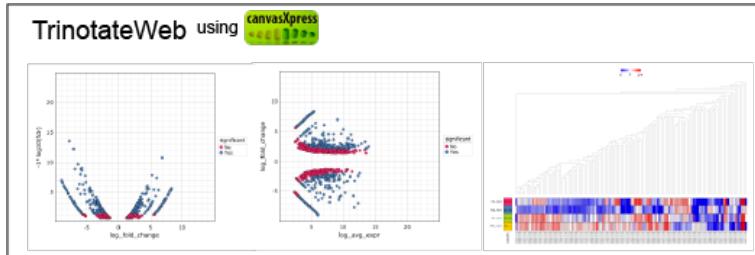
Published online 11 July 2013



# Trinity Framework for De novo Transcriptome Assembly and Analysis



Trinotate



Bioconductor,  
& Trinity

Module

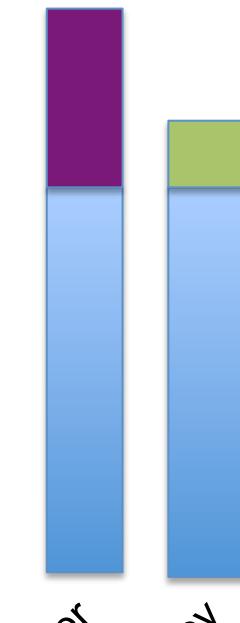
bioinformatics.ca

# Trinotate Functional Annotation Lab

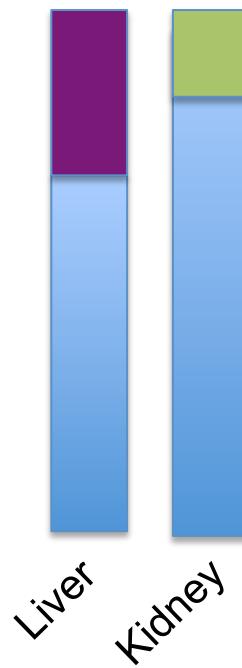


# Why cross-sample normalization is important

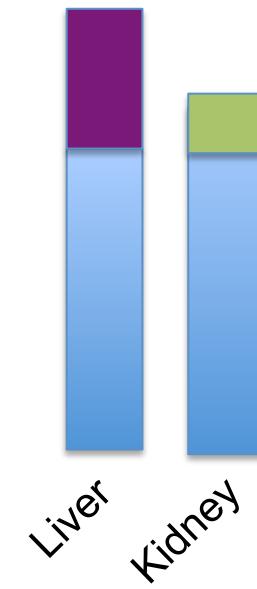
Absolute RNA quantities per cell



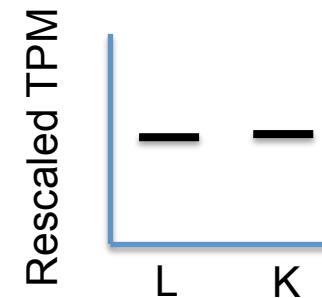
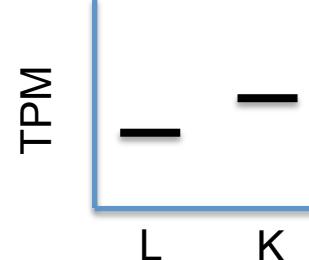
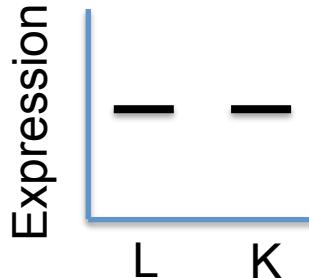
Measured relative abundance via RNA-Seq



Cross-sample normalized (rescaled) relative abundance

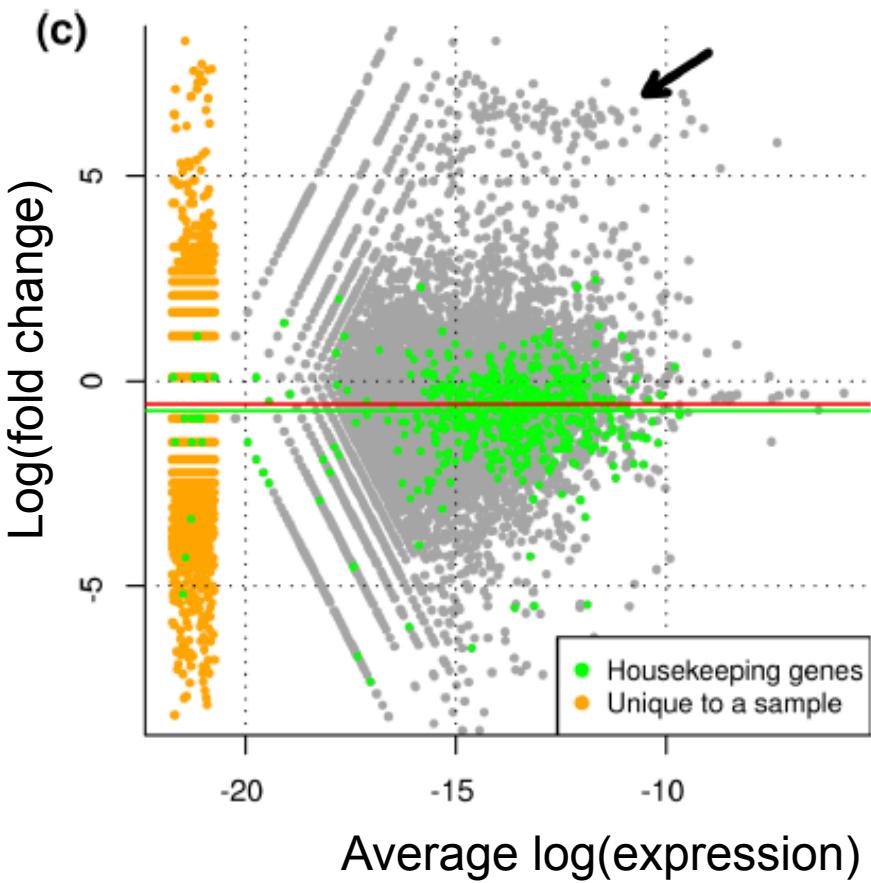
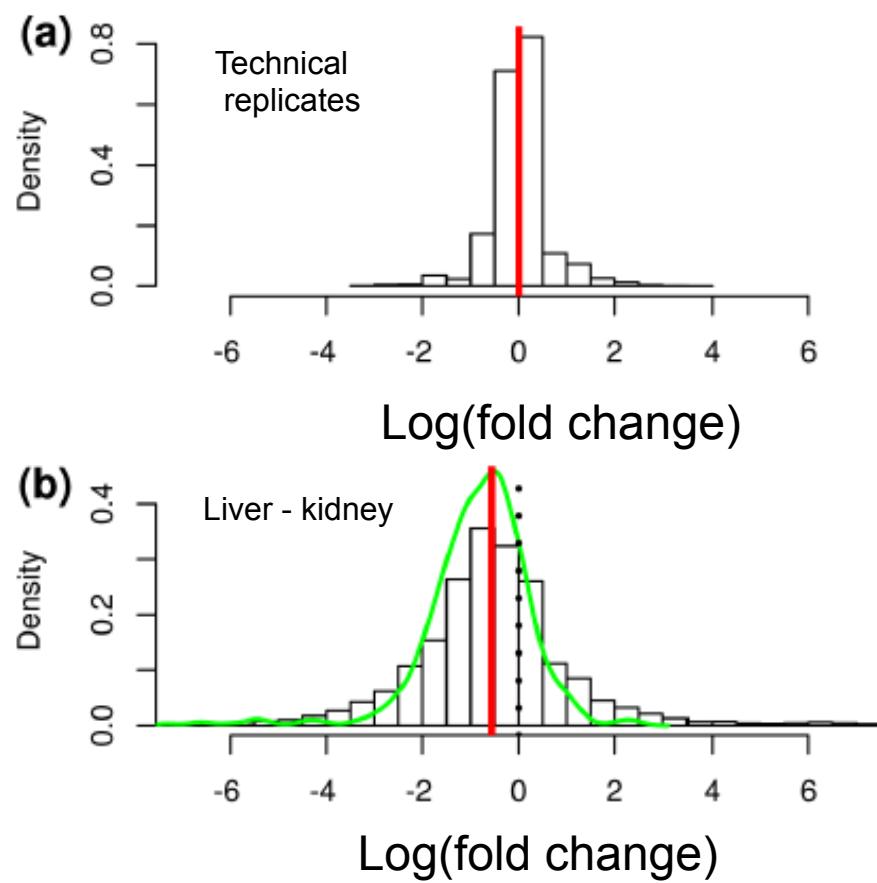


eg. Some housekeeping gene's expression level:



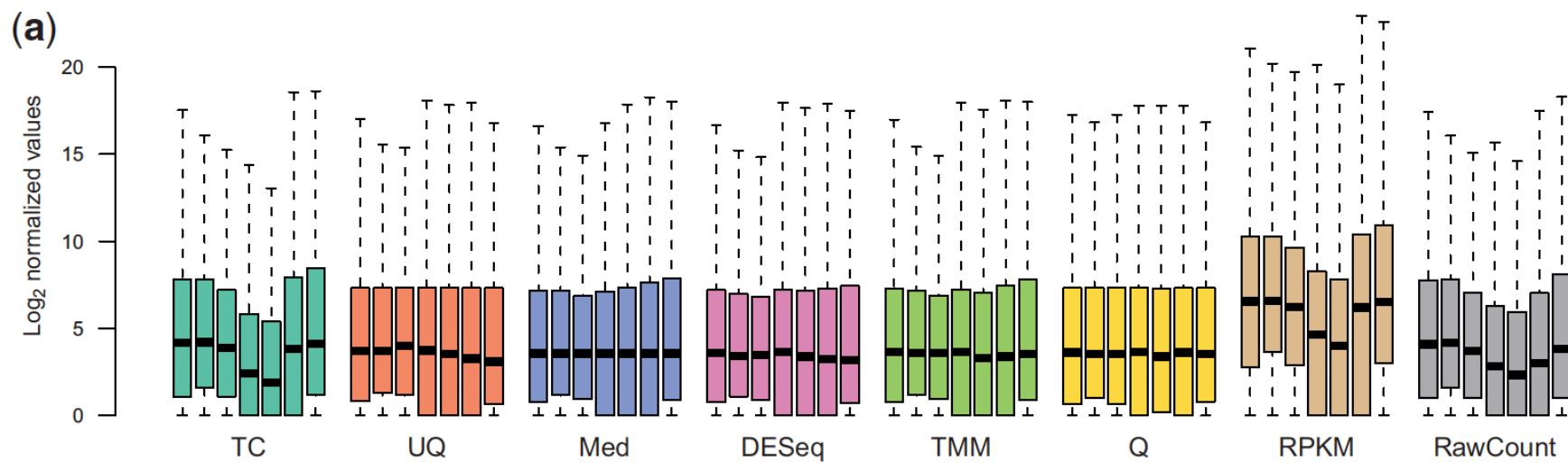
# Cross-sample Normalization Required Otherwise, housekeeping genes look diff expressed due to sample composition differences

Subset of genes  
highly expressed in  
liver



**Figure 1 Normalization is required for RNA-seq data.** Data from [6] comparing log ratios of **(a)** technical replicates and **(b)** liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of the housekeeping genes. **(c)** An M versus A plot comparing liver and kidney shows a clear offset from zero. Green points indicate 545 housekeeping genes, while the green line signifies the median log-ratio of the housekeeping genes. The red line shows the estimated TMM normalization factor. The smear of orange points highlights the genes that were observed in only one of the liver or kidney samples. This indicates a bias in log-fold-changes.

# Normalization methods for Illumina high-throughput RNA sequencing data analysis.



From “A comprehensive evaluation of normalization methods for Illumina high throughput RNA sequencing data analysis” Brief Bioinform. 2013 Nov;14(6):671-83

<http://www.ncbi.nlm.nih.gov/pubmed/22988256>