



# Canadian Bioinformatics Workshops

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

[bioinformaticsdotca.github.io](http://bioinformaticsdotca.github.io)

Supported by



Creative Commons

This page is available in the following languages:

Afrikaans ດັວນລາວ Català Dansk Deutsch Elληνες 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 עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu  
Nederlands Norsk Sesotho sa Leboa polski Português română slovenšči jezik српски srpski (latinica) Sotho svenska  
中文 華語 (台灣) isiZulu

 creative  
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](#)

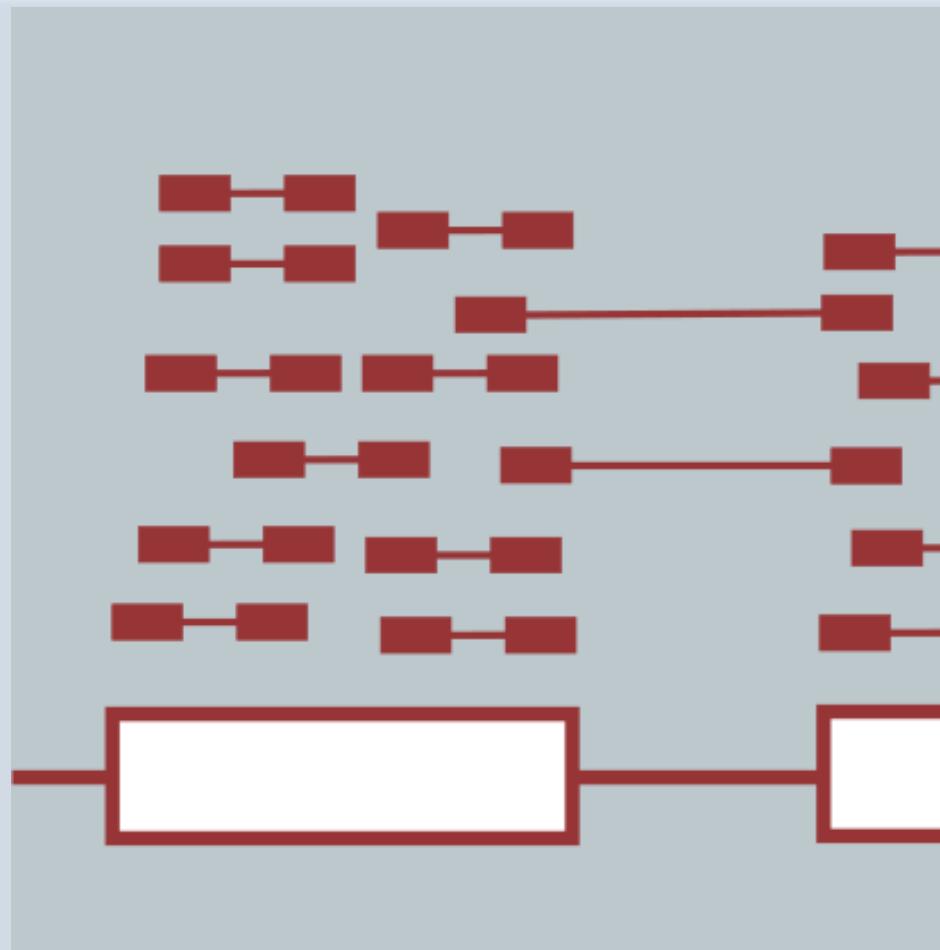
Learn how to distribute your work using this licence

# Functional Annotation and Analysis of Transcripts

Brian Haas

Informatics for RNA-Seq Analysis

June 11-13, 2019



# 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

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGCCCTGGCGCTGCTGCACCAACTCCTGTTGGGCCGTGGTCCT  
TGGAGGCATGCAGTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGAA  
TCTCCCG  
AAAGAC  
GGCTTC  
TGACCT  
GAAAAAC  
TTGTCA  
TCGAC  
TCCCA  
CCTGG  
CCTAA  
TGCTG  
CAGCC  
TTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTGCCAACCGCCCAGACCCCAACACGCCATGGAAGAGACCGTGCAGGGCCA  
TGACCCATGTCATCAACCAGGGATGGCCATGTACTGGGGCACATCAGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTCGCAGTTCAACCTGATCCGCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCTCTGGCGTGCAGGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCTACTCCAGAGCCTCCCTGAAGGGTACCAAGCTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCAGCAGGCCAAGCTGAAGGAACCTGC

Can we gather hints of biological function  
from sequence?

# Methods used to predict function from sequence

- Sequence homology

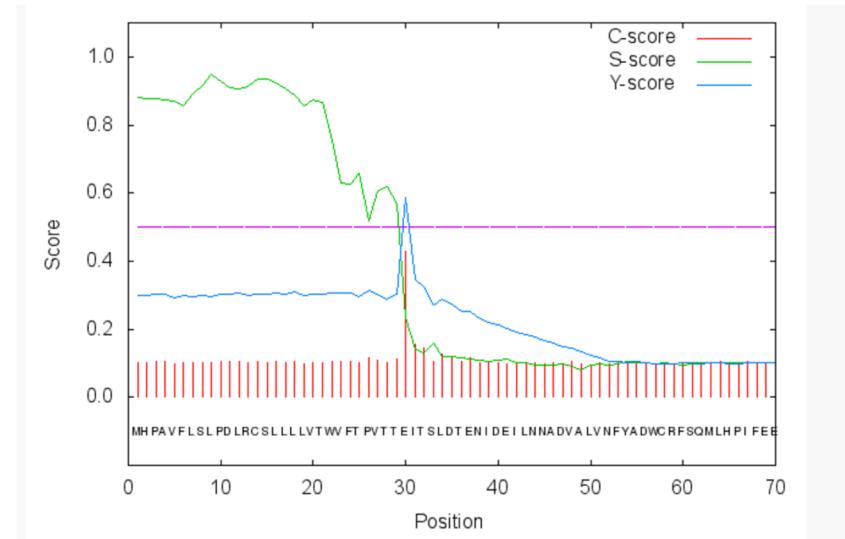
Searching protein database for sequence similarity

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**

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

## 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](#)

**NEWS**

New BLAST Results to become default on Aug 1, 2019  
To help instructors integrate the new design into their lesson plans, we are making the change before the fall semester.  
Thu, 30 May 2019 14:00:00 EST [More BLAST news...](#)

### Web BLAST

**Nucleotide BLAST**  
nucleotide ► nucleotide

**blastx**  
translated nucleotide ► protein

**tblastn**  
protein ► translated nucleotide

**Protein BLAST**  
protein ► protein

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

← → C 🔒 https://www.uniprot.org



UniProt

UniProtKB ▾

Advanced ▾

Search

BLAST Align Retrieve/ID mapping Peptide search

Help Contact

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  
(560,292)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL  
(158,257,522)

Automatically annotated and not reviewed.

Records that await full manual

(as of June, 2019)

## UniRef

Sequence clusters



## UniParc

Sequence archive



## Proteomes



## Supporting data

Literature citations



Cross-ref. databases



Taxonomy



Diseases



Subcellular locations



Keywords



## News



### Forthcoming changes

There are currently no changes planned

### UniProt release 2019\_05

Love's Labour (nearly) Lost

### UniProt release 2019\_04

A pox on your messenger | Removal of the cross-references to HOVERGEN, ProteinModelPortal and UniGene

### News archive

M

Text search

Our basic text search allows you to search all the



UniProt data

Download latest release

Get the UniProt data

Protein spotlight



Twisting Fate

May 2019

a

# 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 Entry information Miscellaneous

Function

Protein | Fructosamine-3-kinase  
Gene | FN3K  
Organism | Homo sapiens (Human)  
Status | Reviewed - Annotation score: ●●●●○ - Experimental evidence at protein level<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...

Keywords<sup>i</sup>

Molecular Kinase Transferase

Gene Ontology (GO): Structured vocabulary for defining molecular functions, biological processes, and cellular components.

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

www.ebi.ac.uk/QuickGO/GTerm?id=GO:0030387#te...

Quick GO Click for example search Search! Web Services Dataset Term Basket: 0

Term Information Ancestor Chart Child Terms Protein Annotation Co-occurring Terms Change Log

This chart is interactive; you can click on the term boxes and legend for more information.

Display

Gene Ontology terms are organized into a directed acyclic graph. Terms are organized from general (top) to more specific (bottom).

The GO structure enables computations such as exploring function enrichment among sets of transcripts.

QuickGO - http://www.ebi.ac.uk/QuickGO

Legend:

- A → B: Is a
- A → B: Part of
- A → B: Regulates
- A → B: Positively regulates
- A → B: Negatively regulates
- A → B: Occurs in
- A → B: Capable of
- A → B: Capable of part of

# 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?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGCCCTGGCGCTGCTGCACCAACTCCTGTTGGGCCGTGGTCCT  
TGGAGGCATGCAGTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTACAGGAATCTGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTGATAACGGGGAGGTCTACGCTGCTG  
GAAAAGCTGAAGTGGTATTAGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAAGATCTCTGGGTGGAAAAGCGGAGACTGAGAGAGGGCTTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTGCCAACCGCCCAGACCCCAACACGCCATGGAAGAGACCGTGCAGGGCCA  
TGACCCATGTCATCAACCAGGGATGGCCATGTACTGGGGCACATCAGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTGGCAGTTCAACCTGATCCGCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCTCTGGCGTGCAGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCTACTCCAGAGCCTCCCTGAAGGGTACCAAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCAGCAGGCCAAGCTGAAGGAAGTGC

# Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGCCCTGGCGCTGCTGCACCAACTCCTGTTGGGCCGTGGTCCT  
TGGAGGCATGCAGTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTACAGGAATCTGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTGATAACGGGGAGGTCTACGCTGCTG  
GAAAAGCTGAAGTGGTATTAGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAAGATCTCTGGGTGGAAAAGCGGAGACTGAGAGAGGGCTTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTGCCAACCGCCCAGACCCCAACACGCCATGGAAGAGACCGTGCAGGGCCA  
TGACCCATGTCATCAACCAGGGATGGCCATGTACTGGGGCACATCAGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTGGCAGTTCAACCTGATCCGCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCTCTGGCGTGCAGCTCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCTACTCCAGAGCCTCCCTGAAGGGTACCAAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCAGCAGGCCAAGCTGAAGGAACCTGC

# Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCAC**ATGGGGCCCTGGCGCTTGCTGCACCAACTCCTGTTGGGCCGTGGTCCT**  
**TGGAGGCATGCAGTCAGCAGACAGTGA**CTCAGCCATCCACCCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTACAGGAATCTGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTGATAACGGGGAGGTCTACGCTGCTG  
GAAAAGCTGAAGTGGTATTAGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAAGATCTCTGGGTGGAAAAGCGGAGACTGAGAGAGGGCTTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTGCCAACCGCCCAGACCCCAACACGCCATGGAAGAGACCGTGCAGGGCCA  
TGACCCATGTCATCAACCAGGGATGGCCATGTACTGGGGCACATCAGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTGGCAGTTCAACCTGATCCGCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCTCTGGCGTGCAGCTCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCTACTCCAGAGCCTCCCTGAAGGGTACCAAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCAGCAGGCCAAGCTGAAGGAAGTGC

# 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:

```
GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGCCCTAGGGCCCTGGTTTGTAGTCAGTGCA  
GTTGCTGCACATGGGCCCTGGCGCTTGCTGCACCAACTTCCCTGGGCCCTGGTCC  
TGGAGGCATGCAGTTCAAGCAGACTGACTCAGCCATCCACCCAACATGCGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCATGTATCCGAATCAACCACGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCA  
AAAGACAGCTCCAGTTTACAGGAATCTGGGCAAATCTGGCCTTGGGTCTCTGGCTGG  
GGCTTGGAACATGGGTGACCTCGGGGGCAGATCAGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTCGATACGGCGGAGGTCTACGCTGCTG
```

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 the NCBI logo, 'Resources', 'How To', and 'Sign in to NCBI'. The main title 'ORFfinder' is on the left, with a dropdown menu set to 'PubMed'. A search bar and a 'Search' button are on the right. The main content area is titled 'Open Reading Frame Viewer'. It displays a sequence viewer with a green track at the top labeled '1: 1..1.8K (1.8Kbp)'. Below it, several red arrows indicate the direction of 12 predicted ORFs (ORF4 through ORF12). A large black box contains the text: 'Sequence ORFs can appear in random sequence – so further analysis is required'. Above this box, status information says 'ORFs found: 12 Genetic code: 1 Start codon: 'ATG' only'. The bottom of the viewer shows a progress bar and the word 'Ready'.

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

Add six-frame translation track

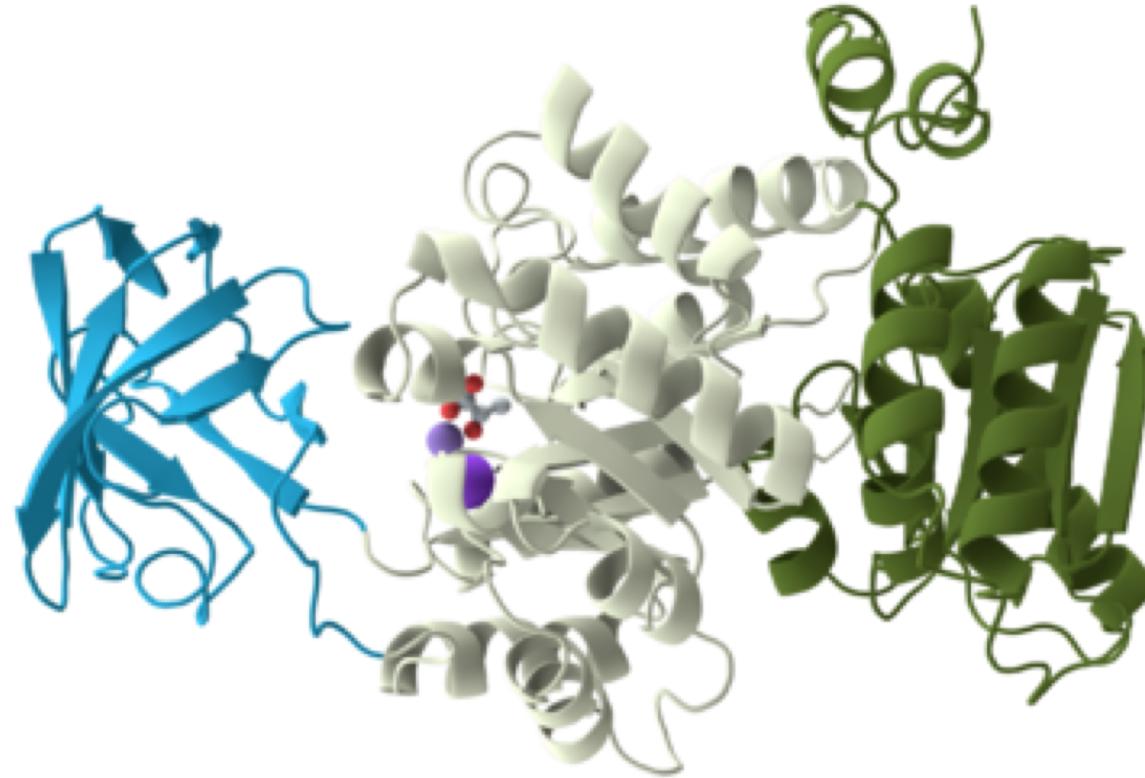
ORF5 (367 aa)    Display ORF as...    Mark

```
>1cl1|ORF5
MYPESTTGSPARLSLRQTRGSPGMIVSTRYGPSPKRQLQFYR
NLGKSGRLRVSLGLGTWVTFGGQITDEMAEHLMTLAYDNG
INLFDTAEVYAAKGKAEVVLGNIIKKKGWRRSSLVITTKIF
WGGKAETERGLSRKHIIEGLKASLERLQLEYVDVVFANRP
DPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYS
VARQFNLIPICEQAEEYHMFQREKVEVQLPELFHKIGVGA
MTWSPLACGIVSGKYDGSGLPPYSRASLKGYQWLKDILSE
EGRRQQAKLKEQAIAPERLGCTLPQLAIWCLRNEGVS
LLGASNAEQLMENIGAIQVLPKLSSSIVHEIDSILGNKPY
SKKDYRS
```

Mark subset...    Marked: 0    Download marked set as Protein FA

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
ORF8	-	2	848	618	231   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

The screenshot shows the Pfam 31.0 (March 2017, 16712 entries) homepage. At the top, there's a navigation bar with links for HOME, SEARCH, BROWSE, FTP, HELP, and ABOUT. The EMBL-EBI logo is on the left, and the Pfam logo with a search bar is on the right. Below the header, a main section titled "Pfam 31.0 (March 2017, 16712 entries)" explains that the database contains multiple sequence alignments and hidden Markov models (HMMs). A "More..." link is present.

**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.

```
METGGARTGTGTPQPAAPGWRARPAAGGGGGGASSWLLDGNSWLLCYGFLY  
LALYAQVSQSKPCERTGSFCGVNSTCLCDPGWVGQDCQHCQGRFKLT  
EPSGYLTDPINVKYKTKCTWLIEGYPNAVLRLRFNHFATECSWHDHMYV  
DGDSIYAPILAIVLGLIVPEIRGNETVPEVVTTSYALLHFSDAYNL  
GFNIFYSINSCPNNCSGHGKCTTSVSPSQVYCECDKYWKGEACDIPYCK  
ANCNGSPDHGYCDLTGEKLCVCNDSWQGPDCSLNPSTESYWILPNVKPSF  
PSVGRASHKSLVHGKFMWVIGGYTFNYSFQMVLYNLESSIWNVGTPSR  
GPLQRYGHSLALYQEENIFMVGRIETNDGNTDELWVNFIHSQSWSKTP  
TVLGHGQQYAVEGHSAHIMELSDRVMIIFGYSAIYGTSSIQEYHIS  
SNTWLPETKGAVQGGYGHTSVYDEITKSIYHGGYKALPGNKYGLVDD  
LYKYEVNKTWTILKESGFRYLHSAVLINGAMLIFGGNTHTNDTLSNGA  
KCFSAFLAYDIACDEWKLKPKNLHRDVNRFGHSAVINGSMYIFGGFS  
SVLNDILVYKPPNCKAFRDEELCKNAGPGIKCVWNKNHCESWESGNNTN  
ILRAKCPPDKTADASDDRCYRYADCASCTANTNGCQWCDDKKCISANNCNM  
SVKNTKCHVRNEQICNKLTSCKSCSLNQNCWDQRQQECQALPAHLCGE  
GWSHIGDACLRVNSSRENYDNALKYCYNLSGNLASLTTSKVEFVLDI  
KYTQQKVSPWVGLRKINISYWGWEDEMSPFTNTTLQWLPGEPDNSGFCAYL  
ERAAGVGLKANPCTSMANGLVCEKPVVSPNQNARPCKPKCSRSTSNCN  
SNGMECMWCSSTKRCVDSNAYISFPYGGCLEWQTATCSPQNCSGLRTCG  
QCLEQPGCGWCNDPSNTGRGHIEGSSRGPMKIGMHSEMVLDTNLCPK  
EKNYEWSFIQCPACQCNHGHTCINNNVCEQCKNLTGKQCDCMPGYYGD  
PTNGQQCTACTCSGHANICLHTGKCFCTTKGIKGDQCLCDSENRYVGN  
PLRGTCTYSSLIDYQFTFSILLQEDDRHHTAINFIANPEQSQNKNLDISINA  
SNNFNLNITWSVGSTAGTISGEETSVSKNNIKEYRDSFSYEKFNFRSNP  
NITFYVVSNFSWPKIQIAFSQHNTIMDLVQFFVTFSCFLSLLLVA  
VWKIKQTCAWSRRREQLLRERQQMASRPFASVDFALEVGAEQTEFLRGPL  
EGAPKPIAIEPCAGNRAAVLTFLCLPRGSSGAPPPGQSGLAIASALIDI  
SQQKASDSKDTSVGRNRKHLSTRQGTCV
```

Go      Example

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

← → C ⓘ pfam.xfam.org/search/sequence

EMBL-EBI 

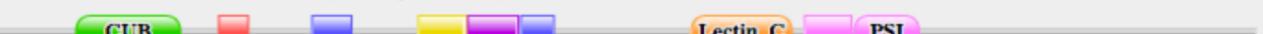
[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)

Pfam 

## 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

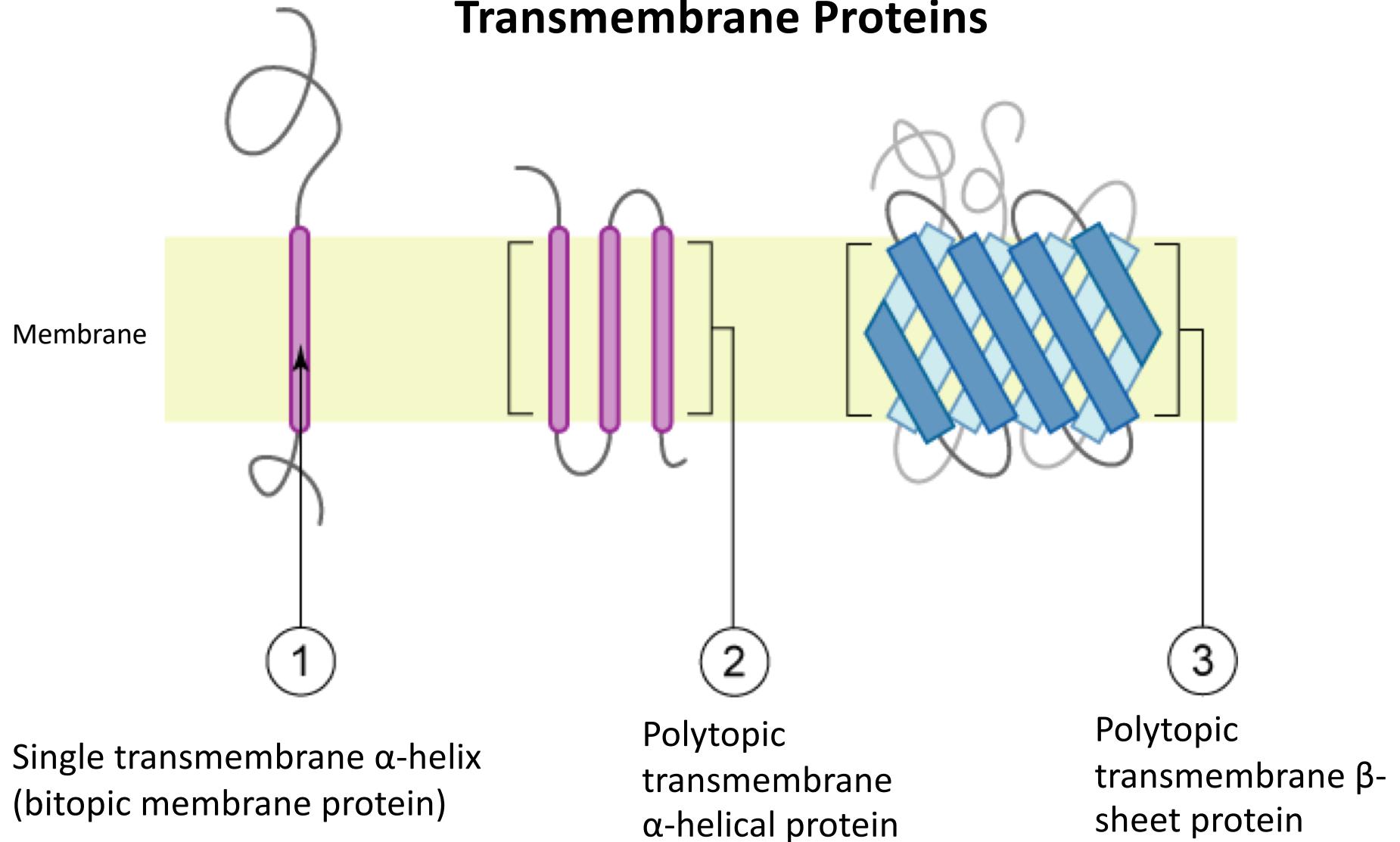
[Show](#) or [hide](#) all alignments.

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



Single transmembrane  $\alpha$ -helix  
(bitopic membrane protein)

Polytopic  
transmembrane  
 $\alpha$ -helical protein

Polytopic  
transmembrane  $\beta$ -  
sheet protein

# Using TMHMM to identify putative transmembrane proteins

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

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

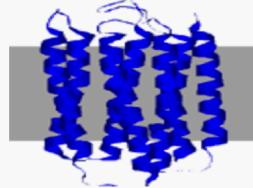
EVENTS NEWS RESEARCH GROUPS CBS PREDICTION SERVERS CBS DATA SETS PUBLICATIONS EDUCATION

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)

Choose File No file chosen

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

```
MEILCEDNTSLSSIPNSLMQVGDGSGLYRNDFNSRDANSSDASNWTIDGENRTNLSEG  
YLPPTCLSLHLQEKNSALLTAVVIIAGNIVMAVSLEKKLQNATNYFLMSLAIADMLL  
GFLMPVPSMLTLYGYRWPLPSKLCAWIYLDVLFSTASIMHLC AISLDRYVAIQNPIHHSR  
FNSRTKAFLKIIAVWTISVGVSMPVIPVFGLQDDSKVFKQGSCLADDNFVLIGSFVAFFIPLTI  
MVITYFLTIKSLQKEATLCVSDLSTRAKLASFSFL
```

**Output format:**

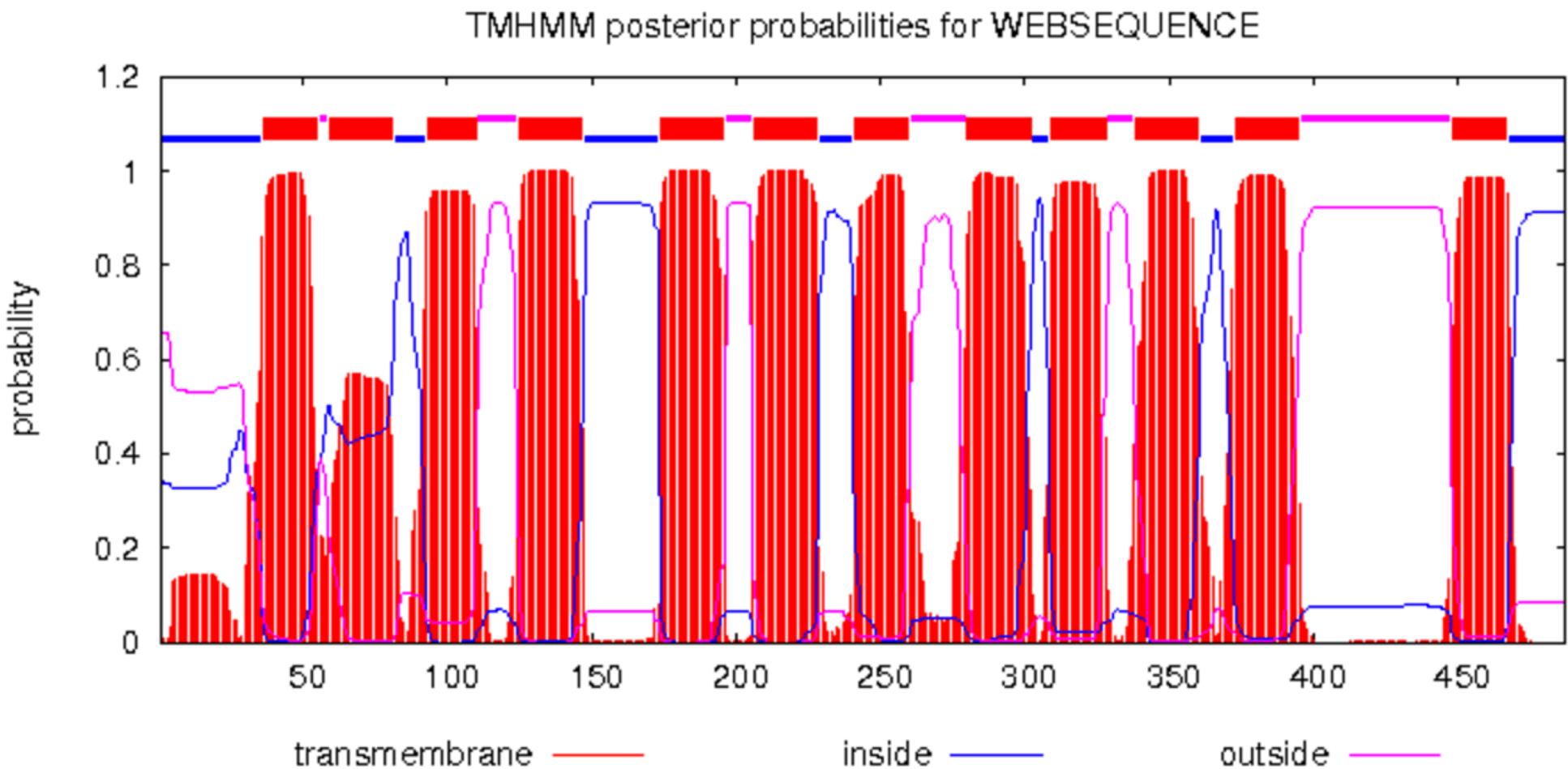
Extensive, with graphics  
 Extensive, no graphics  
 One line per protein

**Other options:**

Use old model (version 1)

Submit  Clear

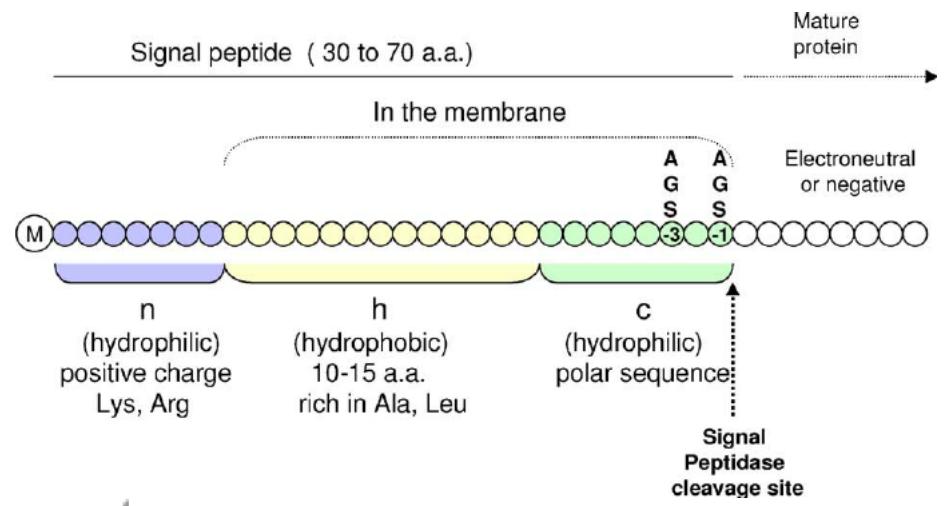
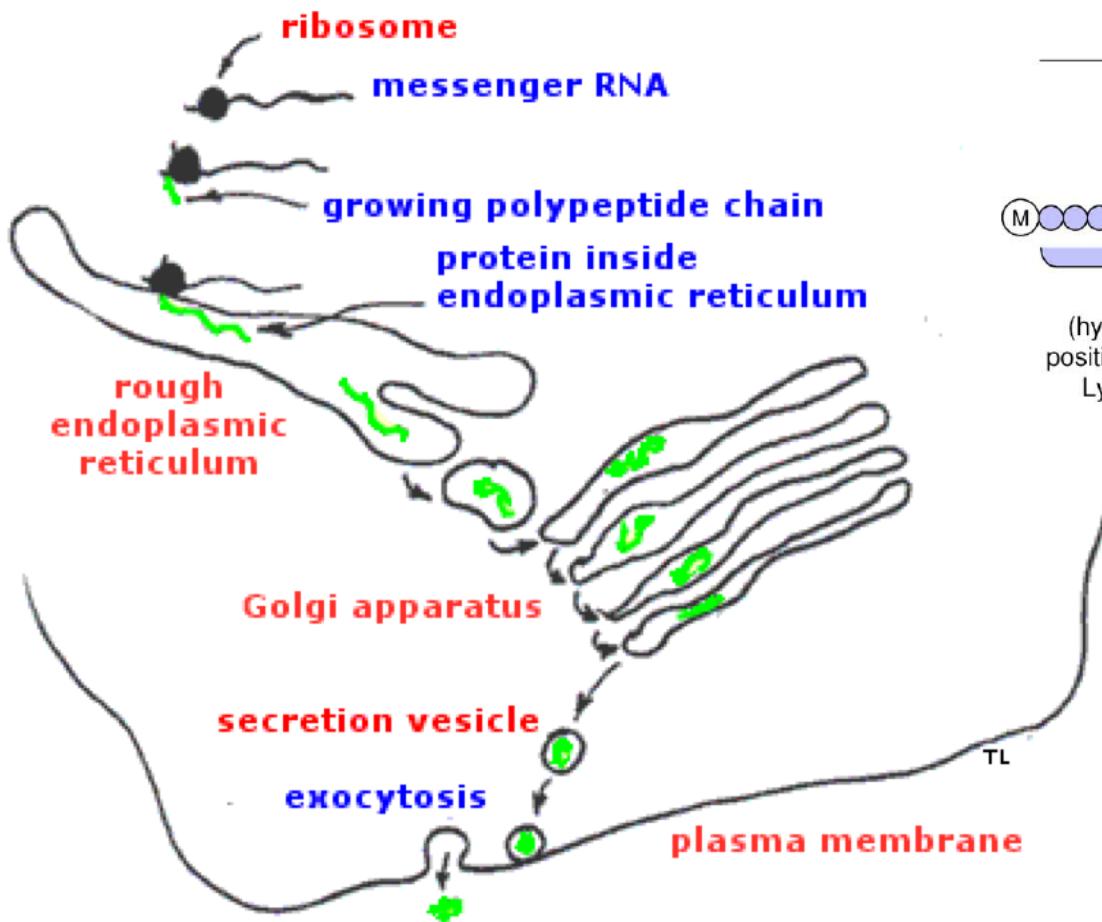
# Trans-membrane Domains via TmHMM



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/

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](#) >> [SignalP](#)

### 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](#).

[FAQ](#) [Article abstracts](#) [Instructions](#) [Output format](#) [Performance](#) [Data](#)

#### SUBMISSION

Paste a single amino acid sequence or several sequences in **FASTA** format into the field below:

```
MHPAVFLSLPDLRCSLLLLTVWFTPVTEITSLDTENIDEILNNADVALVNFYADWCRFSQMLHPIFEASDVIKEFPNENQWFARVDCDQHSDIAQRYRISKYPTLKLFRNGMMMKREYRGQRSVKALADYIRQQKSDPIQEIRDLAETTLDRSKRNIIGYFEQKDSNDYRVFERVANILHDDCAFLSAFGDVSKPERYSQGDNIYKPPGHSAPDMVYLGAMTNFDVTYNWIQDKCVPVLVREITFENGELTEEGLPFLILFHMKEDTESLEIFQNNEVARQLISEKGTTINFLHADCDKFRHPLLHIQKTPADCPVIAIDSFRHMYVFGDFKDVLIPGKLQKFVFDLHSGKLHREFHHGPDPTDTAPGEQAQDVASSPPESSFQKLAPSEYRYTLLRDRDEL
```

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

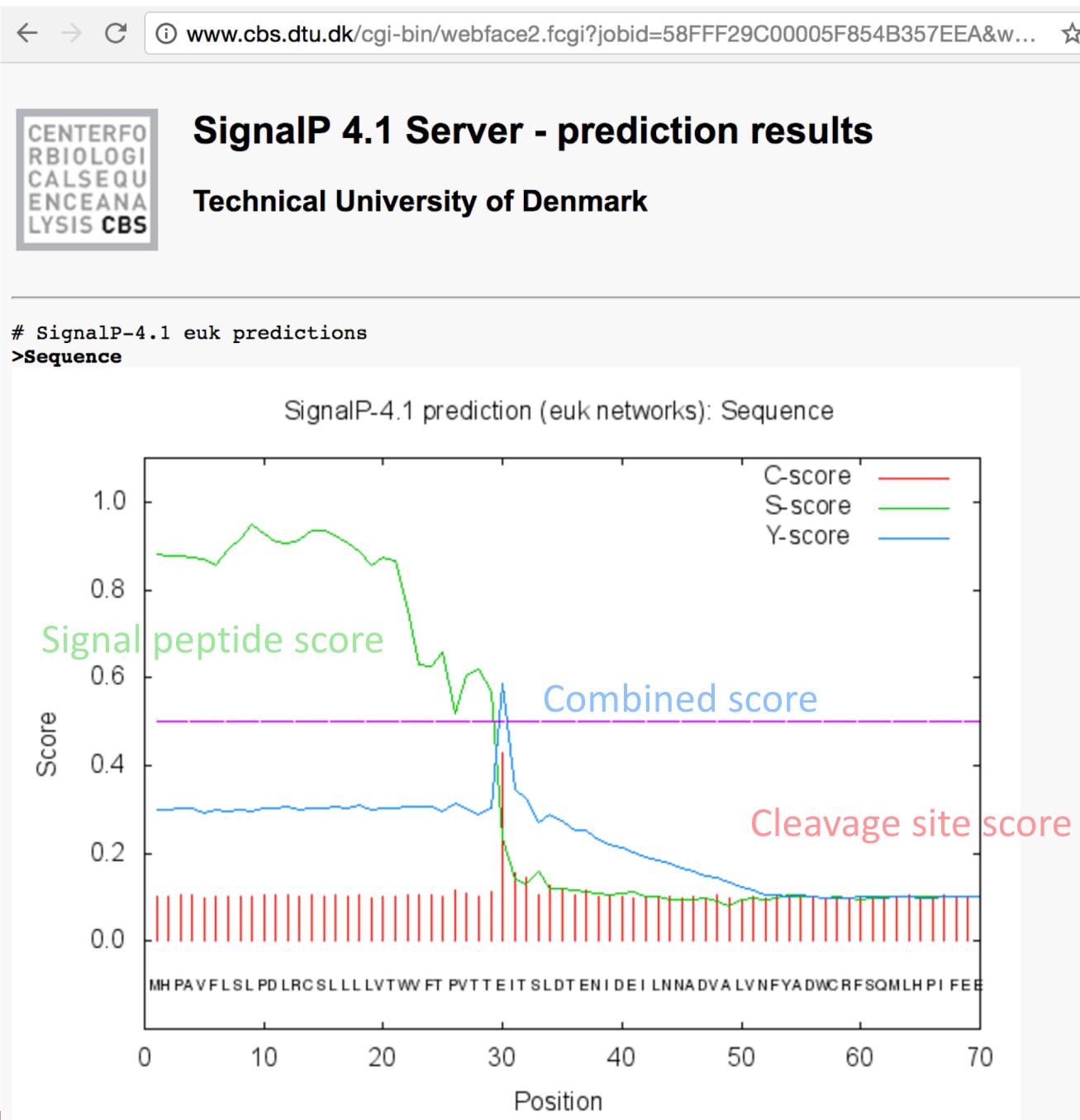
**Graphics output (explain)**  
 No graphics  
 PNG (inline)  
 PNG (inline) and EPS (as links)

**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



# Transcriptome-scale functional annotation using Trinotate



## Trinotate: Transcriptome Functional Annotation and Analysis

# Trinotate



TMHMM

SignalP



TransDecoder



Pfam

eggNOG  
version 3.0



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

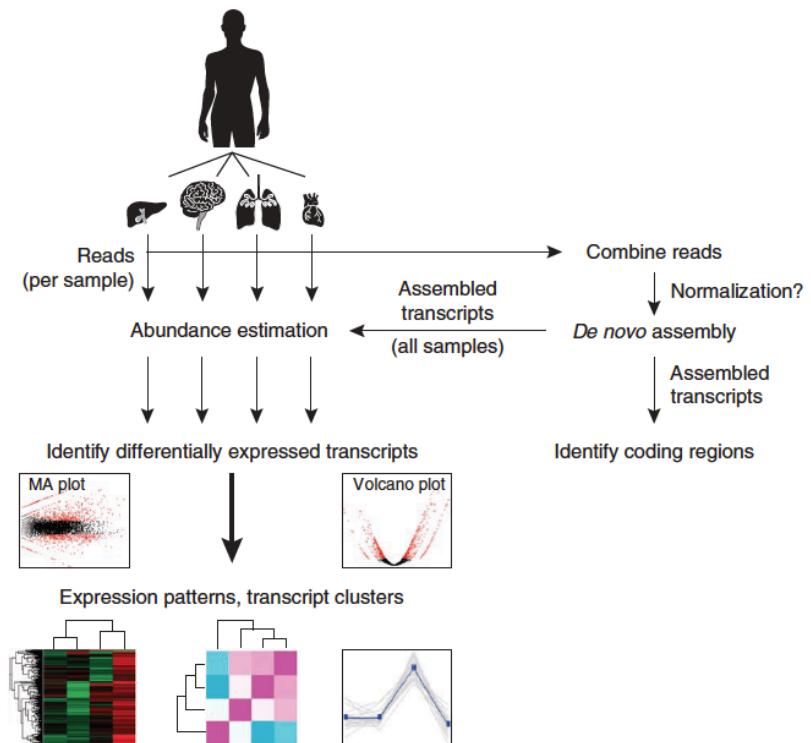
### *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

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

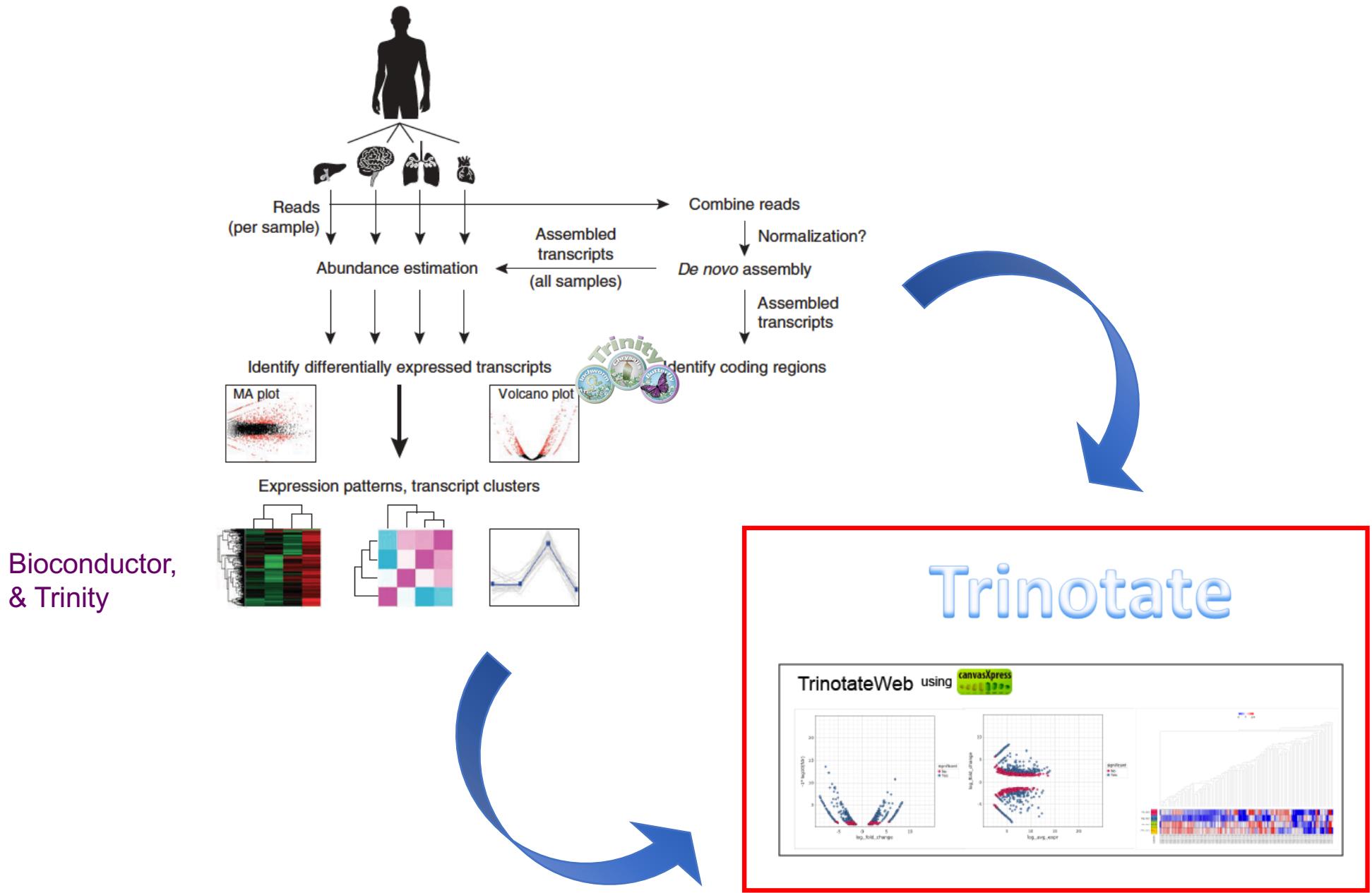
Affiliations | Contributions | Corresponding authors

*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 Functional Annotation Lab



# We are on a Coffee Break & Networking Session

Workshop Sponsors:



Canadian Centre for  
Computational  
Genomics

**MicM** McGill initiative in  
Computational Medicine