21/11/2018

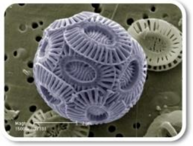
Transcriptome annotation with Tinotate

Journée « Annotation structurale et fonctionnelle des génomes eukaryotes » . PEPI Annot

E. Corre

Plateforme ABiMS

Contexte



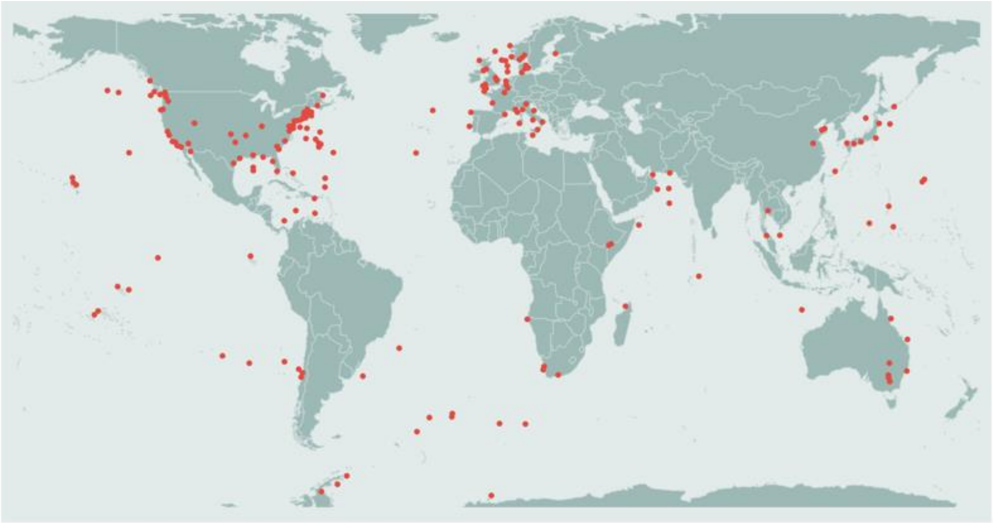
Transcriptome analysis:

- "Krill Arctic" J. Y. Toullec (Roscoff-UMR7144)
- "Chondrus" J. Colleen (Roscoff-UMR8227)
- "Marine Pico Eukaryotes " RmetaT Colomban de Vargas (Roscoff-UMR7144)
- "Radiolaires" F. Not (Roscoff-UMR7144)
- " Artemia" C. Lejeune (Roscoff-UMR7144)
- "JAWS" J.Y. Sire (UPMC Paris)
- "Sepia, Pterois, Heterotis , Saculine" J. Henry C.Gaudin (univ. Caen)
- "Sepia" Laure Bonneau (Paris)
- "Vers tubifères" P.J. Lopez Museum (Paris)
- "Sabellaria alveolata" F. Nunes (Brest)
- "Ormeaux" F. Nunes (Brest)
- "Mucor" A. Le Breton/ JL. Jany / L. Meslet Cladiere (Brest)
- Quercus. J.P. Mevy (Marseille)
- Canard Foie gras. C. Diot (Rennes)

MMETSP

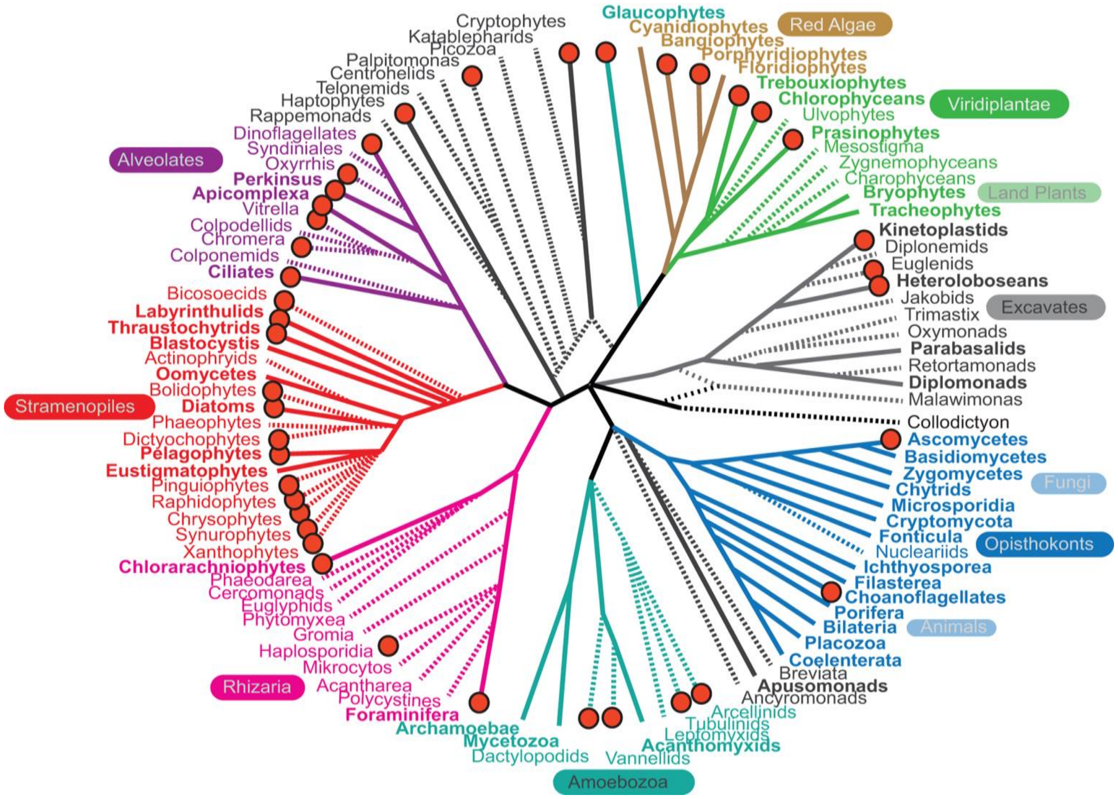
The Moore initiative : Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) was funded to broaden the diversity of sequenced marine protists (study of their evolution and their roles in marine ecosystems)

With data from species spanning more than 40 eukaryotic phyla, the MMETSP provides one of the largest publicly-available collections of RNAseq data from a wide diversity of species.

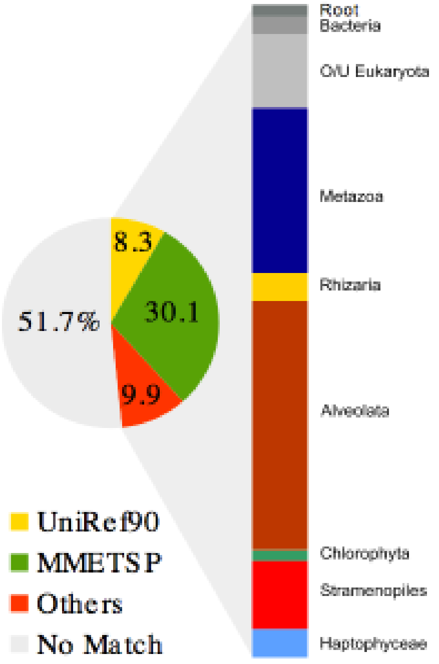


Caron et al. 2016
PMID: 27867198

MMETSP vs MATOU ressource



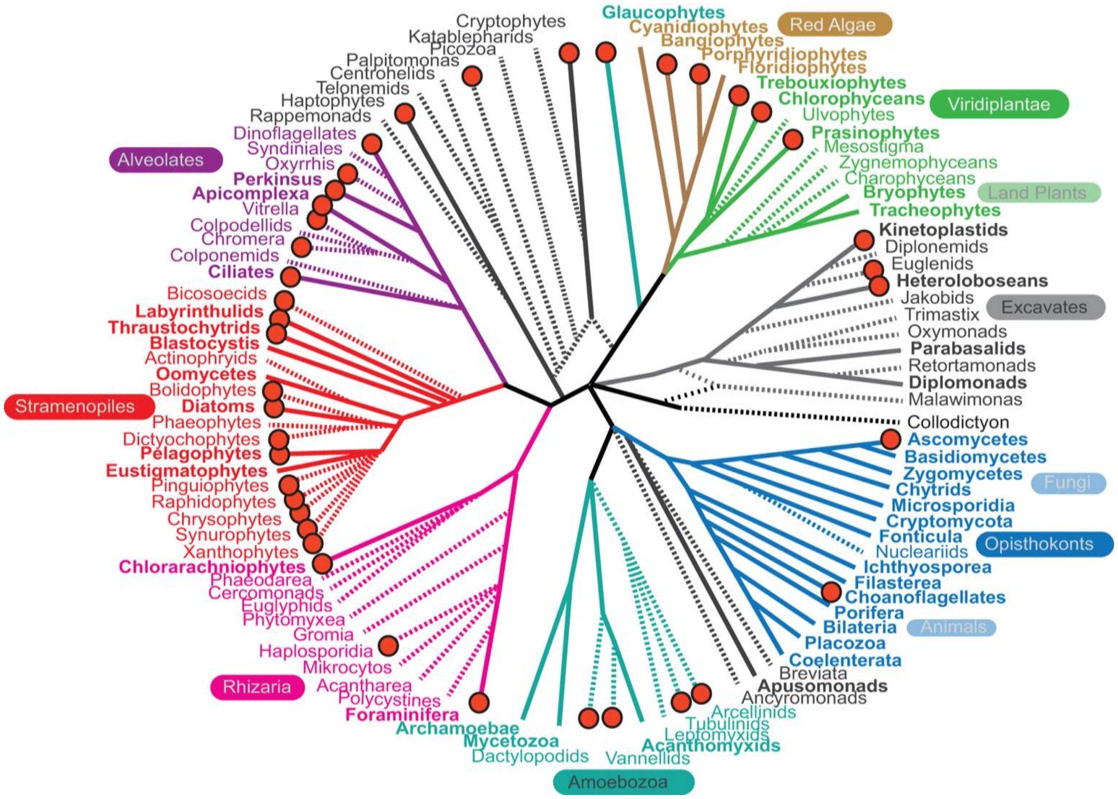
678 transcriptomes from 411 organisms.
First assemblies provided by the NGCR



~ 50% of MATOU genes have no matches in current databases, and 30% have their best similarities with MMETSP database.

Quentin Carradec, Eric Pelletier, et. al - Nature Communications vol.9, 373 (2018). the Marine Atlas of **Tara** Oceans Unigenes (**MATOU**)

MMETSP vs TARA/RCC initiative



TARA OCEANS

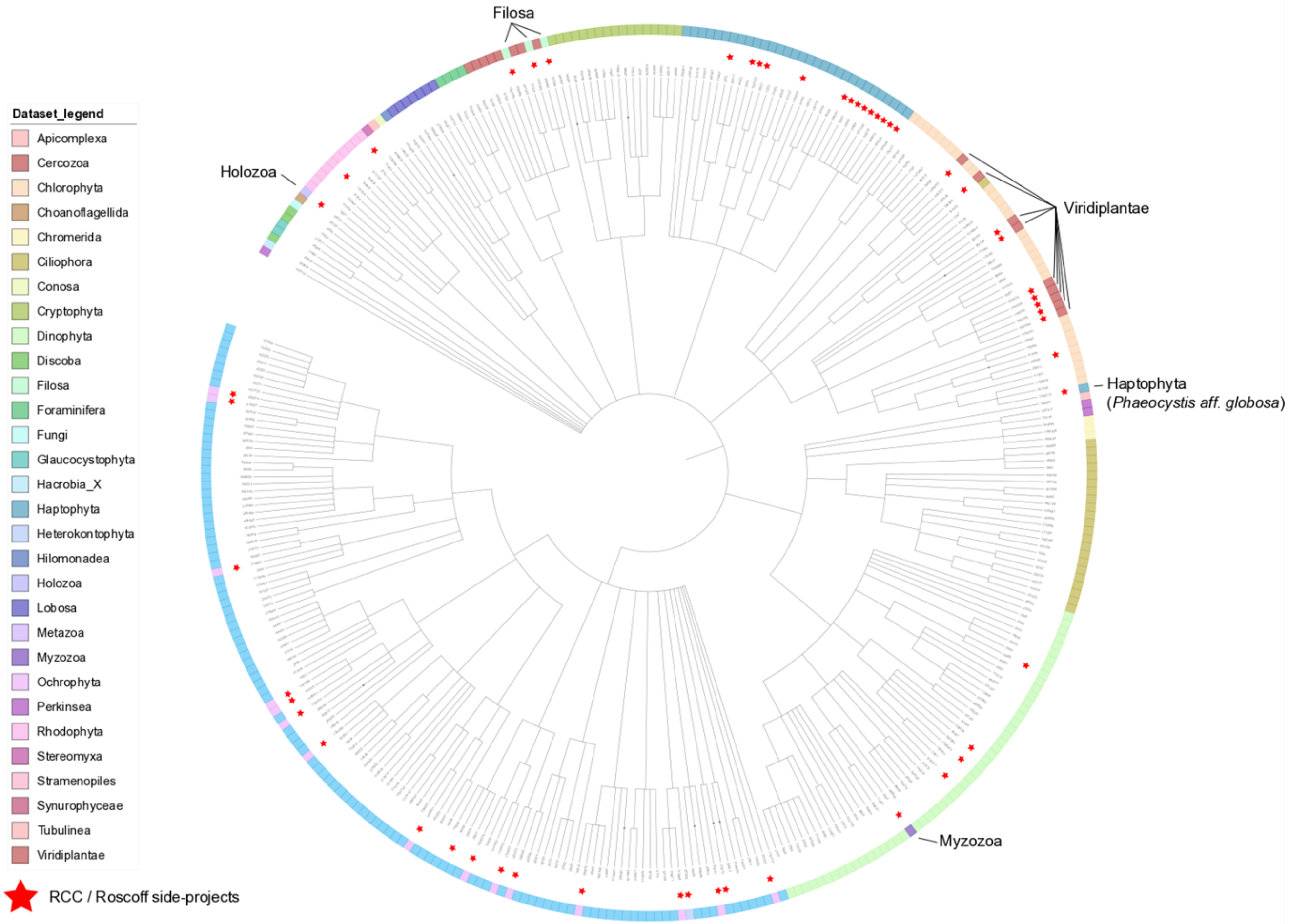
Roscoff RCC Culture Collection

678 transcriptomes from 411 organisms.
First assemblies provided by the NGCR

~ 100 new taxa sequenced by Genoscope and assembled in Roscoff Marine station

elixir FRANCE

WP6 : Marine Micro Eucaryotes Transcriptomes Reference Database



RCC additional species

- Haptophyta
- Dinophyta
- Viridiplantae
- Filosa

-> Develop an homogenous assembly and annotation pipeline

Erwan Corre , Rob Finn, Xi Liu, Arnaud Meng, Guita Niang, Eric Pelletier, Maxim Scheremetjew

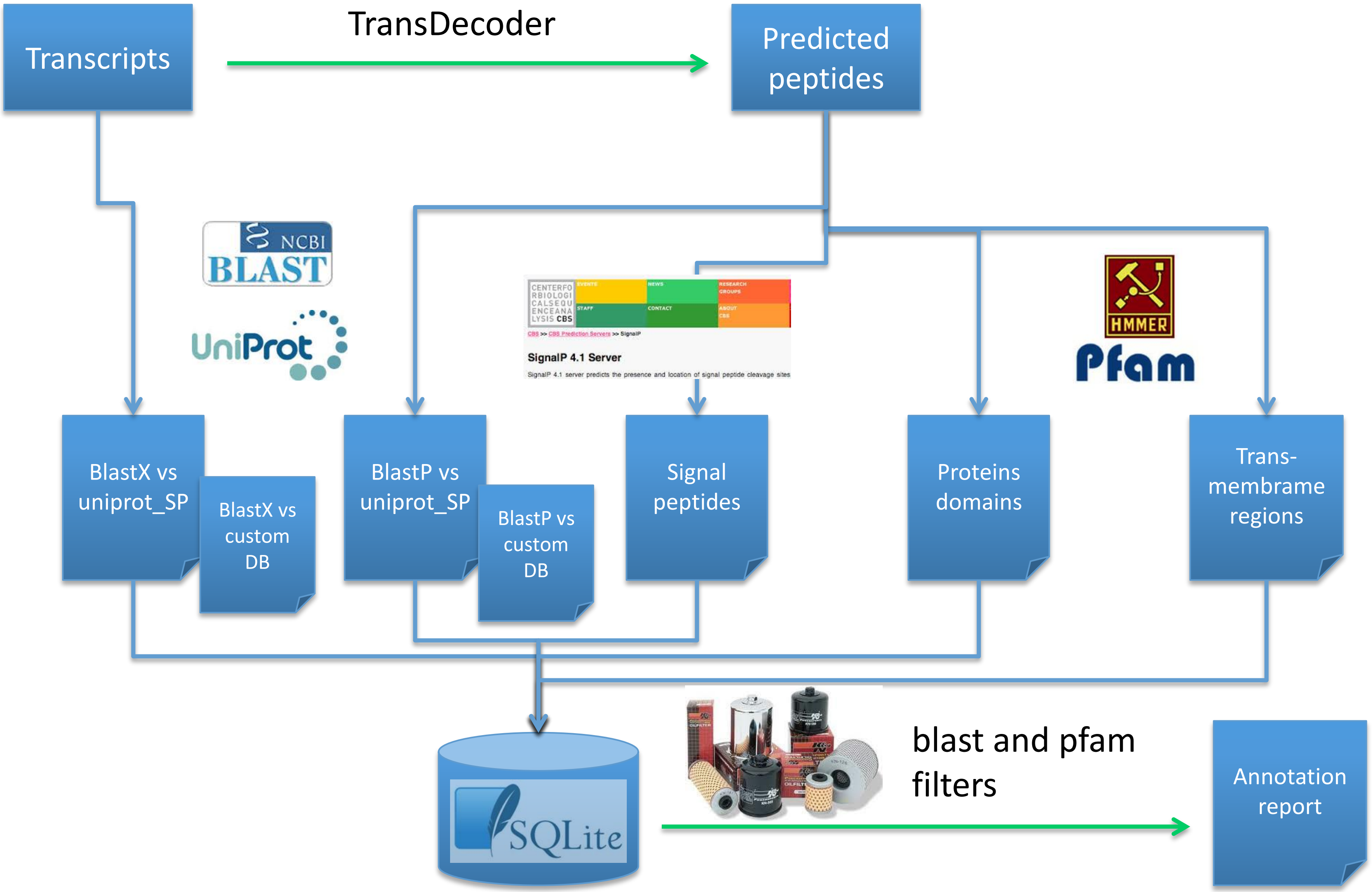
Trinotate pipeline

Trinotate
Trinity
Inchworm
Chrysalis
Butterfly
NCBI
BLAST
HMMER
Pfam
UniProt
eggNOG
version 3.0
SQLite
the Gene Ontology

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

Automated Higher Order Biological Analysis

Trinotate pipeline



Transcripts

TransDecoder

Predicted peptides

BlastX vs uniprot_SP

BlastX vs custom DB

BlastP vs uniprot_SP

BlastP vs custom DB

Signal peptides

Proteins domains

Trans-membrane regions

blast and pfam filters

Annotation report

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

EVENTS NEWS RESEARCH GROUPS STAFF CONTACT ABOUT CBS

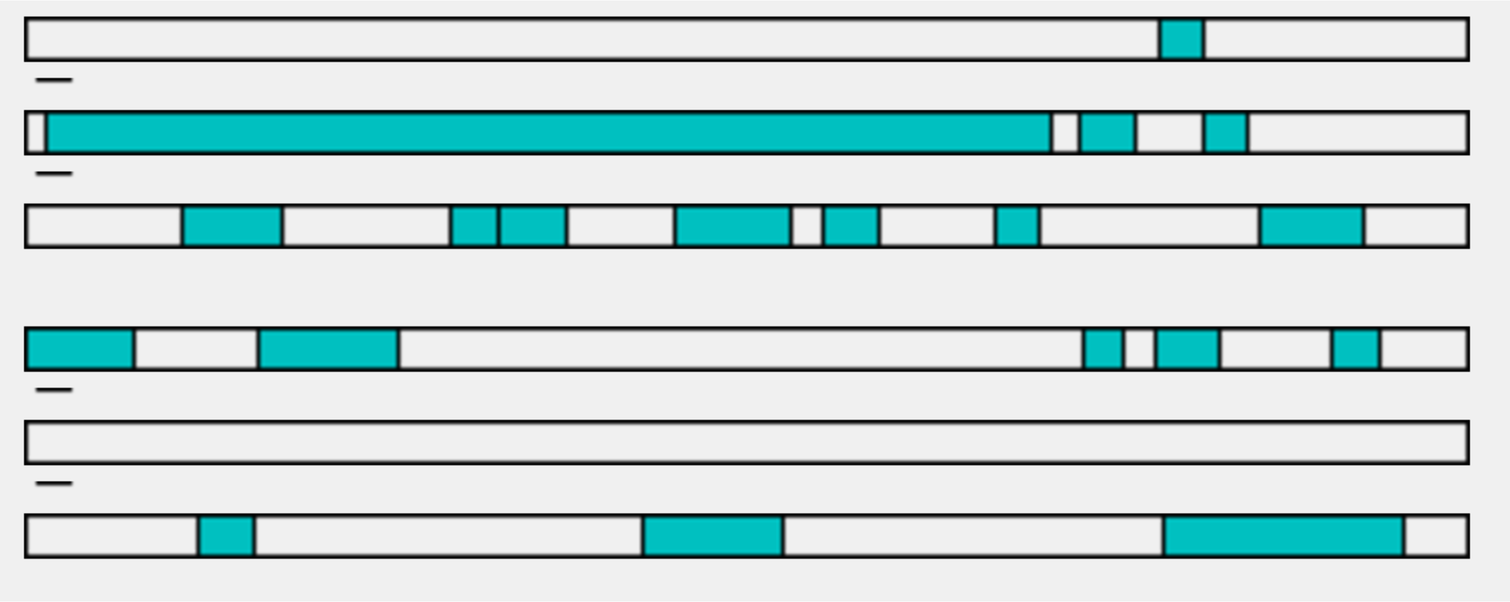
CBS >> CBS Prediction Servers >> SignalP

SignalP 4.1 Server

SignalP 4.1 server predicts the presence and location of signal peptide cleavage sites

Trinotate pipeline

1. Find Likely Coding Regions(using TransDecoder)



The first TransDecoder step identifies all long ORFs.

Score each ORF according to likely coding potential (Markov model)
Report highest scoring ORFs

TransDecoder

TransDecoder identifies likely coding sequences based on the following criteria:

- a minimum length open reading frame (ORF) is found in a transcript sequence
- a log-likelihood score similar to what is computed by the GeneID software is > 0.
- the above coding score is greatest when the ORF is scored in the 1st reading frame as compared to scores in the other 2 forward reading frames.
- if a candidate ORF is found fully encapsulated by the coordinates of another candidate ORF, the longer one is reported. However, a single transcript can report multiple ORFs (allowing for operons, chimeras, etc).
- a PSSM is built/trained/used to refine the start codon prediction.
- **optional** the putative peptide has a match to a Pfam domain above the noise cutoff score. identify ORFs with homology to known proteins via blast or pfam searches

TransDecoder : output

- **transcripts.fasta.transdecoder.pep :** peptide sequences for the final candidate ORFs; all shorter candidates within longer ORFs were removed.
- **transcripts.fasta.transdecoder.cds :** nucleotide sequences for coding regions of the final candidate ORFs
- **transcripts.fasta.transdecoder.gff3 :** positions within the target transcripts of the final selected ORFs
- **transcripts.fasta.transdecoder.bed :** bed-formatted file describing ORF positions, best for viewing using GenomeView or IGV.

Trinotate pipeline

2. Capturing BLASTP and BLASTX Homologies : uniprot-swissprot/uniref 90 and custom DB

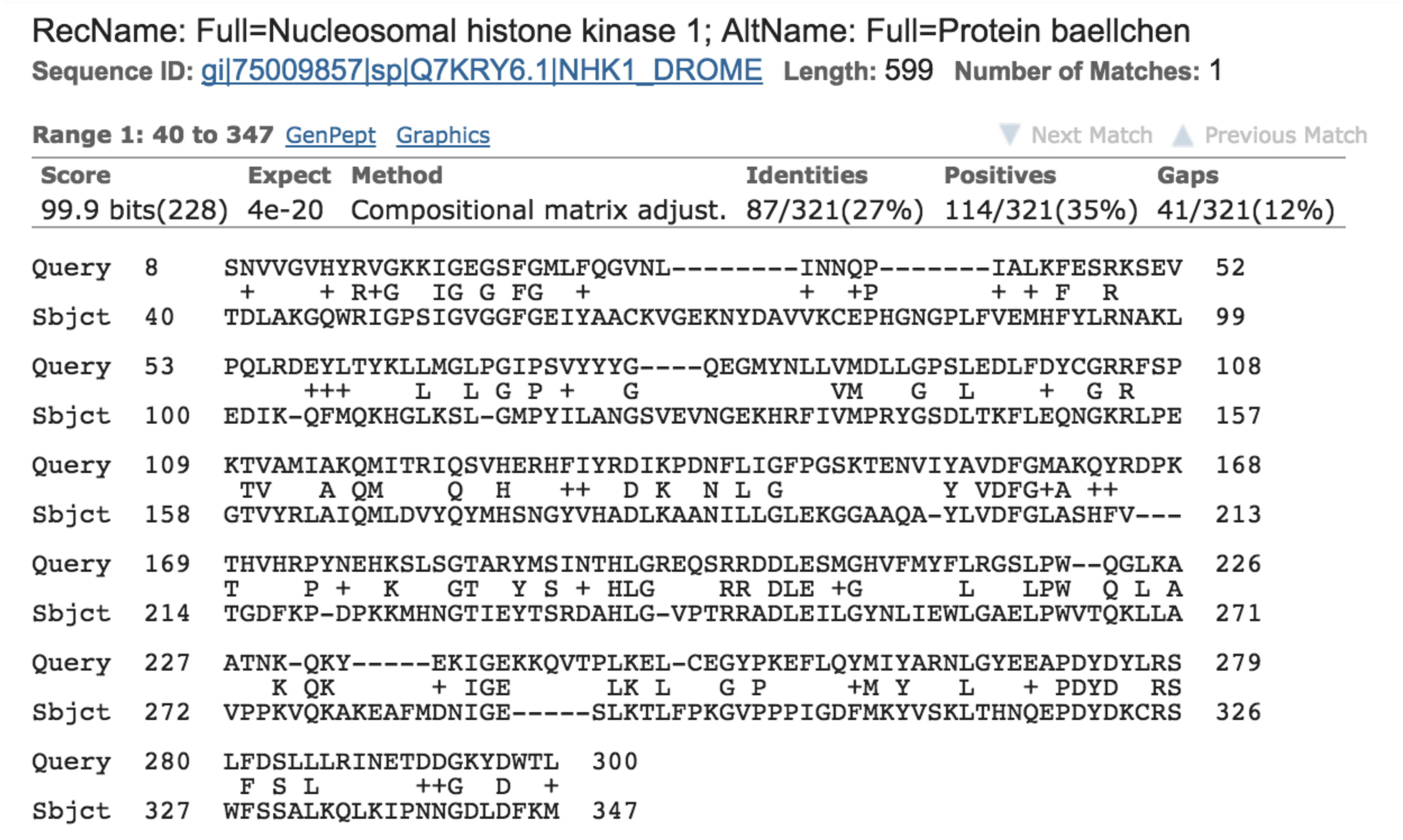
3. Running HMMER to identify protein domains

4. Running signalP to predict signal peptides

5. Running tmHMM to predict transmembrane regions

6. Running Rnammer to detected rRNA

BLAST Uniprot-swissprot



RecName: Full=Nucleosomal histone kinase 1; AltName: Full=Protein baellchen

Sequence ID: gi|75009857|sp|Q7KRY6.1|NHK1_DROME **Length:** 599 **Number of Matches:** 1

Range 1: 40 to 347 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
99.9 bits(228)	4e-20	Compositional matrix adjust.	87/321(27%)	114/321(35%)	41/321(12%)

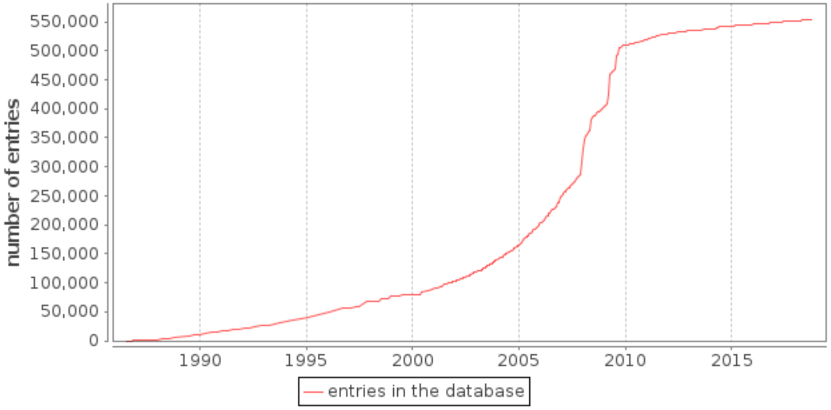
```
Query  8    SNVVGVHYRVGKKIGEGSFGMLFQGVNL--------INNQP-------IALKFESRKSEV  52
             +    + R+G  IG G FG  +             +  +P       + + F  R
Sbjct  40   TDLAKGQWRIGPSIGVGGFGEIYAACKVGEKNYDAVVKCEPHGNGPLFVEMHFYLRNAKL  99

Query  53   PQLRDEYLTYKLLMGLPGIPSVYYYG----QEGMYNLLVMDLLGPSLEDLFDYCGRRFSP  108
                 +++    L  L G P +   G            VM   G  L    +  G R
Sbjct  100  EDIK-QFMQKHGLKSL-GMPYILANGSVEVNGEKHRFIVMPRYGSDLTKFLEQNGKRLPE  157

Query  109  KTVAMIAKQMITRIQSVHERHFIYRDIKPDNFLIGFPGSKTENVIYAVDFGMAKQYRDPK  168
             TV   A QM    Q  H   ++  D K  N L G          Y VDFG+A ++
Sbjct  158  GTVYRLAIQMLDVYQYMHSNGYVHADLKAANILLGLEKGGAAQA-YLVDFGLASHFV---  213

Query  169  THVHRPYNEHKSLSGTARYMSINTHLGREQSRRDDLESMGHVFMYFLRGSLPW--QGLKA  226
            T    P +  K   GT  Y S + HLG    RR DLE +G      L   LPW  Q L A
Sbjct  214  TGDFKP-DPKKMHNGTIEYTSRDAHLG-VPTRRADLEILGYNLIEWLGAELPWVTQKLLA  271

Query  227  ATNK-QKY-----EKIGEKKQVTPLKEL-CEGYPKEFLQYMIYARNLGYEEAPDYDYLRS  279
               K QK      + IGE      LK L   G P     +M Y   L   + PDYD  RS
Sbjct  272  VPPKVQKAKEAFMDNIGE-----SLKTLFPKGVPPPIGDFMKYVSKLTHNQEPDYDKCRS  326

Query  280  LFDSLLLRINETDDGKYDWTL  300
             F S L      ++G  D  +
Sbjct  327  WFSSALKQLKIPNNGDLDFKM  347
```

BLASTX and BLASTP

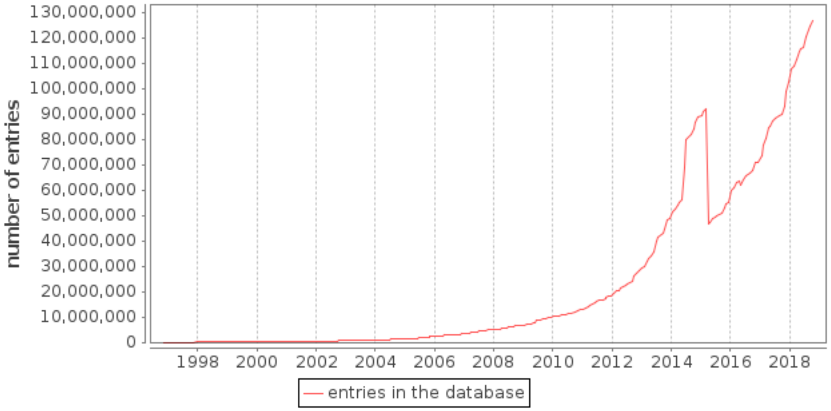
UniProt release 2018_09 consists of two sections:

- **Reviewed (Swiss-Prot) - Manually annotated 558 590 sequences**
  Records with information extracted from literature and curator-evaluated computational analysis.
- **Unreviewed (TrEMBL) - Computationally analyzed 126,780,198 sequences**
  Records that await full manual annotation.

Number of entries in UniProtKB/Swiss-Prot over time



Number of entries in UniProtKB/TrEMBL over time

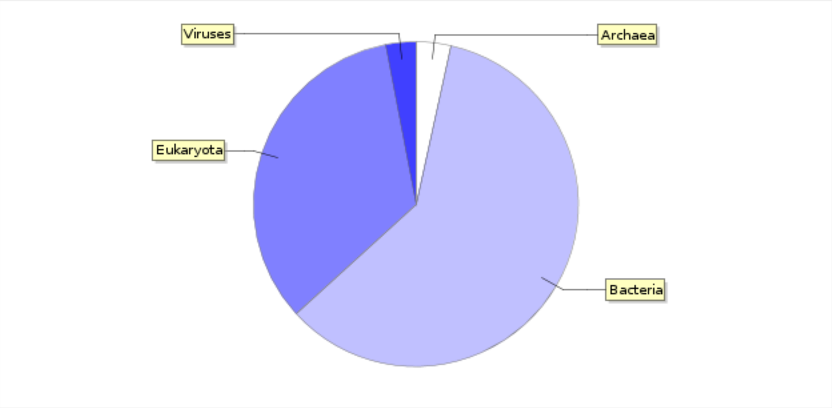


vs

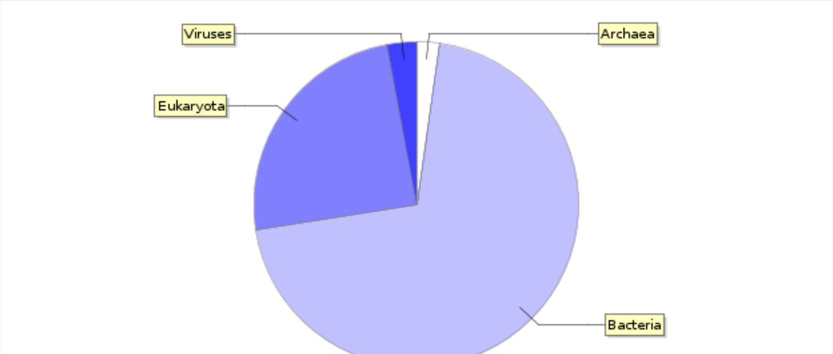
558 590 sequences

TrEMBL : 126 780 198 sequences

Swiss-Prot entries per taxonomic group

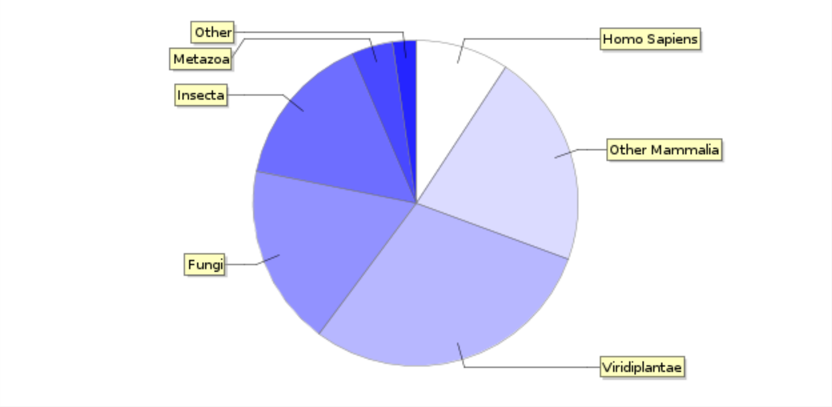


TrEMBL entries per taxonomic group

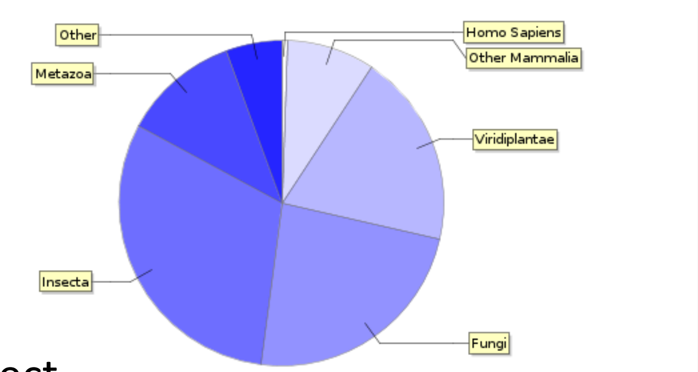


++ bacteria

Swiss-Prot entries in Eukaryota



TrEMBL entries in Eukaryota



++ insect

UniProtKB/TrEMBL: one record for 100% identical full-length sequences in one species;
UniProtKB/Swiss-Prot: one record per gene in one species;

UniParc: one record for **100% identical sequences** over the **entire length**, regardless of the species;
UniRef100: one record for 100% identical sequences, **including fragments**, regardless of the species.

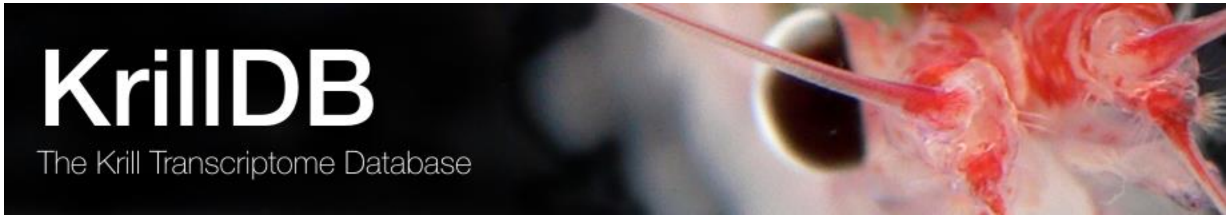
UniRef100 combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry.
UniRef90 is built by clustering UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to, and 80% overlap with, the longest sequence (a.k.a. seed sequence).

UniRef50 (29 636 339)
UniRef90 (80 685 154)
UniRef100 (159 146 034)

http://www.uniprot.org/help/redundancy

Specific databases …

OGA
Ocean Gene Atlas
One Click Marine Gene Biogeography
TARA OCEANS
2009-2012
mio
Mediterranean Institute of Oceanography
OCEANOMICS
ETH zürich
GENOSCOPE

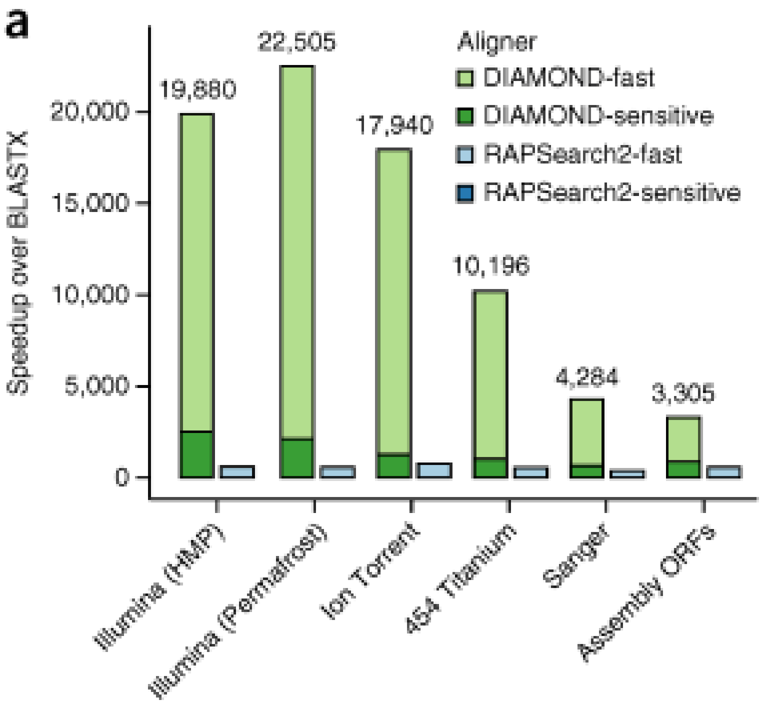


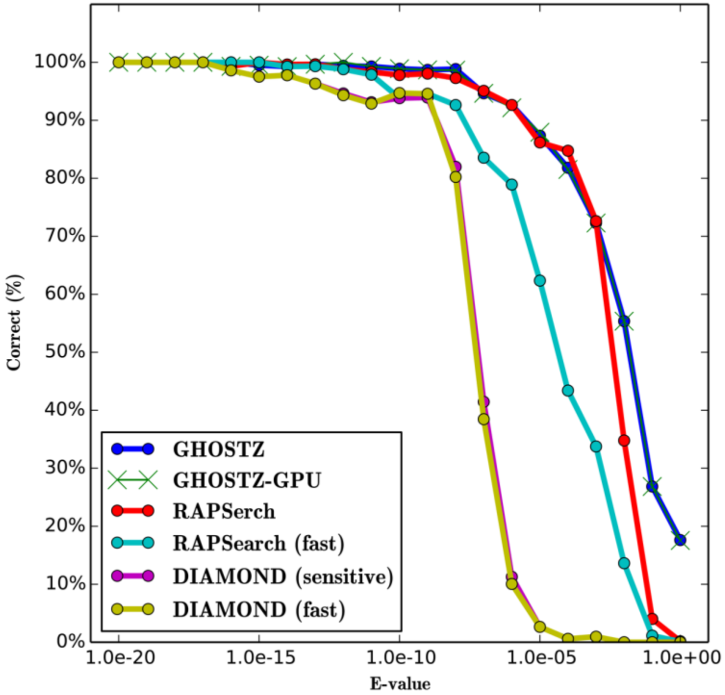
…

BLAST and DIAMOND

DIAMOND : Accelerated BLAST compatible local sequence aligner.

Benjamin Buchfink, Chao Xie & Daniel H. Huson, Fast and Sensitive Protein Alignment using DIAMOND, Nature Methods, 12, 59–60 (2015) doi:10.1038/nmeth.3176.





Diamond X

diamX_uniprot.outfmt6

```
TRINITY_DN97_c0_g1_i1 DNAJ_LACC3 39.7 68 38 1 1102 1296 113 180 1.2e-05 52.8
TRINITY_DN63_c0_g1_i1 PSAC_ACAM1 93.8 81 5 0 62 304 1 81 4.4e-42 171.8
TRINITY_DN67_c0_g1_i1 PUX2_ARATH 28.4 74 51 1 812 1033 176 247 1.1e-04 49.7
TRINITY_DN67_c0_g1_i2 PUX2_ARATH 28.4 74 51 1 678 899 176 247 1.0e-04 49.7
TRINITY_DN85_c0_g2_i1 ANO7_HUMAN 28.2 262 138 6 4 639 320 581 7.2e-22 105.5
TRINITY_DN189_c0_g1_i2 CPSF_ARATH 51.1 92 40 3 121 384 50 140 1.1e-21 104.8
TRINITY_DN118_c0_g1_i1 ARP4_ARATH 37.0 384 218 3 2 1144 77 439 2.9e-64 247.3
TRINITY_DN123_c0_g1_i1 RUBR_SYNY3 48.5 101 48 2 1521 1231 14 114 3.3e-20 101.7
```

diamX_uniref90.outfmt6

```
TRINITY_DN95_c0_g1_i1 UniRef90_W7TYR3 61.4 114 44 0 58 399 9 122 1.3e-34 154.1
TRINITY_DN90_c0_g1_i1 UniRef90_D8LCQ5 44.7 103 55 1 422 114 18 118 2.4e-17 96.3
TRINITY_DN97_c0_g1_i1 UniRef90_D7FKD7 48.6 111 57 0 991 1323 35 145 2.1e-22 114.8
TRINITY_DN15_c0_g1_i1 UniRef90_D7G646 60.0 80 31 1 73 309 243 322 5.2e-18 99.0
TRINITY_DN39_c0_g1_i1 UniRef90_D7FIG4 57.9 392 156 4 218 1393 3 385 8.7e-117 429.5
TRINITY_DN63_c0_g1_i1 UniRef90_A0A088CIH6 91.8 85 7 0 50 304 2 86 1.7e-40 172.9
TRINITY_DN67_c0_g1_i1 UniRef90_D7FV16 65.2 293 102 0 248 1126 32 324 3.6e-95 356.7
TRINITY_DN67_c0_g1_i2 UniRef90_D7FV16 67.6 324 105 0 21 992 1 324 1.6e-110 407.5
TRINITY_DN85_c0_g1_i1 UniRef90_D7FQE2 70.4 125 37 0 376 2 280 404 5.5e-45 188.0
TRINITY_DN85_c0_g2_i1 UniRef90_D7FQE1 75.7 136 31 1 232 639 1 134 1.1e-53 217.6
TRINITY_DN186_c0_g2_i1 UniRef90_D7G5D6 85.8 316 45 0 1 948 125 440 1.3e-147 530.4
TRINITY_DN189_c0_g1_i1 UniRef90_D7FPL2 86.1 36 5 0 58 165 1 36 1.6e-09 70.1
```

Saccharina latissima annotation

DiamondX vs uniprot-swissprot

```
TRINITY_DN10004_c0_g1_i1 ALPL_ARATH 20.9 263 193 8 420 1193 103 355 5.4e-10 67.8
```

DiamondP vs uniprot-swissprot

```
TRINITY_DN10004_c0_g1::TRINITY_DN10004_c0_g1_i1::g.17011::m.17011 ALPL_ARATH 20.7 305 221 10 75 374 67 355 1.1e-11 72.8
```

-> **Protein ALP1-like :** *Arabidopsis thaliana*

DiamondX vs uniprot-uniref90

```
TRINITY_DN10004_c0_g1_i1 UniRef90_D7FSK2 43.8 274 150 3 585 1394 1 274 5.5e-62 246.9
```

-> **Uncharacterized protein Esi_0235_0049** *Ectocarpus siliculosus*

DiamondP vs uniprot-uniref90

```
TRINITY_DN10004_c0_g1::TRINITY_DN10004_c0_g1_i1::g.17011::m.17011 UniRef90_D7FSK2 43.8 274 150 3 172 441 1 274 4.7e-62 246.5
```

-> **Uncharacterized protein Esi_0235_0049** *Ectocarpus siliculosus:* **ALP1-like :** *A. thaliana*

Trinotate pipeline

2. Capturing BLASTP and BLASTX Homologies : uniprot-swissprot/uniref 90

3. Running HMMER to identify protein domains

4. Running signalP to predict signal peptides

5. Running tmHMM to predict transmembrane regions

6. Running Rnammer to detected rRNA

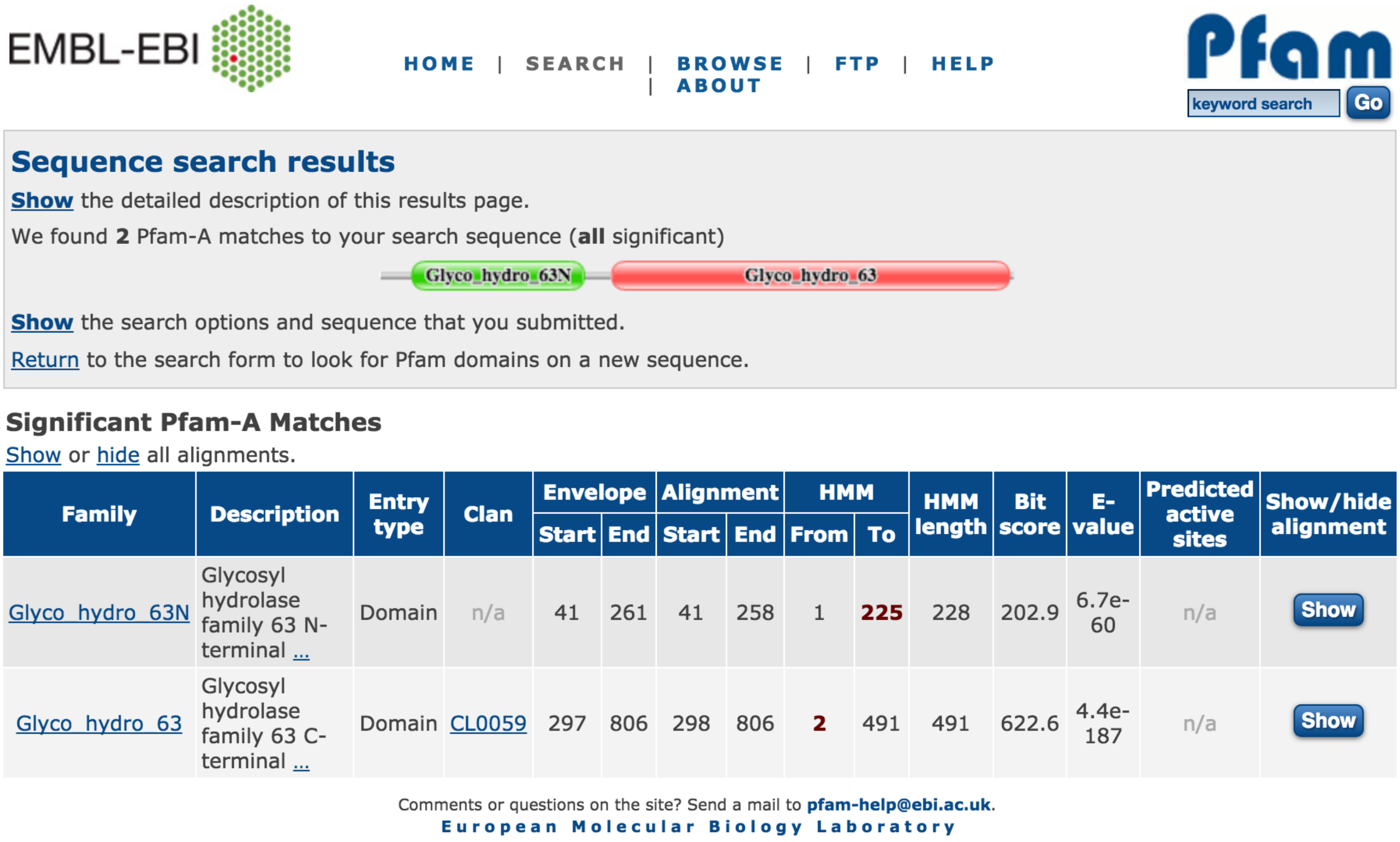
Hmmscan vs Pfam

HMMER is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs).

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs).** The data presented for each entry is based on the UniProt Reference Proteomes

Pfam 32.0 (**Sep 2018**) contains a total of 17929 families and 604 clan

Hmmscan vs Pfam



Sequence search results

Show the detailed description of this results page.

We found **2** 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	Show/hide alignment
				Start	End	Start	End	From	To					
Glyco_hydro_63N	Glycosyl hydrolase family 63 N-terminal ...	Domain	n/a	41	261	41	258	1	**225**	228	202.9	6.7e-60	n/a	Show
Glyco_hydro_63	Glycosyl hydrolase family 63 C-terminal ...	Domain	CL0059	297	806	298	806	**2**	491	491	622.6	4.4e-187	n/a	Show

Comments or questions on the site? Send a mail to **pfam-help@ebi.ac.uk**.
European Molecular Biology Laboratory

Trinity_PFAM.out

```
#                                                                                           --- full sequence --- -------------- this domain ---------
----   hmm coord   ali coord   env coord
# target name        accession   tlen query name                                                 accession   qlen   E-value  score  bias   #  of  c-Evalue  i-Evalue  score
bias  from    to  from    to  from    to  acc description of target
#------------------- ---------- -----                                              -------------------- ---------- ----- --------- ------ ----- --- --- --------- --------- ------ -
---- ----- ----- ----- ----- ----- ----- ---- ---------------------
Plant_tran           PF04827.13   205 TRINITY_DN10004_c0_g1::TRINITY_DN10004_c0_g1_i1::g.17011::m.17011 -            450   5.6e-29  101.1   0.0    1    1    1.4e-32    8.1e-
29  100.6    0.0      3    197    176    374    174    379 0.94 Plant transposon protein

DDE_Tnp_4            PF13359.5    158 TRINITY_DN10004_c0_g1::TRINITY_DN10004_c0_g1_i1::g.17011::m.17011 -            450   4.2e-22    78.4   0.0    1    1    1.2e-25    6.7e-
22    77.7    0.0      2    158    205    372    204    372 0.87 DDE superfamily endonuclease

DDE_Tnp_1            PF01609.20   214 TRINITY_DN10004_c0_g1::TRINITY_DN10004_c0_g1_i1::g.17011::m.17011 -
            450     0.033    13.7    0.7    1    2     0.0036        20     4.6    0.1      9     73    204    270    198    308 0.76 Transposase DDE domain

DDE_Tnp_1            PF01609.20   214 TRINITY_DN10004_c0_g1::TRINITY_DN10004_c0_g1_i1::g.17011::m.17011 -
            450     0.033    13.7    0.7    2    2     0.0007       3.9     7.0    0.1    173    211    330    368    327    373 0.72 Transposase DDE domain

DUF4735              PF15882.4    286 TRINITY_DN10004_c0_g1::TRINITY_DN10004_c0_g1_i1::g.17017::m.17017 -             60     0.055    12.8   0.1    1    1    3.3e-
06     0.055    12.8    0.1    251    285     22     57      3     58 0.77 Domain of unknown function (DUF4735)
```

Trinotate pipeline

2. Capturing BLASTP and BLASTX Homologies : uniprot-swissprot/uniref 90

3. Running HMMER to identify protein domains

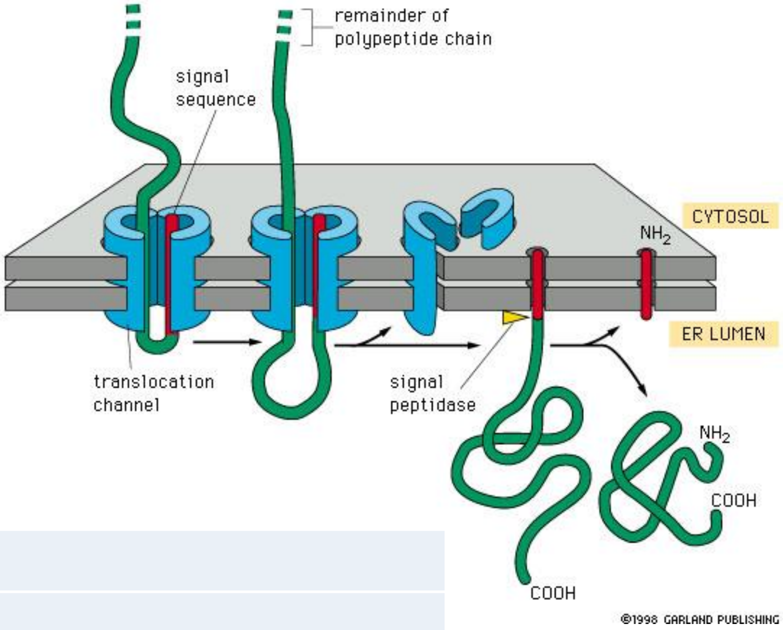
4. Running signalP to predict signal peptides

5. Running tmHMM to predict transmembrane regions

6. Running Rnammer to detected rRNA

SignalP

A signal peptide is a peptide chain of a protein serving to address it to a particular cell (organelle) compartment

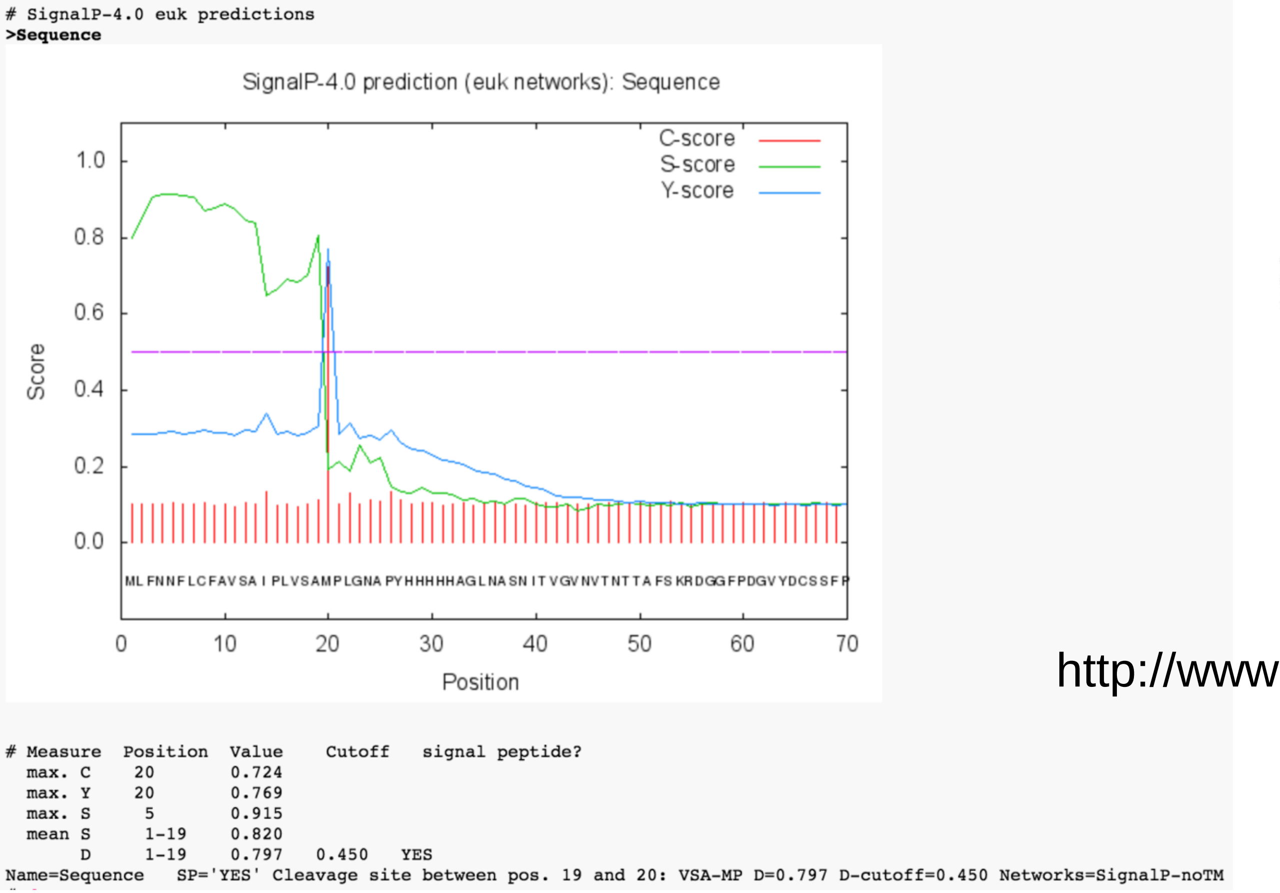


Typical Signal Peptides	
peptide function	Composition
Transport in cellular nucleus (NLS)	-Pro-Pro-Lys-Lys-Lys-Arg-Lys-Val-
Endoplasmic reticulum transport	H_2N-Met-Met-Ser-Phe-Val-Ser-Leu-Leu-Leu-Val-Gly-Ile-Leu-Phe-Trp-Ala-Thr-Glu-Ala-Glu-Gln-Leu-Thr-Lys-Cys-Glu-Val-Phe-Gln-
Endoplasmic reticulum retention	-Lys-Asp-Glu-Leu-COOH
Mitochondrial matrix transport	H_2N-Met-Leu-Ser-Leu-Arg-Gln-Ser-Ile-Arg-Phe-Phe-Lys-Pro-Ala-Thr-Arg-Thr-Leu-Cys-Ser-Ser-Arg-Tyr-Leu-Leu-
Peroxysome (PTS1) transport	-Ser-Lys-Leu-COOH
Peroxysome (PTS2) transport	H_2N-----Arg-Leu-X_5-His-Leu-

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

CNRS • SORBONNE UNIVERSITÉ
Station Biologique de Roscoff
ABiMS

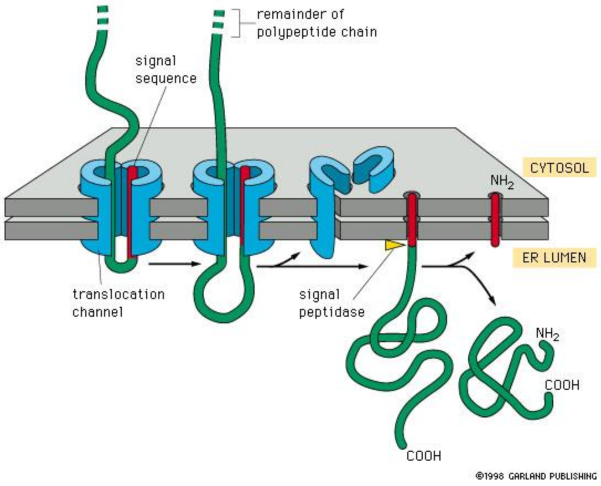
SignalP



```
# SignalP-4.0 euk predictions
>Sequence
```

SignalP-4.0 prediction (euk networks): Sequence
C-score
S-score
Y-score
Score
1.0
0.8
0.6
0.4
0.2
0.0
MLFNNFLCFAVSAIPLVSAMPLGNAPYHHHHHAGLNASNITVGVNVTNTTAFSKRDGGFPDGVYDCSSFP
0
10
20
30
40
50
60
70
Position

```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C     20      0.724
  max. Y     20      0.769
  max. S      5      0.915
  mean S      1-19   0.820
       D      1-19   0.797   0.450    YES
Name=Sequence   SP='YES' Cleavage site between pos. 19 and 20: VSA-MP D=0.797 D-cutoff=0.450 Networks=SignalP-noTM
```

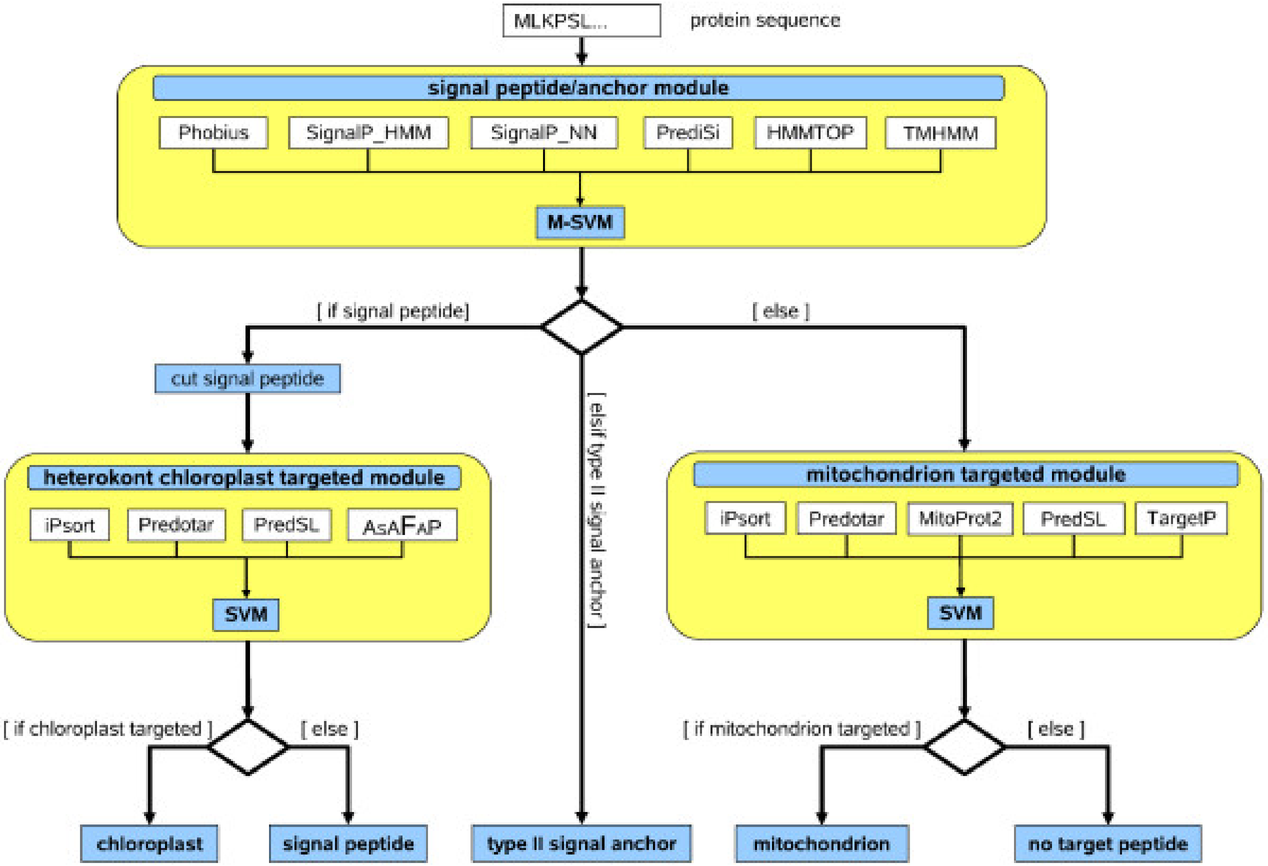


http://www.cbs.dtu.dk/services/SignalP/

```
##gff-version 2
##sequence-name source feature start end score N/A ?
## -----------------------------------------------------------
TRINITY_DN123_c0_g1::TRINITY_DN123_c0_g1_i1::g.213::m.213 SignalP-4.1 SIGNAL 1 20 0.524 . . YES
TRINITY_DN142_c0_g1::TRINITY_DN142_c0_g1_i1::g.238::m.238 SignalP-4.1 SIGNAL 1 18 0.459 . . YES
TRINITY_DN166_c0_g1::TRINITY_DN166_c0_g1_i1::g.284::m.284 SignalP-4.1 SIGNAL 1 28 0.777 . . YES
TRINITY_DN166_c0_g1::TRINITY_DN166_c0_g1_i2::g.290::m.290 SignalP-4.1 SIGNAL 1 28 0.777 . . YES
```

HECTAR

HECTAR (HEterokont subCellular TARgeting) is a statistical prediction method designed to assign proteins to five different categories of subcellular targeting: Signal peptides, type II signal anchors, chloroplast transit peptides, mitochondrion transit peptides and proteins which do not possess any N-terminal target peptide.



Trinotate pipeline

2. Capturing BLASTP and BLASTX Homologies : uniprot-swissprot/uniref 90

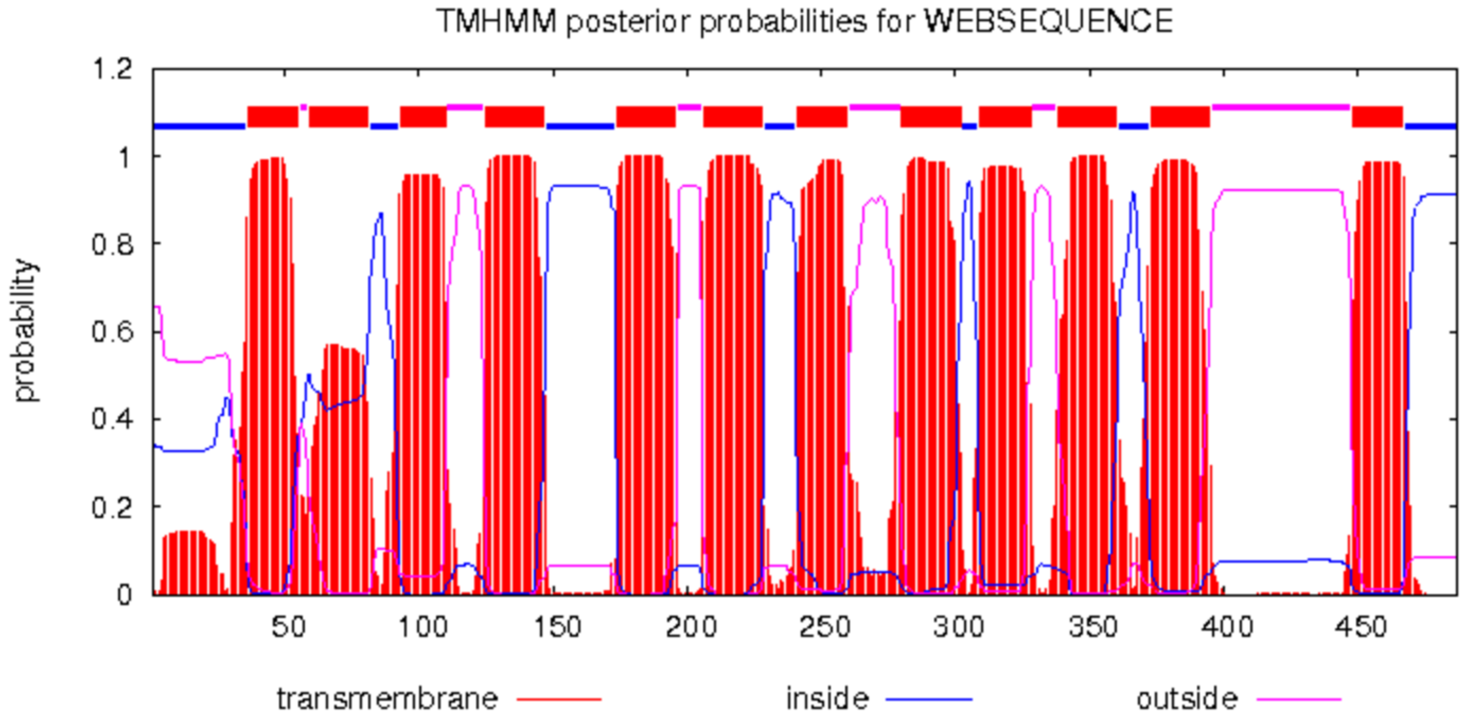
3. Running HMMER to identify protein domains

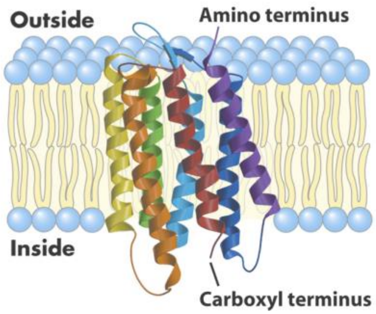
4. Running signalP to predict signal peptides

5. Running tmHMM to predict transmembrane regions

6. Running Rnammer to detected rRNA

TMHMM : Prediction of transmembrane helices in proteins





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

```
TRINITY_DN10013_c0_g2::TRINITY_DN10013_c0_g2_i1::g.17046::m.17046    len=55    ExpAA=0.01     First60=0.01     PredHel=0  Topology=i
TRINITY_DN10016_c0_g1::TRINITY_DN10016_c0_g1_i1::g.17052::m.17052    len=244   ExpAA=12.78    First60=12.76    PredHel=1  Topology=i13-32o
TRINITY_DN10018_c0_g1::TRINITY_DN10018_c0_g1_i1::g.17057::m.17057    len=61    ExpAA=25.61    First60=25.61    PredHel=1  Topology=o4-35i
TRINITY_DN10023_c0_g1::TRINITY_DN10023_c0_g1_i1::g.17077::m.17077    len=84    ExpAA=17.86    First60=17.46    PredHel=0  Topology=o
TRINITY_DN1002_c0_g1::TRINITY_DN1002_c0_g1_i1::g.1928::m.1928        len=106   ExpAA=0.34     First60=0.14     PredHel=0  Topology=o
```

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

Trinotate pipeline

2. Capturing BLASTP and BLASTX Homologies : uniprot-swissprot/uniref 90

3. Running HMMER to identify protein domains

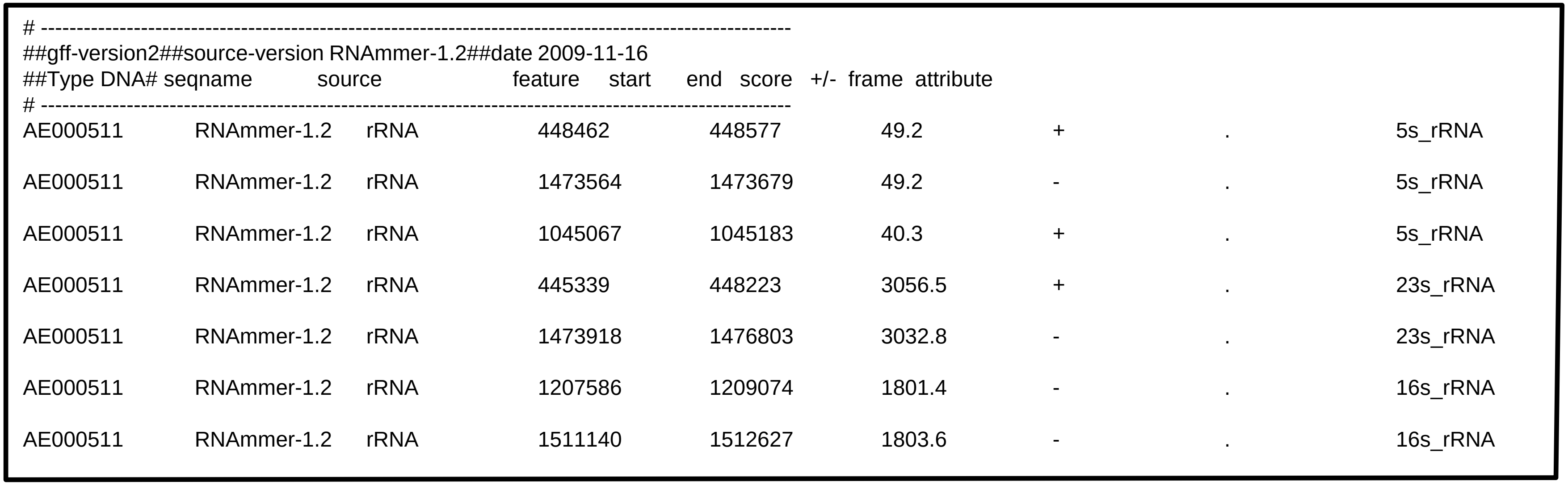
4. Running signalP to predict signal peptides

5. Running tmHMM to predict transmembrane regions

6. Running Rnammer to detected rRNA

RNAMMER

The program uses hidden Markov models trained on data from the 5S ribosomal RNA database and the European ribosomal RNA database project



```
# ---------------------------------------------------------------------------------------------------------
##gff-version2##source-version RNAmmer-1.2##date 2009-11-16
##Type DNA# seqname          source                   feature     start      end   score   +/-  frame  attribute
# ---------------------------------------------------------------------------------------------------------
AE000511        RNAmmer-1.2     rRNA            448462          448577          49.2            +               .               5s_rRNA

AE000511        RNAmmer-1.2     rRNA            1473564         1473679         49.2            -               .               5s_rRNA

AE000511        RNAmmer-1.2     rRNA            1045067         1045183         40.3            +               .               5s_rRNA

AE000511        RNAmmer-1.2     rRNA            445339          448223          3056.5          +               .               23s_rRNA

AE000511        RNAmmer-1.2     rRNA            1473918         1476803         3032.8          -               .               23s_rRNA

AE000511        RNAmmer-1.2     rRNA            1207586         1209074         1801.4          -               .               16s_rRNA

AE000511        RNAmmer-1.2     rRNA            1511140         1512627         1803.6          -               .               16s_rRNA
```

Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T Ussery DW RNammer: consistent annotation of rRNA genes in genomic sequences . Nucleic Acids Res. 2007 Apr 22.

Alternative Barnap :
https://github.com/tseemann/barrnap

CNRS • SORBONNE UNIVERSITÉ
Station Biologique de Roscoff
ABiMS

Trinotate pipeline: annotation report

7. Loading Results into a Trinotate SQLite Database

(perl scripts)

- a boilerplate SQLite database called 'Trinotate.sqlite' that comes pre-populated with a lot of generic data about SWISSPROT records and Pfam domains.
- Need to upload PFAM swissprot database versions specific and synchronized with 'Trinotate.sqlite’ database

Trinotate pipeline: annotation report

7. Loading Results into a Trinotate SQLite Database (perl scripts)

- Trinotate Trinotate.sqlite init --gene_trans_map Trinity.fasta.gene_trans_map --transcript_fasta Trinity.fasta --transdecoder_pep Trinity.fasta.transdecoder.pep

- Trinotate Trinotate.sqlite LOAD_swissprot_blastp blastp.outfmt6 (ou resultats de diamond)
- Trinotate Trinotate.sqlite LOAD_swissprot_blastx blastx.outfmt6 (ou resultats de diamond)
- Trinotate Trinotate.sqlite LOAD_custom_blast --outfmt6 blastx_vs_uniref90.tab --prog blastx --dbtype uniref90
- Trinotate Trinotate.sqlite LOAD_custom_blast --outfmt6 blastp_vs_uniref90.tab --prog blastp --dbtype uniref90
- Trinotate Trinotate.sqlite LOAD_pfam Trinity_PFAM.out
- Trinotate Trinotate.sqlite LOAD_tmhmm Trinity.tmhmm.out
- Trinotate Trinotate.sqlite LOAD_signalp Trinity_signalp.out
- Trinotate Trinotate.sqlite LOAD_rnammer Trinity.fasta.rnammer.gff

Trinotate pipeline: annotation report

8. Threshold the blast and pfam results to be reported

- E-value : maximum blast E-value cutoff
- 'DNC' : domain noise cutoff (default)
- 'DGC' : domain gathering cutoff
- 'DTC' : domain trusted cutoff
- 'SNC' : sequence noise cutoff
- 'SGC' : sequence gathering cutoff
- 'STC' : sequence trusted cutoff

Trinotate pipeline : annotation report

0 #gene_id
1 transcript_id
2 sprot_Top_BLASTX_hit
3 RNAMMER
4 prot_id
5 prot_coords
6 sprot_Top_BLASTP_hit
7 custom_db_nuc_BLASTX
8 custom_db_pep_BLASTP
9 Pfam
10 SignalP
11 TmHMM
12 eggnog
13 Kegg
14 gene_ontology_blast
15 gene_ontology_pfam

16 transcript
17 peptide

Trinotate pipeline : annotation report

0 #gene_id
TRINITY_DN179_c0_g1

1 transcript_id
TRINITY_DN179_c0_g1_i1

2 sprot_Top_BLASTX_hit GCS1_SCHPO^GCS1_SCHPO^Q:53-2476,H:1-808^100%ID^E:0^RecName: Full=Probable mannosyl-oligosaccharide glucosidase;^Eukaryota; Fungi; Dikarya; Ascomycota; Taphrinomycotina; Schizosaccharomycetes; Schizosaccharomycetales; Schizosaccharomycetaceae; Schizosaccharomyces

3 RNAMMER
.

4 prot_id
TRINITY_DN179_c0_g1_i1|m.1

5 prot_coords
2-2479[+]

6 sprot_Top_BLASTP_hit
GCS1_SCHPO^GCS1_SCHPO^Q:18-825,H:1-808^100%ID^E:0^RecName: Full=Probable mannosyl-oligosaccharide glucosidase;^Eukaryota; Fungi; Dikarya; Ascomycota; Taphrinomycotina; Schizosaccharomycetes; Schizosaccharomycetales; Schizosaccharomycetaceae; Schizosaccharomyces

7 custom_db_nuc_BLASTX
SPAC6G10_09_SPAC6G10_09_I_alpha_glucosidase_I_Gls1_predicte^SPAC6G10_09_SPAC6G10_09_I_alpha_glucosidase_I_Gls1_predicte^Q:53-2476,H:1-808^100%ID^E:0^.^.

8 custom_db_pep_BLASTP
SPAC6G10_09_SPAC6G10_09_I_alpha_glucosidase_I_Gls1_predicte^SPAC6G10_09_SPAC6G10_09_I_alpha_glucosidase_I_Gls1_predicte^Q:18-825,H:1-808^100%ID^E:0^.^.

9 Pfam
PF16923.2^Glyco_hydro_63N^Glycosyl hydrolase family 63 N-terminal domain^58-275^E:6.9e-60`PF03200.13^Glyco_hydro_63^Glycosyl hydrolase family 63 C-terminal domain^315-823^E:5.1e-187

10 SignalP
.

11 TmHMM
.

12 eggnog
.

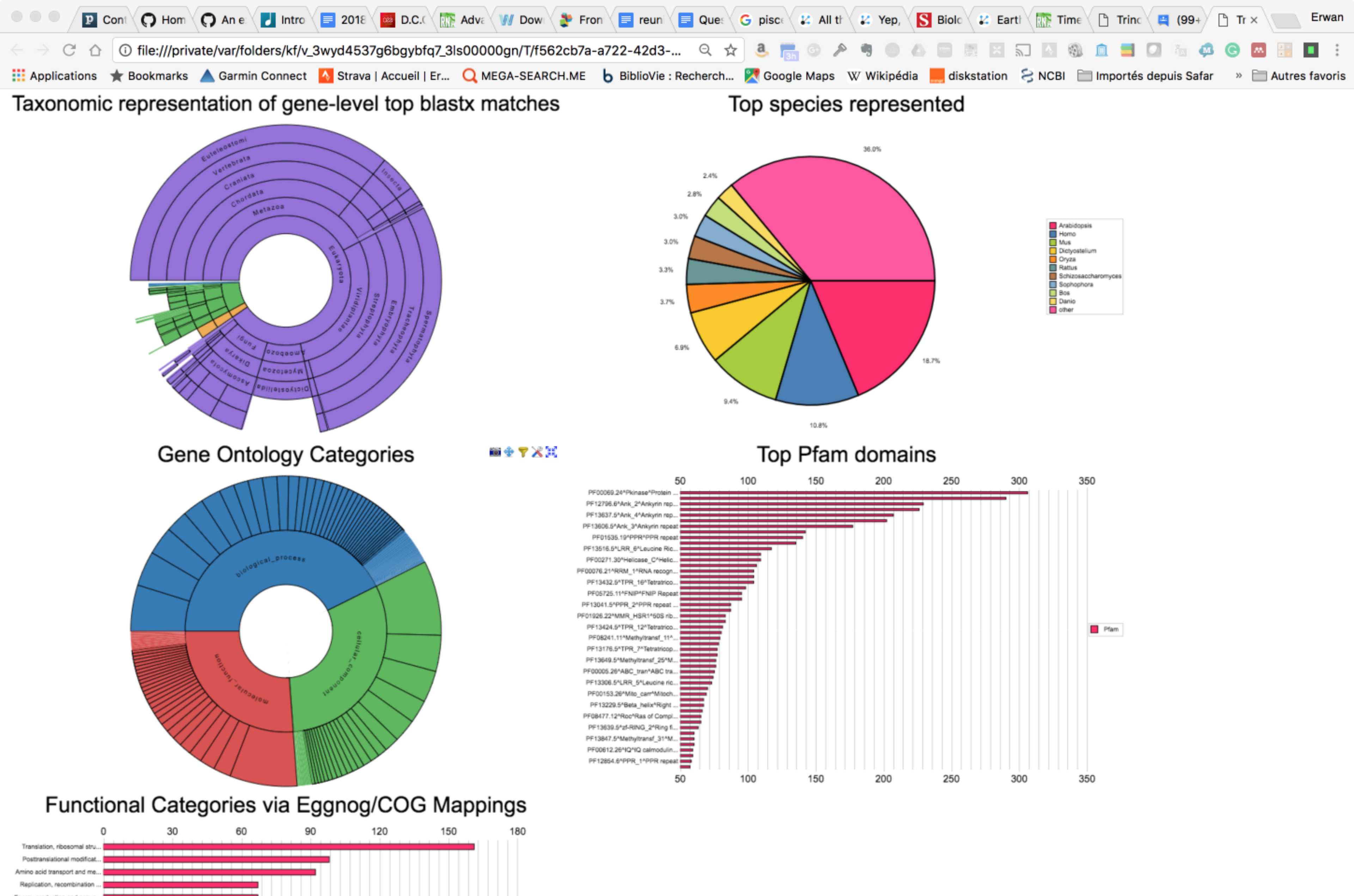
13 Kegg
KEGG:spo:SPAC6G10.09`KO:K01228

14 gene_ontology_blast
GO:0005783^cellular_component^endoplasmic reticulum`GO:0005789^cellular_component^endoplasmic reticulum membrane`GO:0016021^cellular_component^integral component of membrane`GO:0004573^molecular_function^mannosyl-oligosaccharide glucosidase activity`GO:0009272^biological_process^fungal-type cell wall biogenesis`GO:0009311^biological_process^oligosaccharide metabolic process`GO:0006487^biological_process^protein N-linked glycosylation

15 gene_ontology_pfam

16 transcript
17 peptide

New : trinotate_report_summary.pl



file:///private/var/folders/kf/v_3wyd4537g6bybfq7_3ls00000gn/T/f562cb7a-a722-42d3-...

Taxonomic representation of gene-level top blastx matches

Top species represented

Gene Ontology Categories

Top Pfam domains

Functional Categories via Eggnog/COG Mappings

Applications
Bookmarks
Garmin Connect
Strava | Accueil | Er...
MEGA-SEARCH.ME
BiblioVie : Recherch...
Google Maps
Wikipédia
diskstation
NCBI
Importés depuis Safar
Autres favoris
Erwan
Euteleostomi
Vertebrata
Craniata
Chordata
Metazoa
Insecta
Eukaryota
Viridiplantae
Streptophyta
Embryophyta
Tracheophyta
Spermatophyta
Fungi
Dikarya
Ascomycota
Amoebozoa
Mycetozoa
Dictyosteliida
36.0%
2.4%
2.8%
3.0%
3.0%
3.3%
3.7%
6.9%
9.4%
10.8%
18.7%
Arabidopsis
Homo
Mus
Dictyostelium
Oryza
Rattus
Schizosaccharomyces
Sophophora
Bos
Danio
other
biological_process
cellular_component
molecular_function
50
100
150
200
250
300
350
PF00069.24^Pkinase^Protein ...
PF12796.6^Ank_2^Ankyrin rep...
PF13637.5^Ank_4^Ankyrin rep...
PF13606.5^Ank_3^Ankyrin repeat
PF01535.19^PPR^PPR repeat
PF13516.5^LRR_6^Leucine Ric...
PF00271.30^Helicase_C^Helic...
PF00076.21^RRM_1^RNA recogn...
PF13432.5^TPR_16^Tetratrico...
PF05725.11^FNIP^FNIP Repeat
PF13041.5^PPR_2^PPR repeat ...
PF01926.22^MMR_HSR1^50S rib...
PF13424.5^TPR_12^Tetratrico...
PF08241.11^Methyltransf_11^...
PF13176.5^TPR_7^Tetratricop...
PF13649.5^Methyltransf_25^M...
PF00005.26^ABC_tran^ABC tra...
PF13306.5^LRR_5^Leucine ric...
PF00153.26^Mito_carr^Mitoch...
PF13229.5^Beta_helix^Right ...
PF08477.12^Roc^Ras of Compl...
PF13639.5^zf-RING_2^Ring f...
PF13847.5^Methyltransf_31^M...
PF00612.26^IQ^IQ calmodulin...
PF12854.6^PPR_1^PPR repeat
Pfam
0
30
60
90
120
150
180
Translation, ribosomal stru...
Posttranslational modificat...
Amino acid transport and me...
Replication, recombination ...

Automated Execution of Trinotate

TRINOTATE_HOME/auto/autoTrinotate.pl

```
################################################################################
# Required:
#
#--Trinotate_sqlite <string> Trinotate.sqlite boilerplate database
#
#--transcripts <string> transcripts.fasta
#
#--gene_to_trans_map <string> gene-to-transcript mapping file
#
#--conf <string> config file
#
#--CPU <int> number of threads to use.
################################################################################
```

Trinotate web

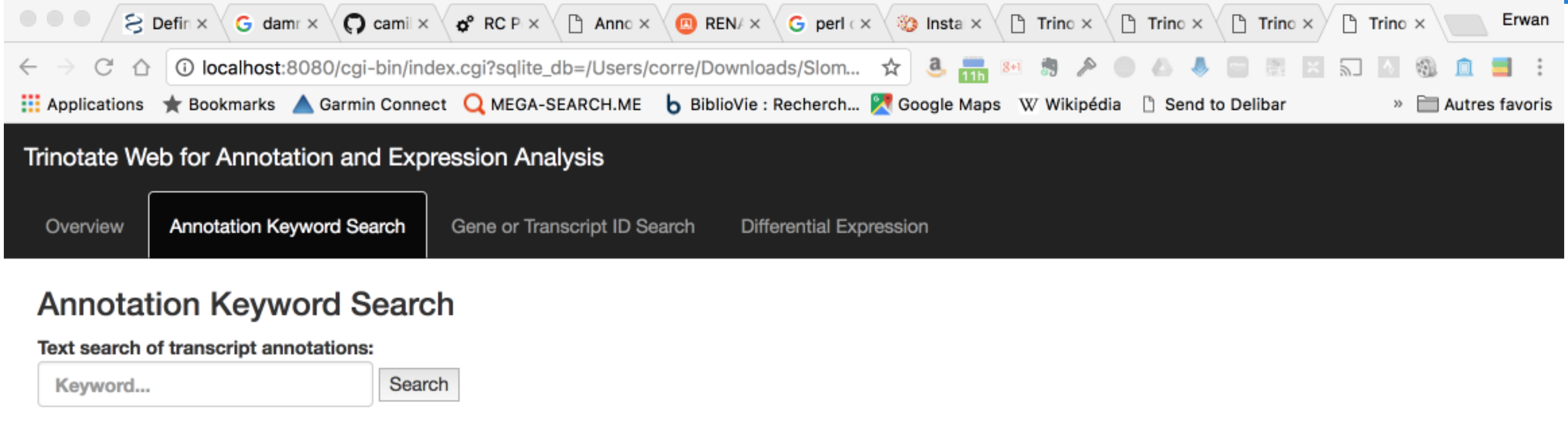
Trinotate web : **Graphical Interface for Navigating Trinotate Annotations and Expression Analyses**

Note, Trinotate is not yet a full-featured application, but is instead in a very early state of development since 5-6 years .. :/

Dependancy
Lighttpd

Perl
Perl DBI, Perl URI, Perl CGI, Perl HTML::Template, Perl DBD::SQLite

Trinotate web



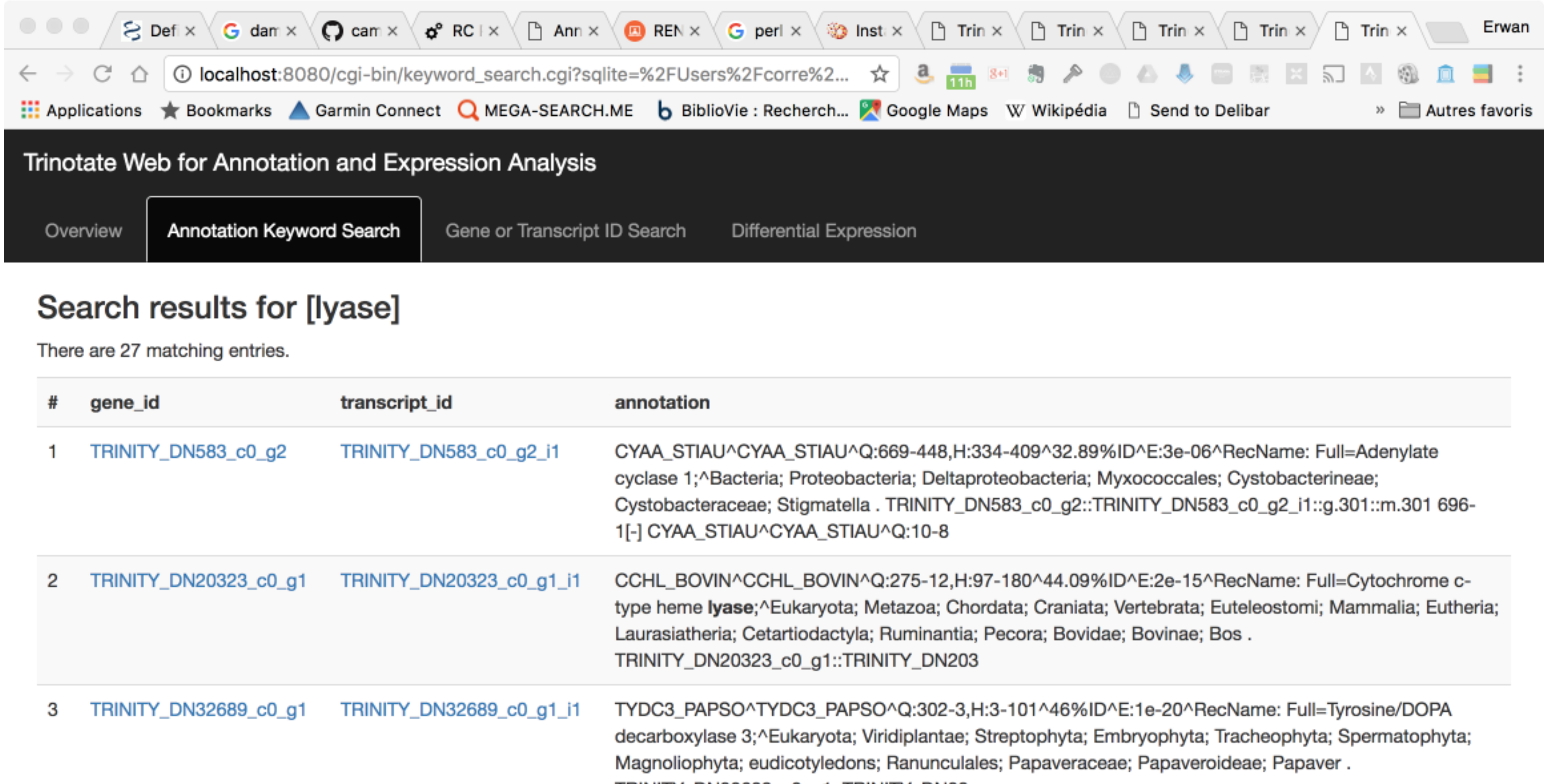
localhost:8080/cgi-bin/index.cgi?sqlite_db=/Users/corre/Downloads/Slom...

Trinotate Web for Annotation and Expression Analysis

Annotation Keyword Search

Text search of transcript annotations:

Keyword...



localhost:8080/cgi-bin/keyword_search.cgi?sqlite=%2FUsers%2Fcorre%2...

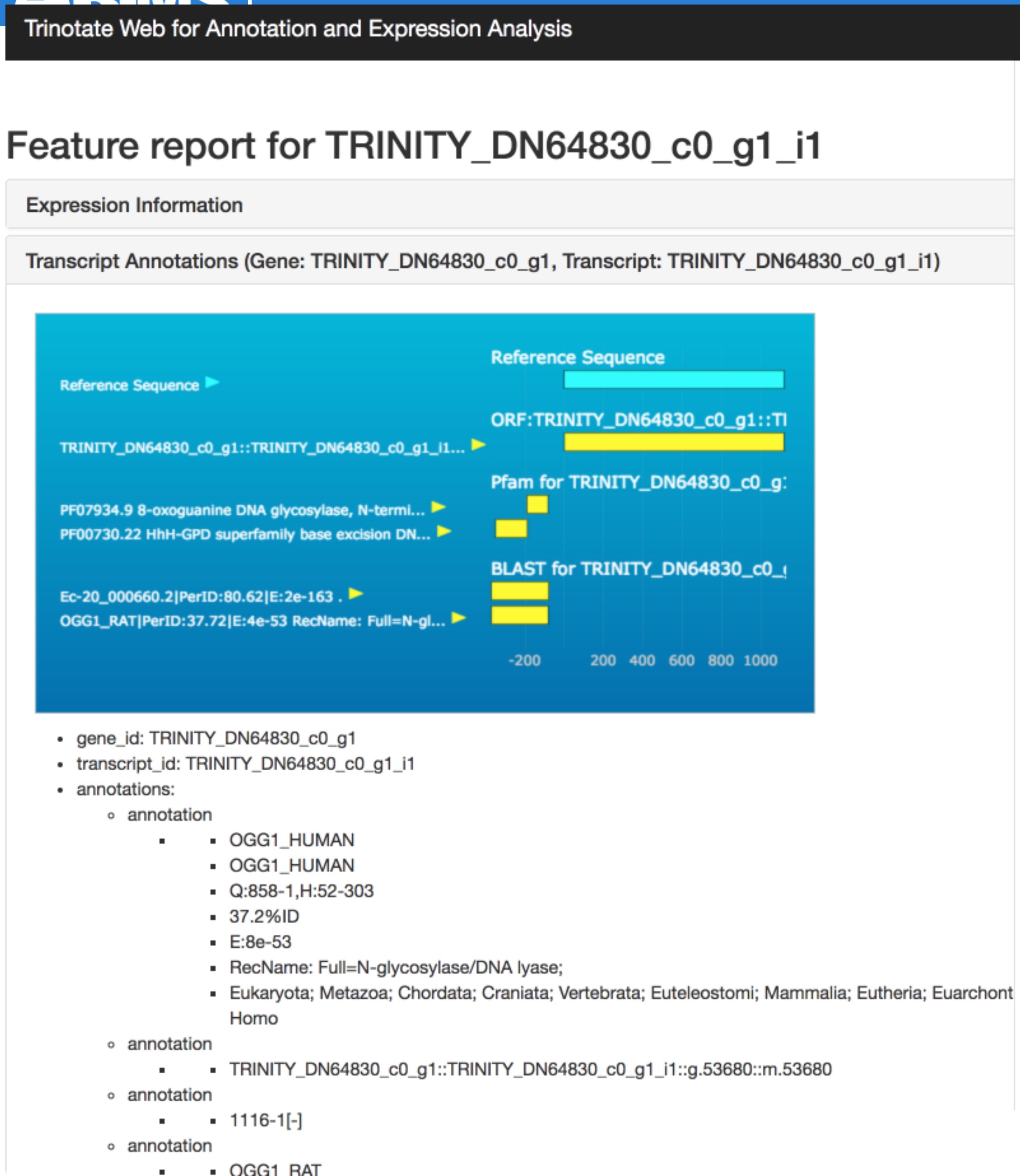
Trinotate Web for Annotation and Expression Analysis

Search results for [lyase]

There are 27 matching entries.

#	gene_id	transcript_id	annotation
1	TRINITY_DN583_c0_g2	TRINITY_DN583_c0_g2_i1	CYAA_STIAU^CYAA_STIAU^Q:669-448,H:334-409^32.89%ID^E:3e-06^RecName: Full=Adenylate cyclase 1;^Bacteria; Proteobacteria; Deltaproteobacteria; Myxococcales; Cystobacterineae; Cystobacteraceae; Stigmatella . TRINITY_DN583_c0_g2::TRINITY_DN583_c0_g2_i1::g.301::m.301 696-1[-] CYAA_STIAU^CYAA_STIAU^Q:10-8
2	TRINITY_DN20323_c0_g1	TRINITY_DN20323_c0_g1_i1	CCHL_BOVIN^CCHL_BOVIN^Q:275-12,H:97-180^44.09%ID^E:2e-15^RecName: Full=Cytochrome c-type heme **lyase**;^Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos . TRINITY_DN20323_c0_g1::TRINITY_DN203
3	TRINITY_DN32689_c0_g1	TRINITY_DN32689_c0_g1_i1	TYDC3_PAPSO^TYDC3_PAPSO^Q:302-3,H:3-101^46%ID^E:1e-20^RecName: Full=Tyrosine/DOPA decarboxylase 3;^Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Ranunculales; Papaveraceae; Papaveroideae; Papaver .

Trinotate web



Trinotate Web for Annotation and Expression Analysis

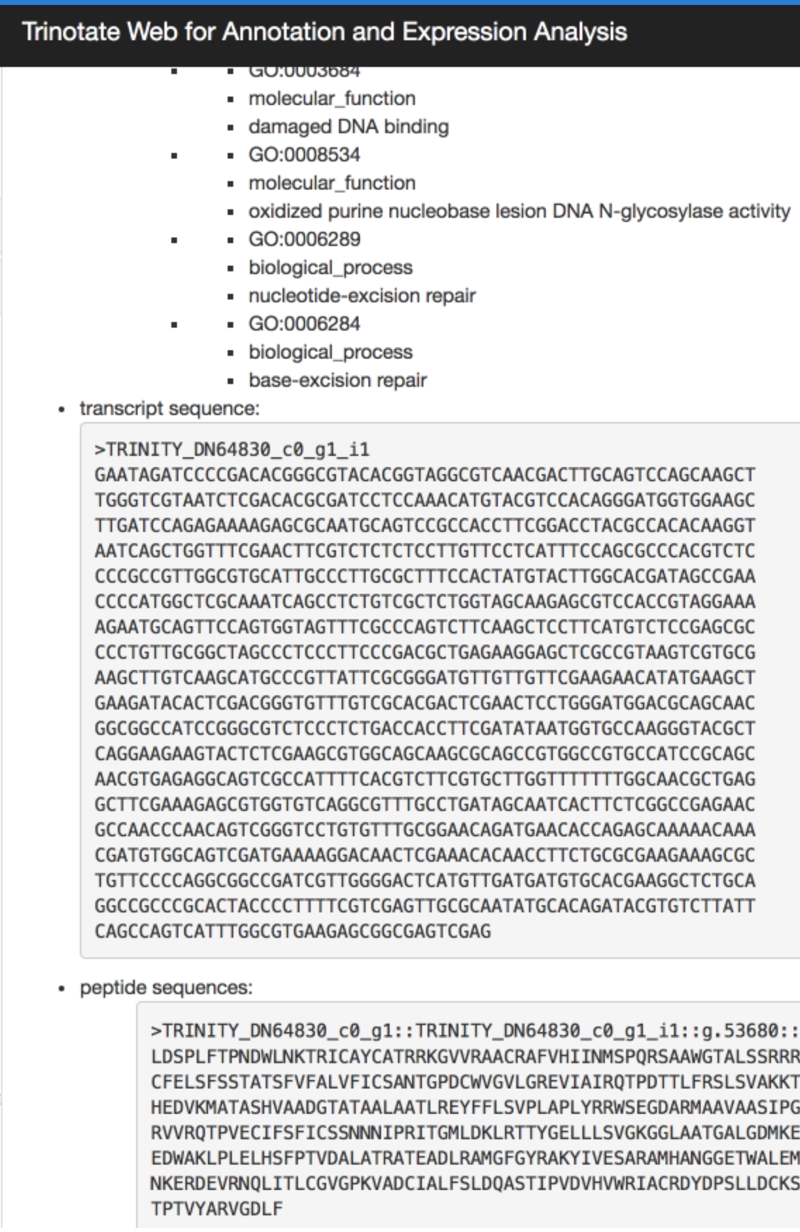
Feature report for TRINITY_DN64830_c0_g1_i1

Expression Information

Transcript Annotations (Gene: TRINITY_DN64830_c0_g1, Transcript: TRINITY_DN64830_c0_g1_i1)

Reference Sequence
Reference Sequence
TRINITY_DN64830_c0_g1::TRINITY_DN64830_c0_g1_i1...
ORF:TRINITY_DN64830_c0_g1::TI
PF07934.9 8-oxoguanine DNA glycosylase, N-termi...
PF00730.22 HhH-GPD superfamily base excision DN...
Pfam for TRINITY_DN64830_c0_g
Ec-20_000660.2|PerID:80.62|E:2e-163 .
OGG1_RAT|PerID:37.72|E:4e-53 RecName: Full=N-gl...
BLAST for TRINITY_DN64830_c0_
-200
200
400
600
800
1000

- gene_id: TRINITY_DN64830_c0_g1
- transcript_id: TRINITY_DN64830_c0_g1_i1
- annotations:
  - annotation
    -
      - OGG1_HUMAN
      - OGG1_HUMAN
      - Q:858-1,H:52-303
      - 37.2%ID
      - E:8e-53
      - RecName: Full=N-glycosylase/DNA lyase;
      - Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchont Homo
  - annotation
    -
      - TRINITY_DN64830_c0_g1::TRINITY_DN64830_c0_g1_i1::g.53680::m.53680
  - annotation
    -
      - 1116-1[-]
  - annotation
    -
      - OGG1_RAT



Trinotate Web for Annotation and Expression Analysis

-
  - GO:0003684
  - molecular_function
  - damaged DNA binding
-
  - GO:0008534
  - molecular_function
  - oxidized purine nucleobase lesion DNA N-glycosylase activity
-
  - GO:0006289
  - biological_process
  - nucleotide-excision repair
-
  - GO:0006284
  - biological_process
  - base-excision repair

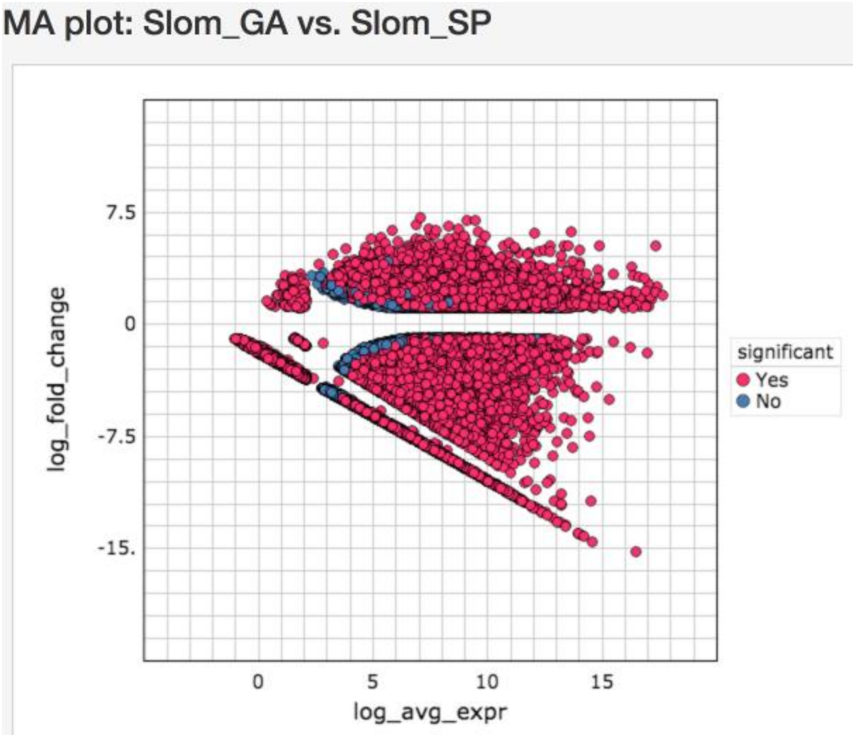
- transcript sequence:

```
>TRINITY_DN64830_c0_g1_i1
GAATAGATCCCCGACACGGGCGTACACGGTAGGCGTCAACGACTTGCAGTCCAGCAAGCT
TGGGTCGTAATCTCGACACGCGATCCTCCAAACATGTACGTCCACAGGGATGGTGGAAGC
TTGATCCAGAGAAAAGAGCGCAATGCAGTCCGCCACCTTCGGACCTACGCCACACAAGGT
AATCAGCTGGTTTCGAACTTCGTCTCTCTCCTTGTTCCTCATTTCCAGCGCCCACGTCTC
CCCGCCGTTGGCGTGCATTGCCCTTGCGCTTTCCACTATGTACTTGGCACGATAGCCGAA
CCCCATGGCTCGCAAATCAGCCTCTGTCGCTCTGGTAGCAAGAGCGTCCACCGTAGGAAA
AGAATGCAGTTCCAGTGGTAGTTTCGCCCAGTCTTCAAGCTCCTTCATGTCTCCGAGCGC
CCCTGTTGCGGCTAGCCCTCCCTTCCCGACGCTGAGAAGGAGCTCGCCGTAAGTCGTGCG
AAGCTTGTCAAGCATGCCCGTTATTCGCGGGATGTTGTTGTTCGAAGAACATATGAAGCT
GAAGATACACTCGACGGGTGTTTGTCGCACGACTCGAACTCCTGGGATGGACGCAGCAAC
GGCGGCCATCCGGGCGTCTCCCTCTGACCACCTTCGATATAATGGTGCCAAGGGTACGCT
CAGGAAGAAGTACTCTCGAAGCGTGGCAGCAAGCGCAGCCGTGGCCGTGCCATCCGCAGC
AACGTGAGAGGCAGTCGCCATTTTCACGTCTTCGTGCTTGGTTTTTTTGGCAACGCTGAG
GCTTCGAAAGAGCGTGGTGTCAGGCGTTTGCCTGATAGCAATCACTTCTCGGCCGAGAAC
GCCAACCCAACAGTCGGGTCCTGTGTTTGCGGAACAGATGAACACCAGAGCAAAAACAAA
CGATGTGGCAGTCGATGAAAAGGACAACTCGAAACACAACCTTCTGCGCGAAGAAAGCGC
TGTTCCCCAGGCGGCCGATCGTTGGGGACTCATGTTGATGATGTGCACGAAGGCTCTGCA
GGCCGCCCGCACTACCCCTTTTCGTCGAGTTGCGCAATATGCACAGATACGTGTCTTATT
CAGCCAGTCATTTGGCGTGAAGAGCGGCGAGTCGAG
```

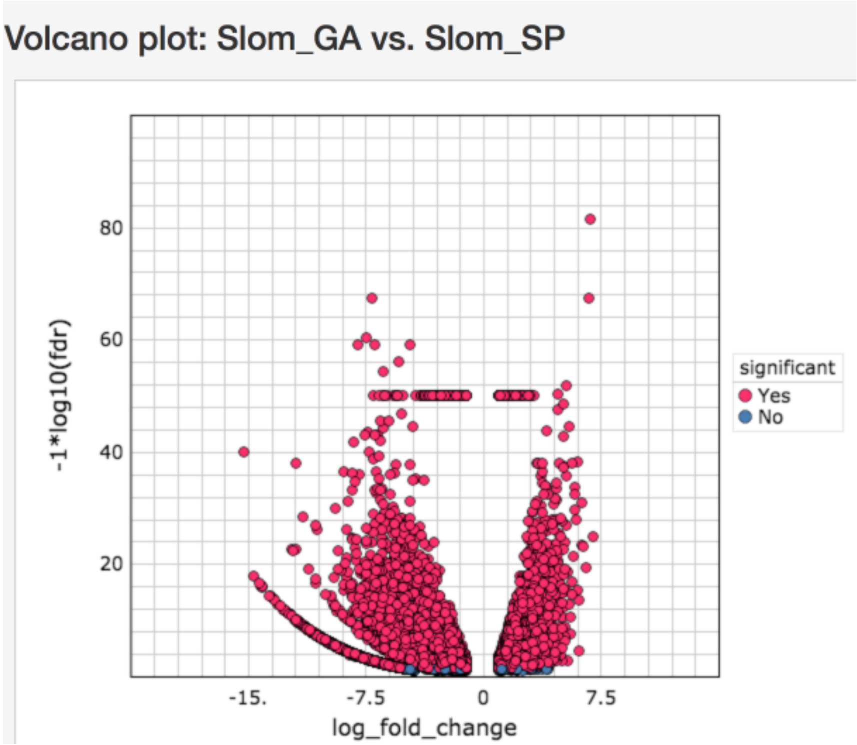
- peptide sequences:

```
>TRINITY_DN64830_c0_g1::TRINITY_DN64830_c0_g1_i1::g.53680::
LDSPLFTPNDWLNKTRICAYCATRRKGVVRAACRAFVHIINMSPQRSAAWGTALSSRRR
CFELSFSSTATSFVFALVFICSANTGPDCWVGVLGREVIAIRQTPDTTLFRSLSVAKKT
HEDVKMATASHVAADGTATAALAATLREYFFLSVPLAPLYRRWSEGDARMAAVAASIPG
RVVRQTPVECIFSFICSSNNNIPRITGMLDKLRTTYGELLLSVGKGGLAATGALGDMKE
EDWAKLPLELHSFPTVDALATRATEADLRAMGFGYRAKYIVESARAMHANGGETWALEM
NKERDEVRNQLITLCGVGPKVADCIALFSLDQASTIPVDVHVWRIACRDYDPSLLDCKS
TPTVYARVGDLF
```

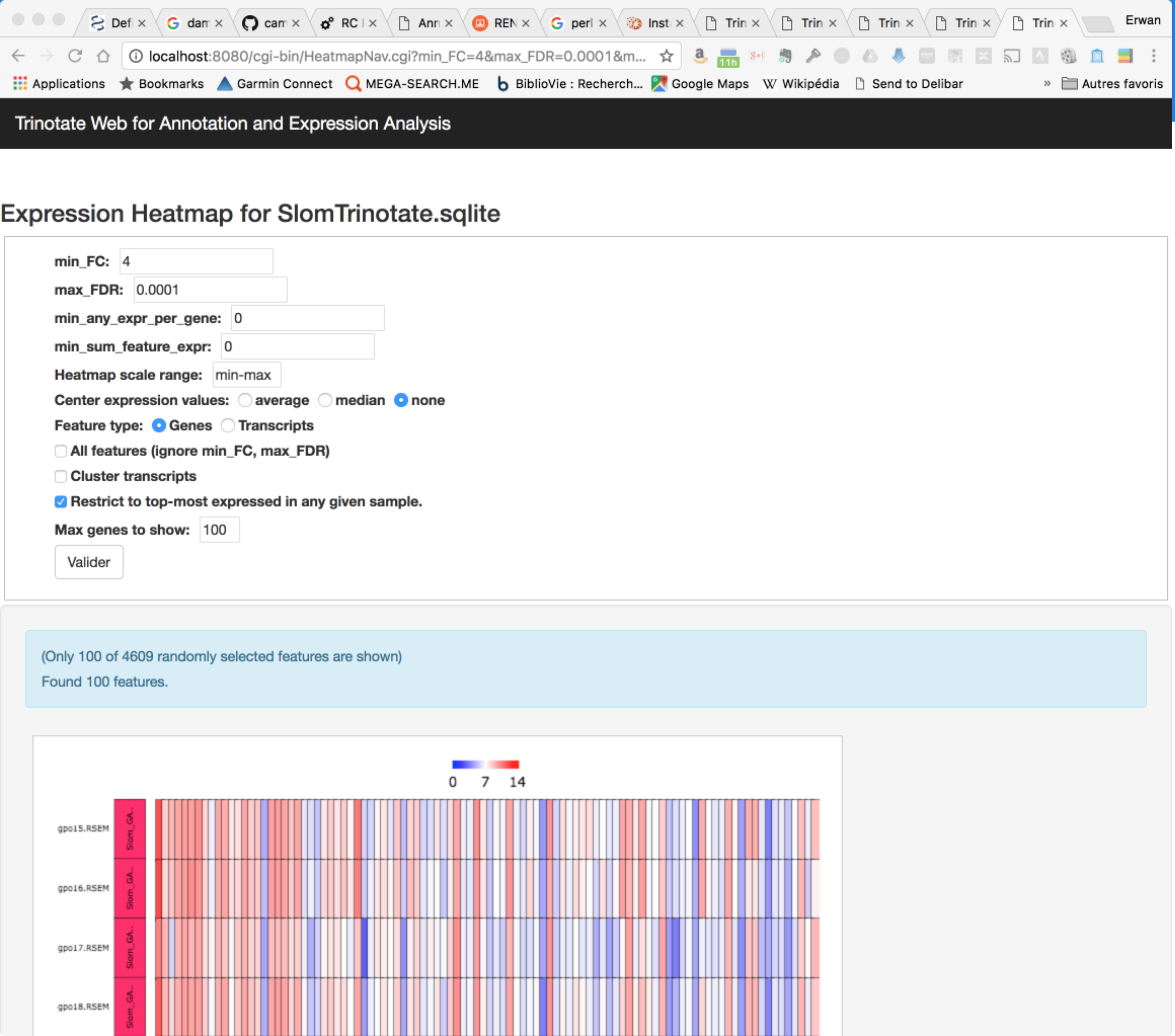
Trinotate web



MA plot: Slom_GA vs. Slom_SP



Volcano plot: Slom_GA vs. Slom_SP



Expression Heatmap for SlomTrinotate.sqlite

min_FC: 4

max_FDR: 0.0001

min_any_expr_per_gene: 0

min_sum_feature_expr: 0

Heatmap scale range: min-max

Center expression values: ○ average ○ median ● none

Feature type: ● Genes ○ Transcripts

□ **All features (ignore min_FC, max_FDR)**

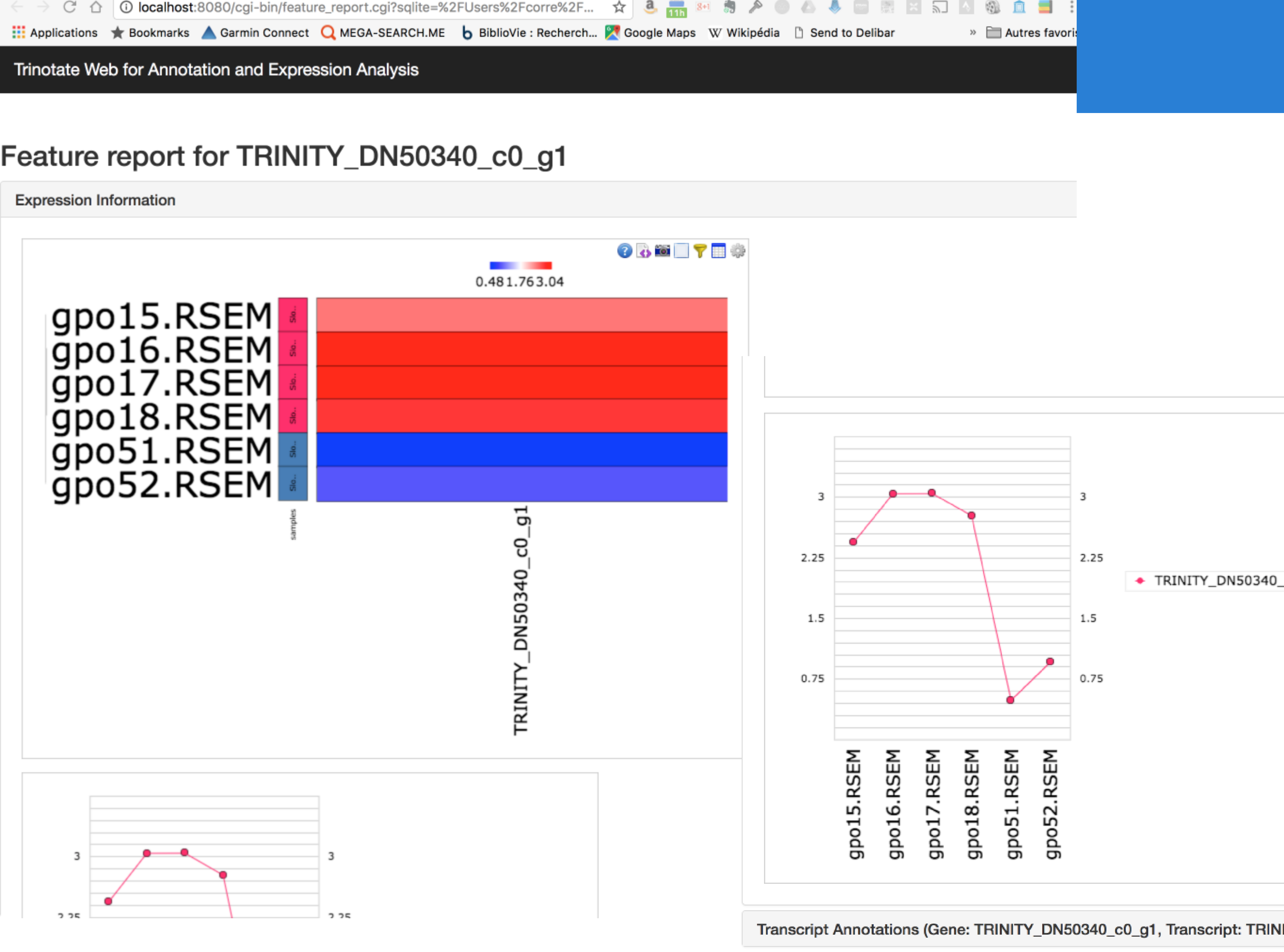
□ **Cluster transcripts**

☑ **Restrict to top-most expressed in any given sample.**

Max genes to show: 100

(Only 100 of 4609 randomly selected features are shown)

Found 100 features.



Trinotate Web for Annotation and Expression Analysis

Feature report for TRINITY_DN50340_c0_g1

Expression Information

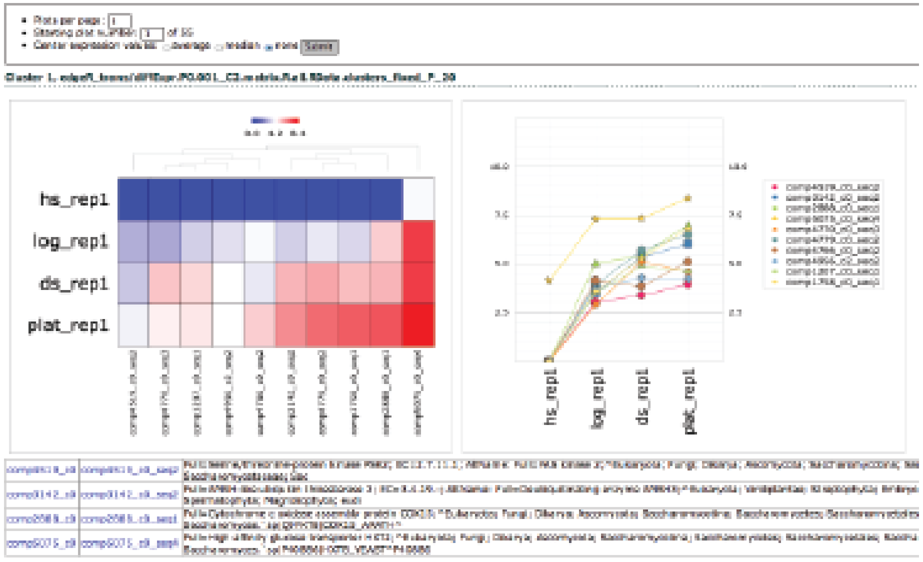
Transcript Annotations (Gene: TRINITY_DN50340_c0_g1, Transcript: TRIN

CNRS • SORBONNE UNIVERSITÉ
Station Biologique de Roscoff

ABiMS

Transcript/Protein Annotation Report

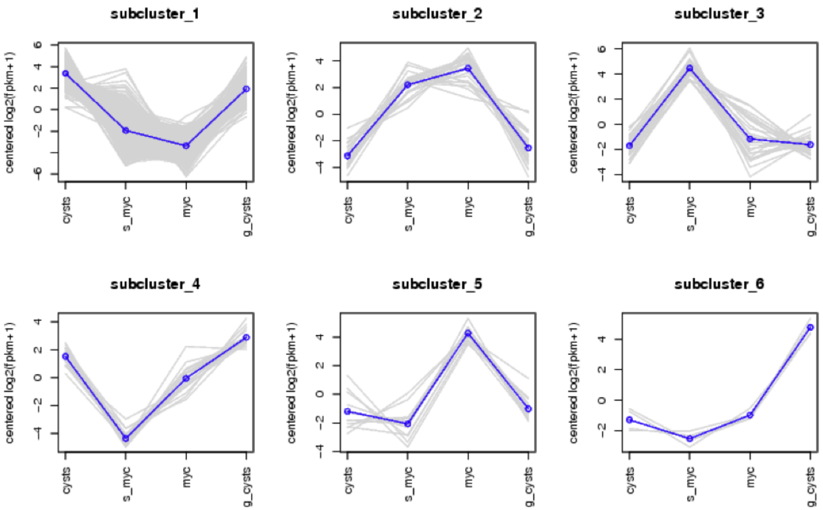
Clustered Expression Profiles



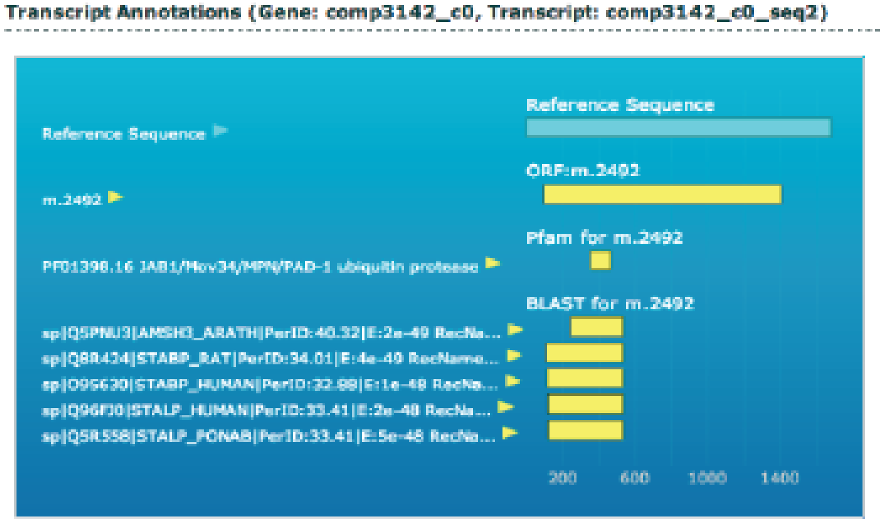
- Plots per page: 1
- Starting plot number: 1 of 35
- Center expression values: ○ average ○ median ● none [Submit]

Cluster 1, edgeR_trans/diffExpr.P0.001_C2.matrix.RData.clusters_fixed_P_20

comp4328_c0	comp4328_c0_seq2	Full=Serine/threonine-protein kinase [illegible]; EC 2.7.11.1; AltName: Full=[illegible] kinase 2;^Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; [illegible] Saccharomycetales [illegible]
comp3142_c0	comp3142_c0_seq2	Full=AMSH-like ubiquitin thioesterase 3; EC=3.4.19.-; AltName: Full=Deubiquitinating enzyme AMSH3;^Eukaryota; Viridiplantae; Streptophyta; Embryophyta; [illegible] Spermatophyta; Magnoliophyta; eudi
comp2888_c0	comp2888_c0_seq1	Full=Cytochrome c oxidase assembly protein COX15;^Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomyces; [illegible]
comp5075_c0	comp5075_c0_seq4	Full=High-affinity glucose transporter HXT2;^Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces [illegible]

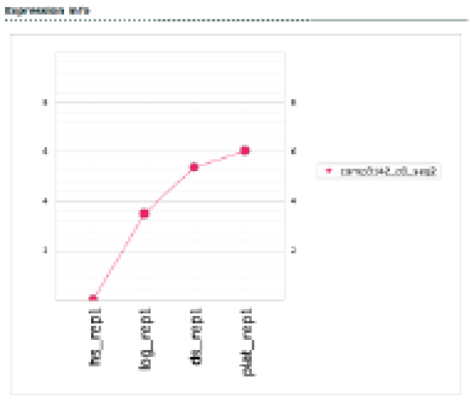


Blast Hits, Pfam Domains, etc.



Transcript Annotations (Gene: comp3142_c0, Transcript: comp3142_c0_seq2)

Reference Sequence
Reference Sequence
m.2492
ORF:m.2492
PF01398.16 JAB1/Mov34/MPN/PAD-1 ubiquitin protease
Pfam for m.2492
BLAST for m.2492
sp|Q5PNU3|AMSH3_ARATH|PerID:40.32|E:2e-49 RecNa...
sp|Q8R424|STABP_RAT|PerID:34.01|E:4e-49 RecName...
sp|O95630|STABP_HUMAN|PerID:32.88|E:1e-48 RecNa...
sp|Q96FJ0|STALP_HUMAN|PerID:33.41|E:2e-48 RecNa...
sp|Q5R558|STALP_PONAB|PerID:33.41|E:5e-48 RecNa...
200
600
1000
1400



Expression info

- transcript sequence:

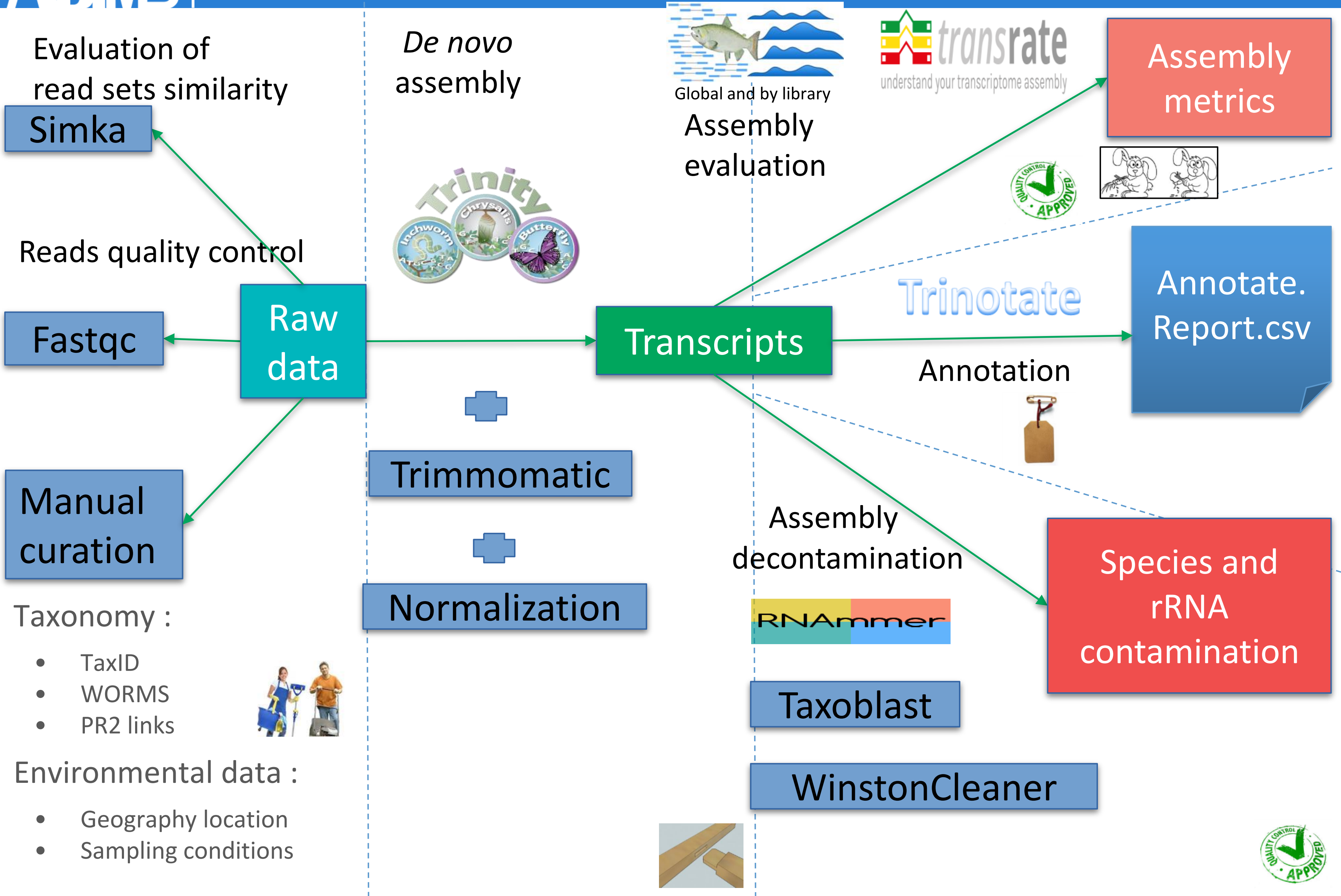
>comp3142_c0_seq2
[illegible]

- peptide sequences:
  - m.2492
[illegible]

Individual Transcript Expression Profiles

Transcript and Protein Sequence

Homogeneous pipeline



Evaluation of read sets similarity

Simka

Reads quality control

Fastqc

Raw data

Manual curation

Taxonomy :

- TaxID
- WORMS
- PR2 links

Environmental data :

- Geography location
- Sampling conditions

De novo assembly

Trimmomatic

Normalization

Transcripts

Global and by library

Assembly evaluation

Assembly metrics

Annotation

Annotate. Report.csv

Assembly decontamination

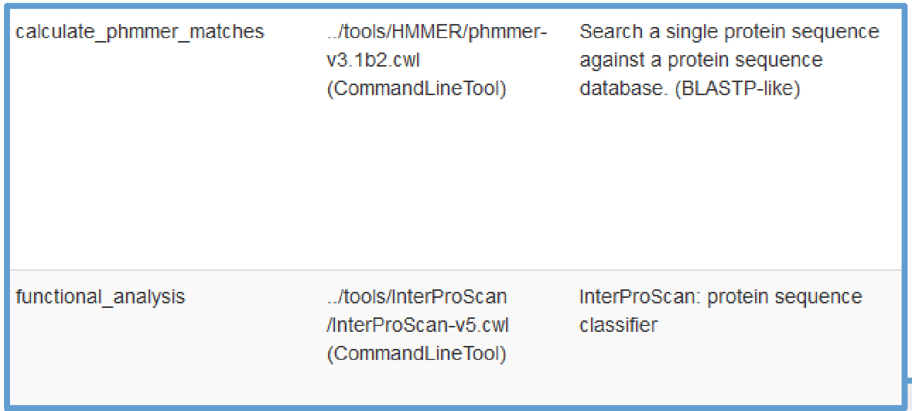
Taxoblast

WinstonCleaner

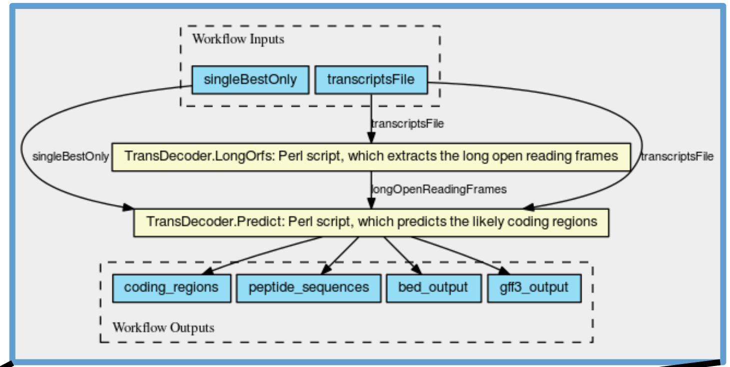
Species and rRNA contamination

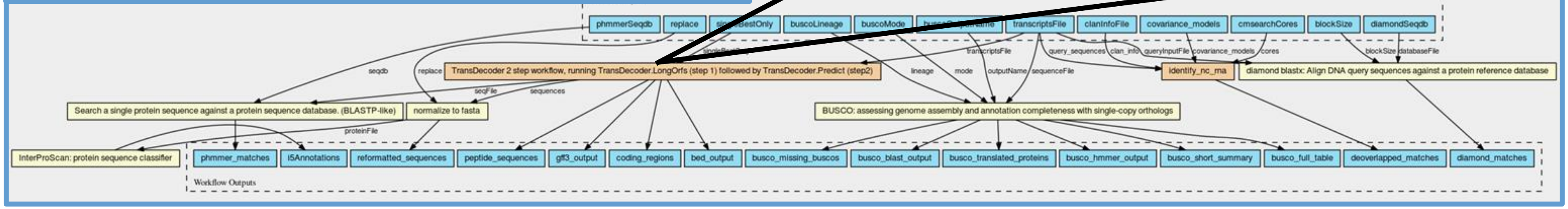
Transcripts annotation workflow - Visualisation

Reusing and enhancing Common Workflow Language (CWL) described analysis pipelines to enrich marine reference data



calculate_phmmer_matches	../tools/HMMER/phmmer-v3.1b2.cwl (CommandLineTool)	Search a single protein sequence against a protein sequence database. (BLASTP-like)
functional_analysis	../tools/InterProScan/InterProScan-v5.cwl (CommandLineTool)	InterProScan: protein sequence classifier





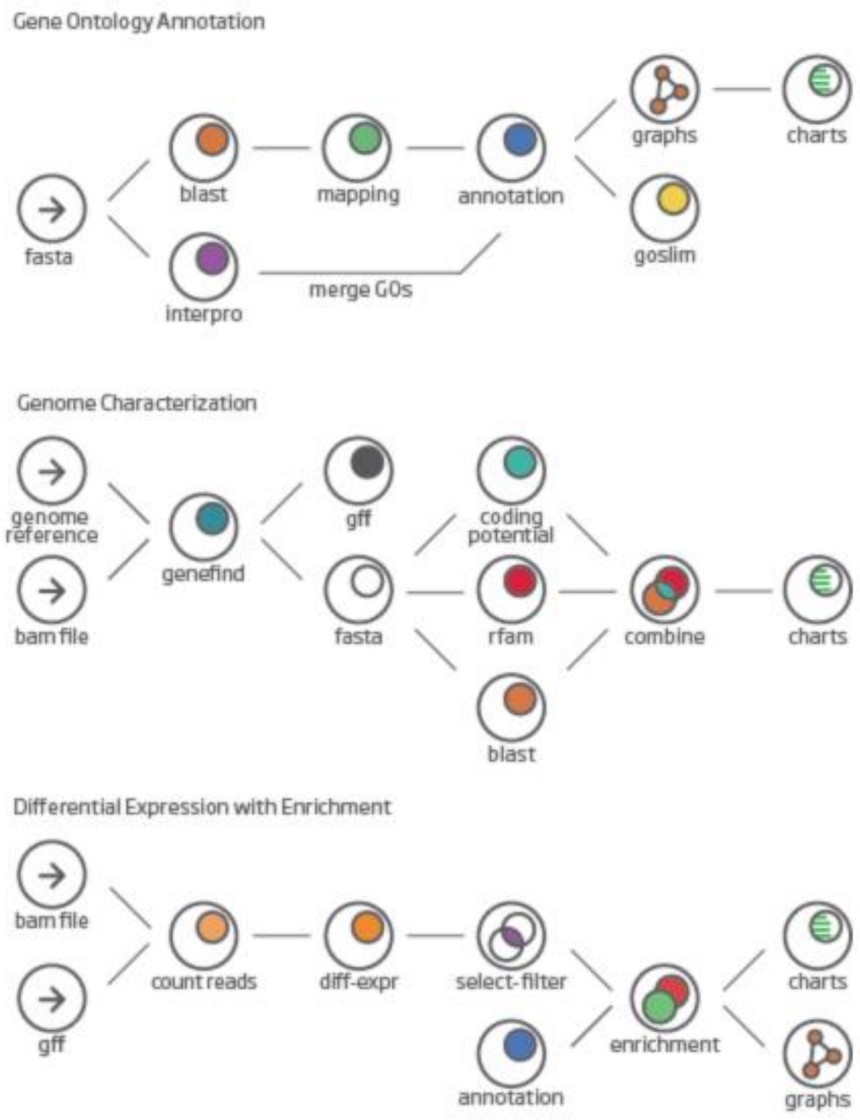
Galaxy PROJECT

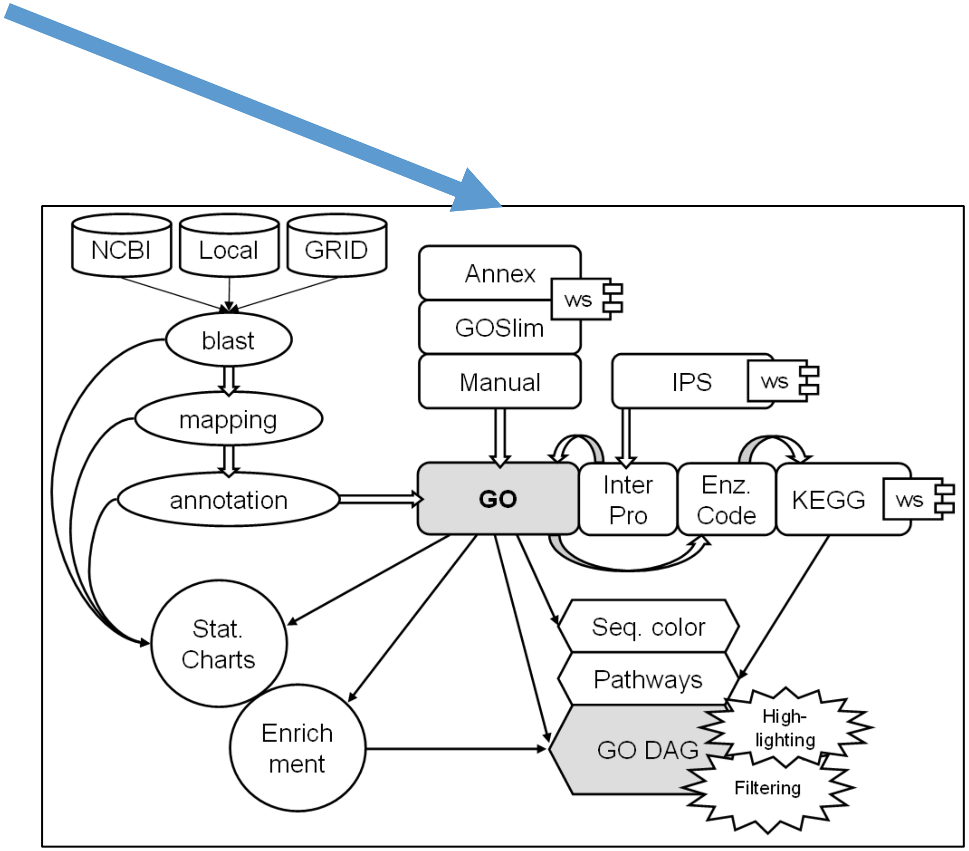
nextflow

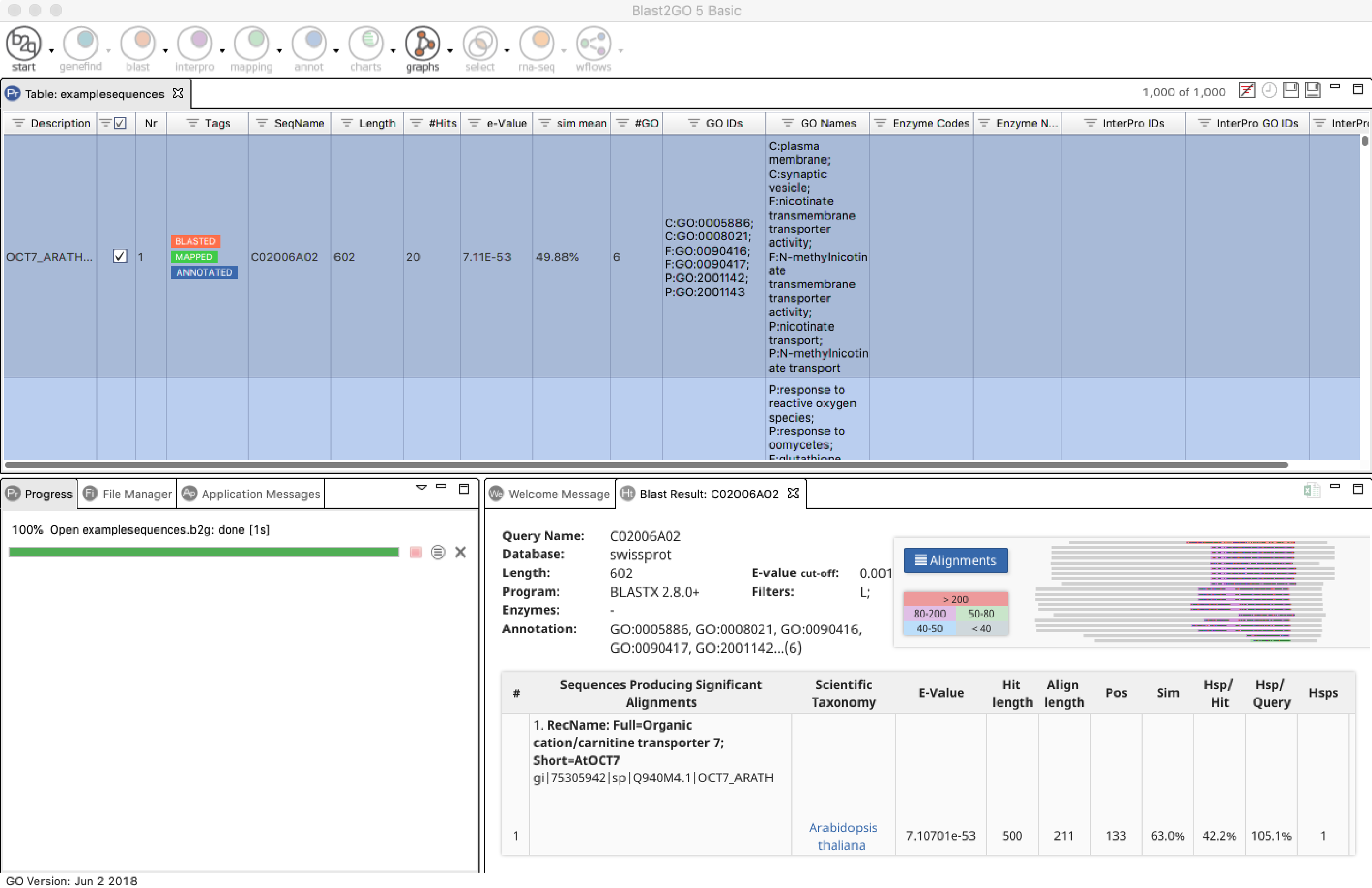
Some alternatives

- Blast2Go
- FunctionAnnotator
- Annoscript
- Dammit
- KOBAS
- EnTAP

Blast2GO Schema







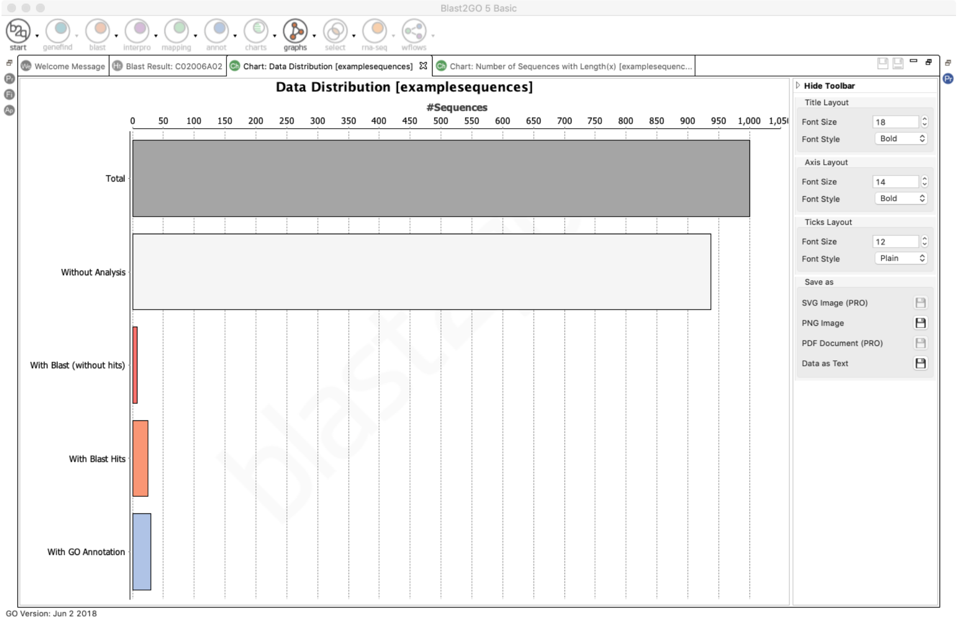
Description	☑	Nr	Tags	SeqName	Length	#Hits	e-Value	sim mean	#GO	GO IDs	GO Names	Enzyme Codes	Enzyme N...	InterPro IDs	InterPro GO IDs	InterPr
OCT7_ARATH...	☑	1	BLASTED MAPPED ANNOTATED	C02006A02	602	20	7.11E-53	49.88%	6	C:GO:0005886; C:GO:0008021; F:GO:0090416; F:GO:0090417; P:GO:2001142; P:GO:2001143	C:plasma membrane; C:synaptic vesicle; F:nicotinate transmembrane transporter activity; F:N-methylnicotinate transmembrane transporter activity; P:nicotinate transport; P:N-methylnicotinate transport					
											P:response to reactive oxygen species; P:response to oomycetes; F:glutathione					

100% Open examplesequences.b2g: done [1s]

Query Name: C02006A02
Database: swissprot
Length: 602 **E-value cut-off:** 0.001
Program: BLASTX 2.8.0+ **Filters:** L;
Enzymes: -
Annotation: GO:0005886, GO:0008021, GO:0090416, GO:0090417, GO:2001142...(6)

#	Sequences Producing Significant Alignments	Scientific Taxonomy	E-Value	Hit length	Align length	Pos	Sim	Hsp/ Hit	Hsp/ Query	Hsps
1	1. **RecName: Full=Organic cation/carnitine transporter 7; Short=AtOCT7** gi|75305942|sp|Q940M4.1|OCT7_ARATH	Arabidopsis thaliana	7.10701e-53	500	211	133	63.0%	42.2%	105.1%	1

GO Version: Jun 2 2018



Blast2GO 5 Basic

Data Distribution [examplesequences]

Hide Toolbar

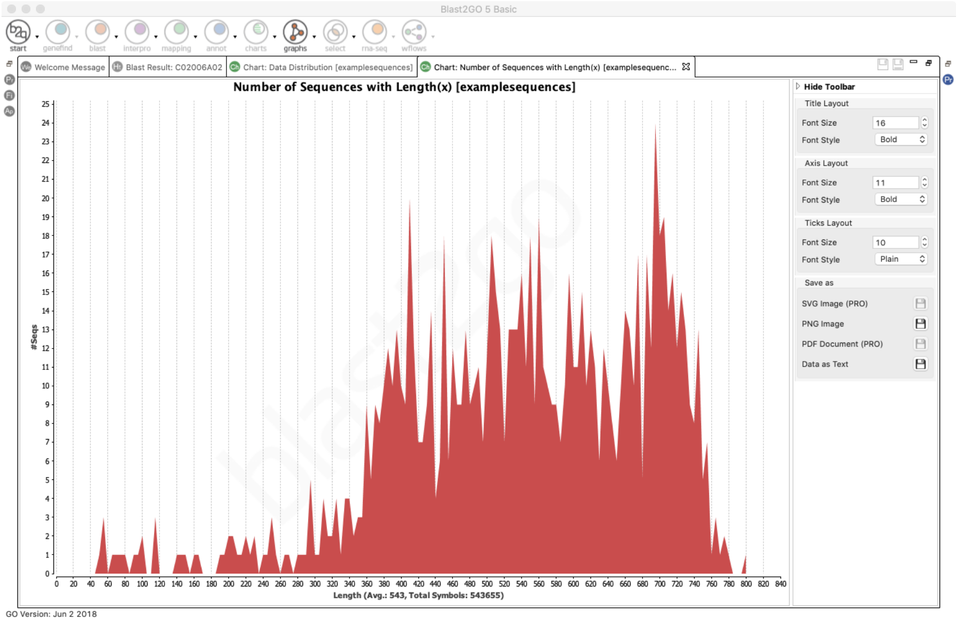
Title Layout
Font Size 18
Font Style Bold

Axis Layout
Font Size 14
Font Style Bold

Ticks Layout
Font Size 12
Font Style Plain

Save as
SVG Image (PRO)
PNG Image
PDF Document (PRO)
Data as Text

GO Version: Jun 2 2018



Blast2GO 5 Basic

Number of Sequences with Length(x) [examplesequences]

Hide Toolbar

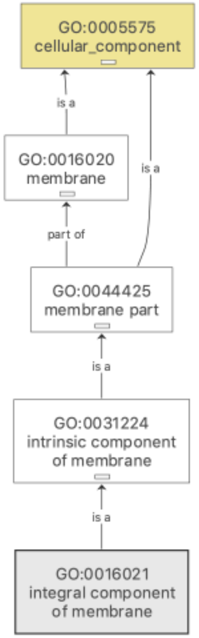
Title Layout
Font Size 16
Font Style Bold

Axis Layout
Font Size 11
Font Style Bold

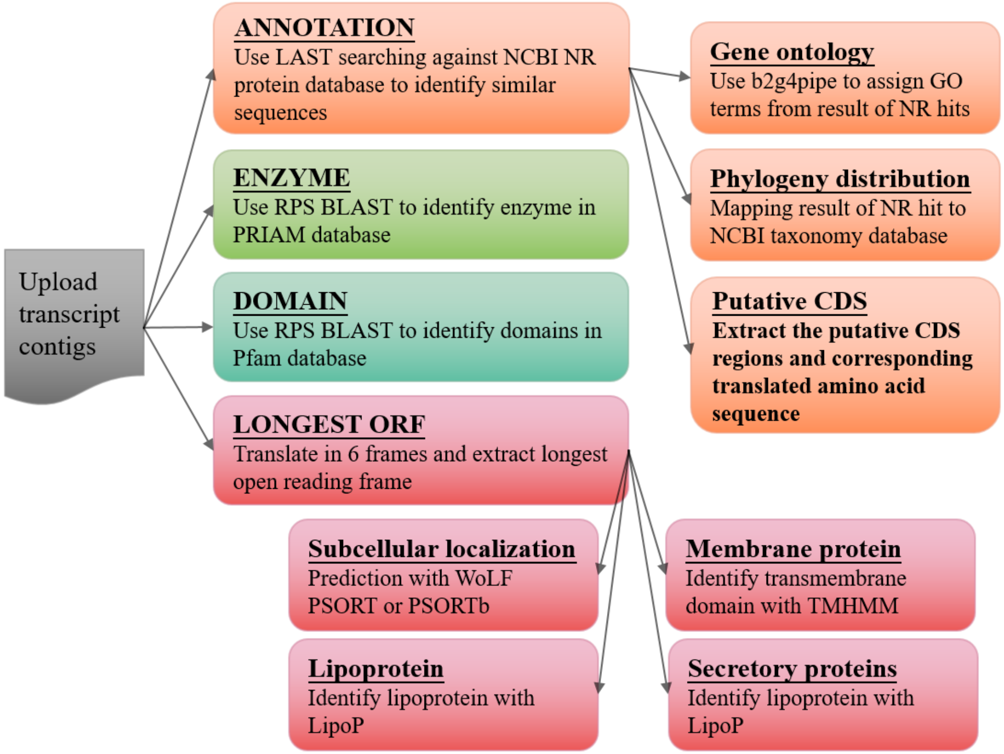
Ticks Layout
Font Size 10
Font Style Plain

Save as
SVG Image (PRO)
PNG Image
PDF Document (PRO)
Data as Text

GO Version: Jun 2 2018



*Function*Annotator



Upload transcript contigs

ANNOTATION
Use LAST searching against NCBI NR protein database to identify similar sequences

ENZYME
Use RPS BLAST to identify enzyme in PRIAM database

DOMAIN
Use RPS BLAST to identify domains in Pfam database

LONGEST ORF
Translate in 6 frames and extract longest open reading frame

Gene ontology
Use b2g4pipe to assign GO terms from result of NR hits

Phylogeny distribution
Mapping result of NR hit to NCBI taxonomy database

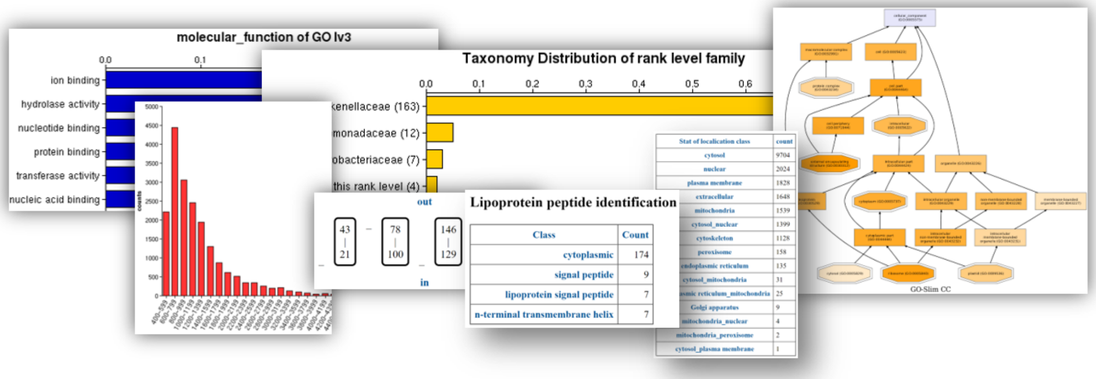
Putative CDS
Extract the putative CDS regions and corresponding translated amino acid sequence

Subcellular localization
Prediction with WoLF PSORT or PSORTb

Membrane protein
Identify transmembrane domain with TMHMM

Lipoprotein
Identify lipoprotein with LipoP

Secretory proteins
Identify lipoprotein with LipoP



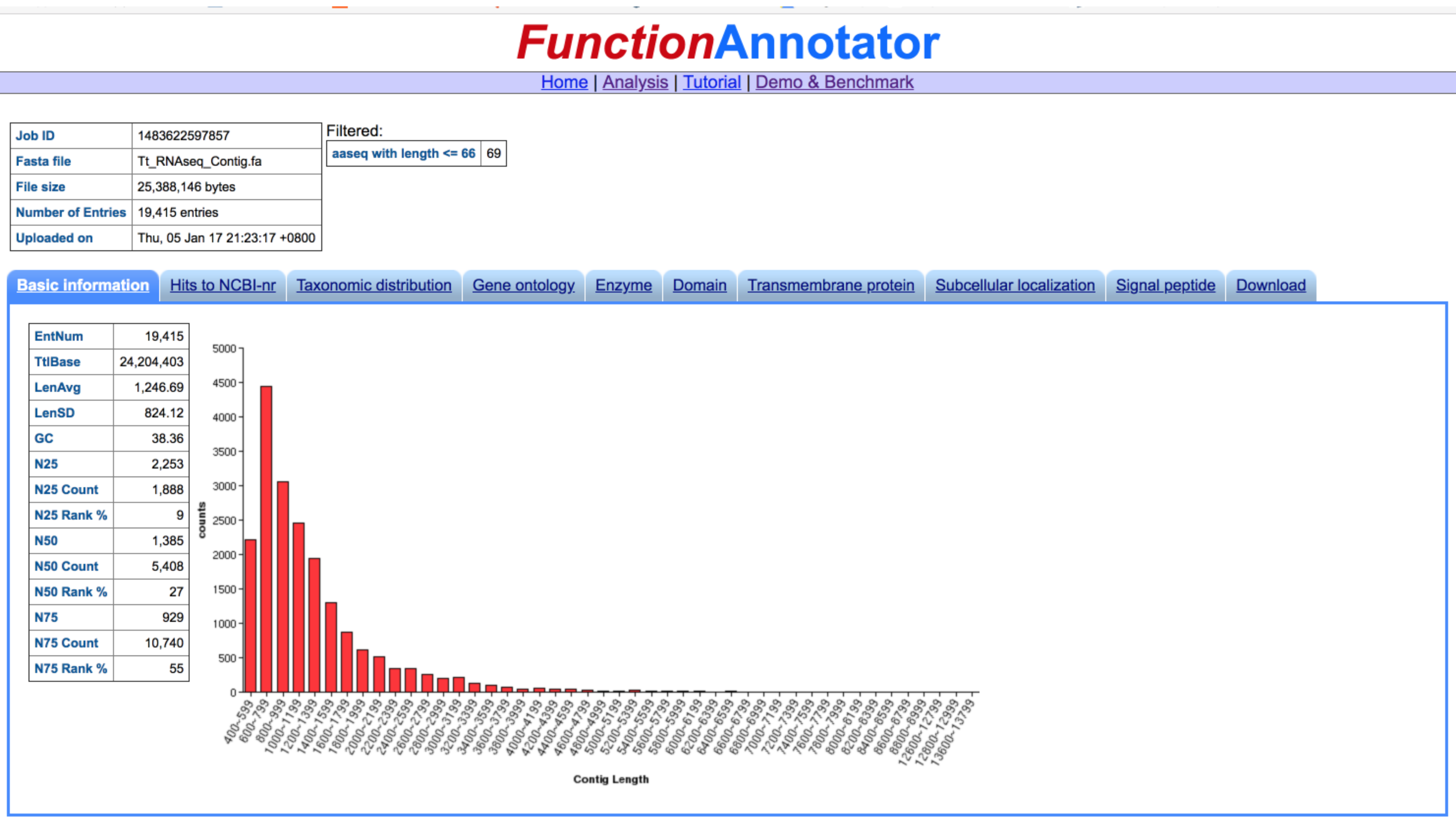
Lipoprotein peptide identification

Class	Count
cytoplasmic	174
signal peptide	9
lipoprotein signal peptide	7
n-terminal transmembrane helix	7

Stat of localication class	count
cytosol	9704
nuclear	2024
plasma membrane	1828
extracellular	1648
mitochondria	1539
cytosol_nuclear	1399
cytoskeleton	1128
peroxisome	158
endoplasmic reticulum	135
cytosol_mitochondria	31
asmic reticulum_mitochondria	25
Golgi apparatus	9
mitochondria_nuclear	4
mitochondria_peroxisome	2
cytosol_plasma membrane	1

Chen TW et al., (2017). FunctionAnnotator, a versatile and efficient web tool for non-model organism annotation. Scientific Reports

*Function*Annotator



*Function*Annotator

Home | Analysis | Tutorial | Demo & Benchmark

Job ID	1483622597857
Fasta file	Tt_RNAseq_Contig.fa
File size	25,388,146 bytes
Number of Entries	19,415 entries
Uploaded on	Thu, 05 Jan 17 21:23:17 +0800

Filtered:

aaseq with length <= 66	69

Basic information | Hits to NCBI-nr | Taxonomic distribution | Gene ontology | Enzyme | Domain | Transmembrane protein | Subcellular localization | Signal peptide | Download

EntNum	19,415
TtlBase	24,204,403
LenAvg	1,246.69
LenSD	824.12
GC	38.36
N25	2,253
N25 Count	1,888
N25 Rank %	9
N50	1,385
N50 Count	5,408
N50 Rank %	27
N75	929
N75 Count	10,740
N75 Rank %	55

Method

The pipeline allows the creation of a comprehensive user-friendly table containing all the annotations produced for each transcript.

The user can choose to annotate her/his transcriptome against selected organisms or the complete database.

https://github.com/frankMusacchia/Annocript

Version 2.0 : April 2018

Method

The proteins most similar to the transcripts are given by the **blastx (blastp** if you use peptides) analyses against the UniProt databases **SwissProt and TrEMBL** (or UniRef).

Blastn (tblastn) against a concatenation of the **SILVA database** (small and large subunits ribosomal RNAs) and the **Rfam database** allows to check for ribosomal and other short noncoding RNAs.

Rpstblastn (rpsblast) returns information about **the Conserved Domains Database** within each transcript.

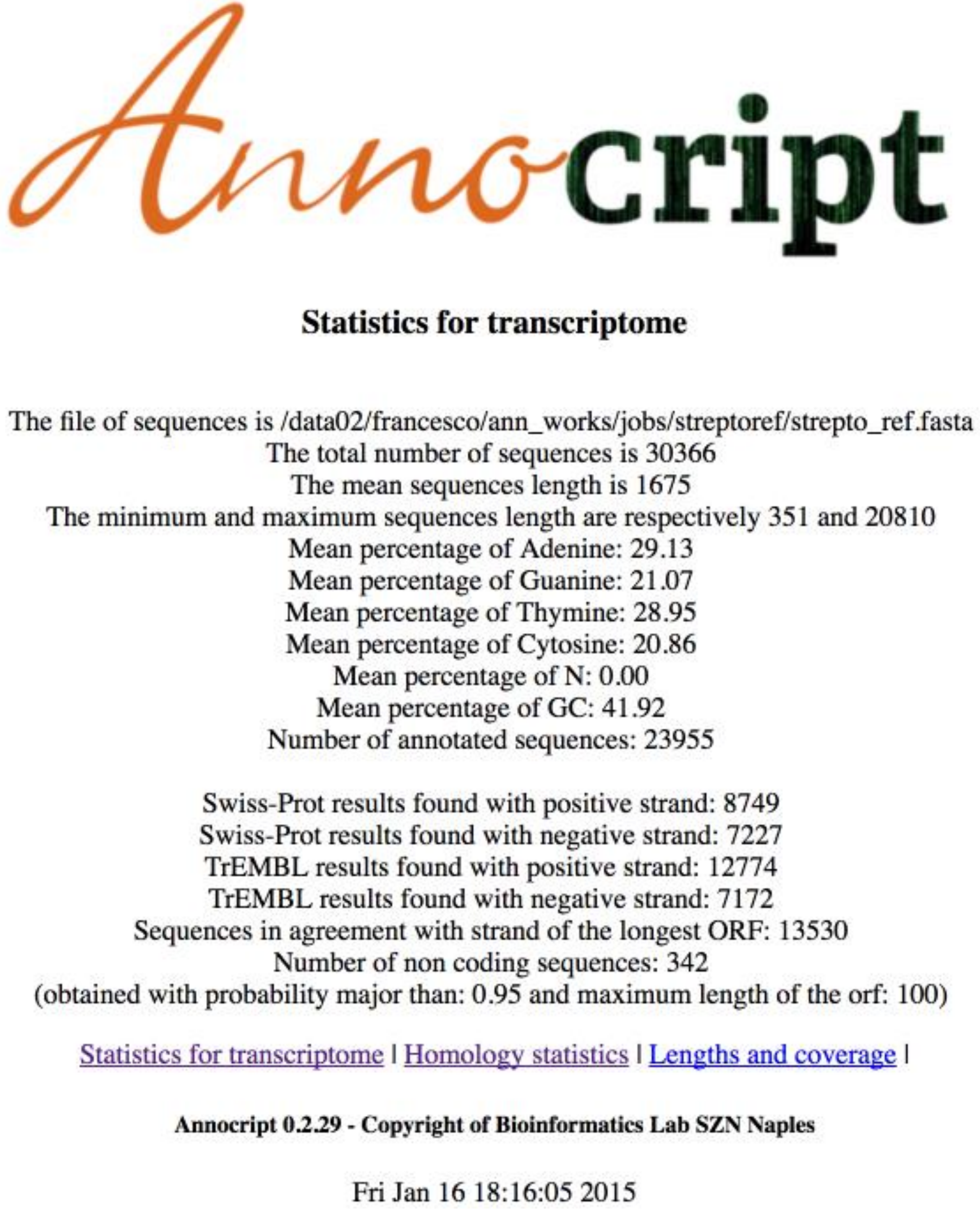
Mapping of GO functional classification is shown using the **best matches between SwissProt and TrEMBL**. If UniRef is used, the GO terms are always taken associated to its result. GO terms can be also associated to Pfam Domains

Mapping of Enzyme Commission IDs and Pathways descriptions are always given associated only to **the SwissProt** id, if present.

Portrait measures the **probability that a sequence is coding or non-coding** and its score, together with a final heuristic, based on the integration of all the results, makes Annocript capable to also identify bona-fide noncoding transcripts.

https://github.com/frankMusacchia/Annocript

Results: statistics



Annocript

Statistics for transcriptome

The file of sequences is /data02/francesco/ann_works/jobs/streptoref/strepto_ref.fasta
The total number of sequences is 30366
The mean sequences length is 1675
The minimum and maximum sequences length are respectively 351 and 20810
Mean percentage of Adenine: 29.13
Mean percentage of Guanine: 21.07
Mean percentage of Thymine: 28.95
Mean percentage of Cytosine: 20.86
Mean percentage of N: 0.00
Mean percentage of GC: 41.92
Number of annotated sequences: 23955

Swiss-Prot results found with positive strand: 8749
Swiss-Prot results found with negative strand: 7227
TrEMBL results found with positive strand: 12774
TrEMBL results found with negative strand: 7172
Sequences in agreement with strand of the longest ORF: 13530
Number of non coding sequences: 342
(obtained with probability major than: 0.95 and maximum length of the orf: 100)

Statistics for transcriptome | Homology statistics | Lengths and coverage |

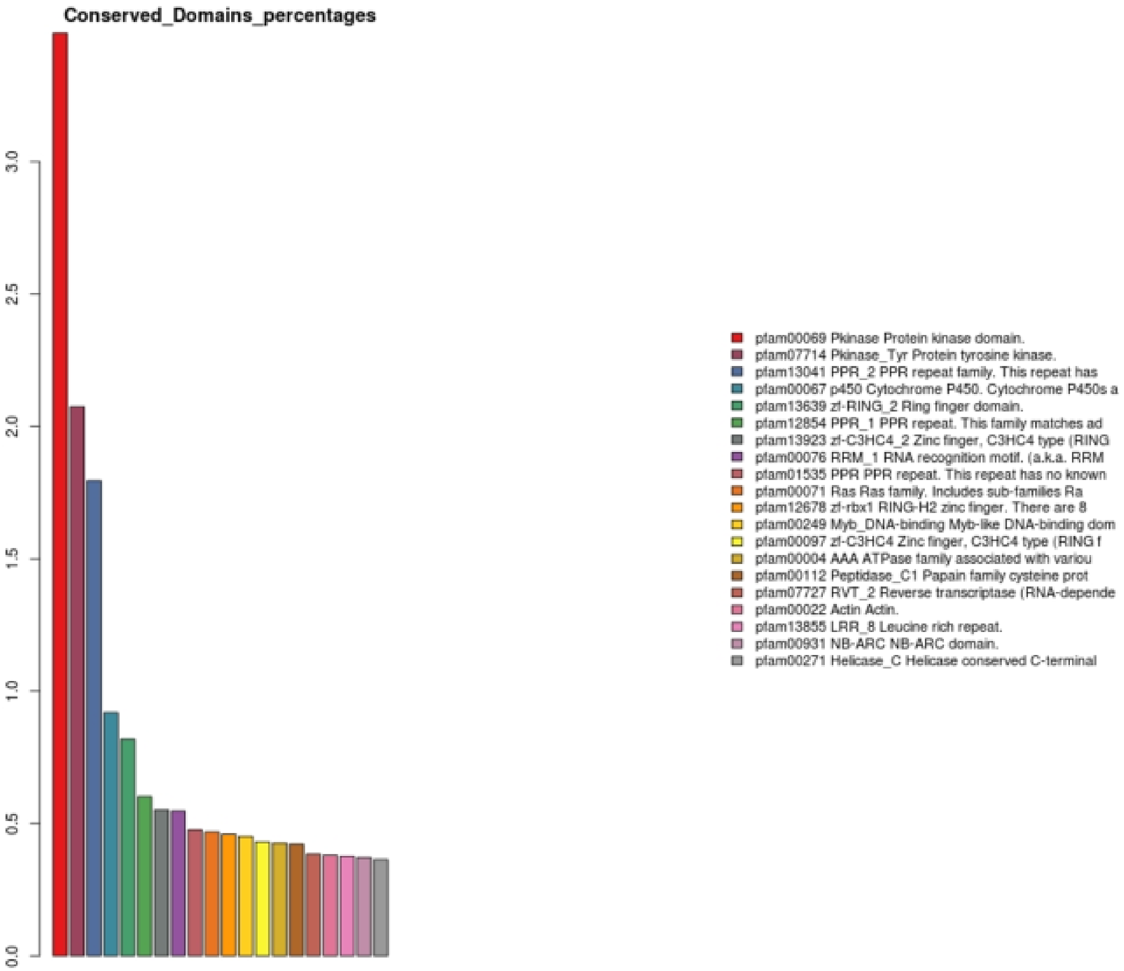
Annocript 0.2.29 - Copyright of Bioinformatics Lab SZN Naples

Fri Jan 16 18:16:05 2015

Results: graphical representation

Annocript

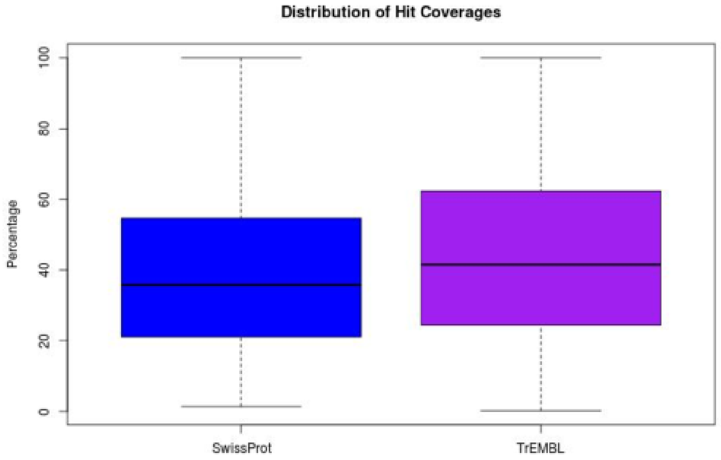
Homology statistics

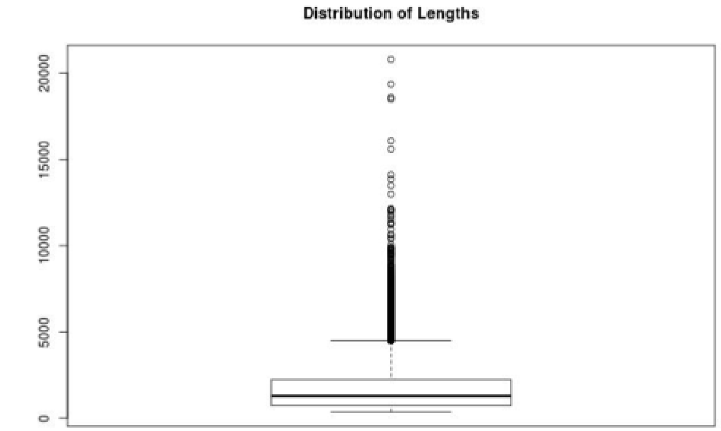


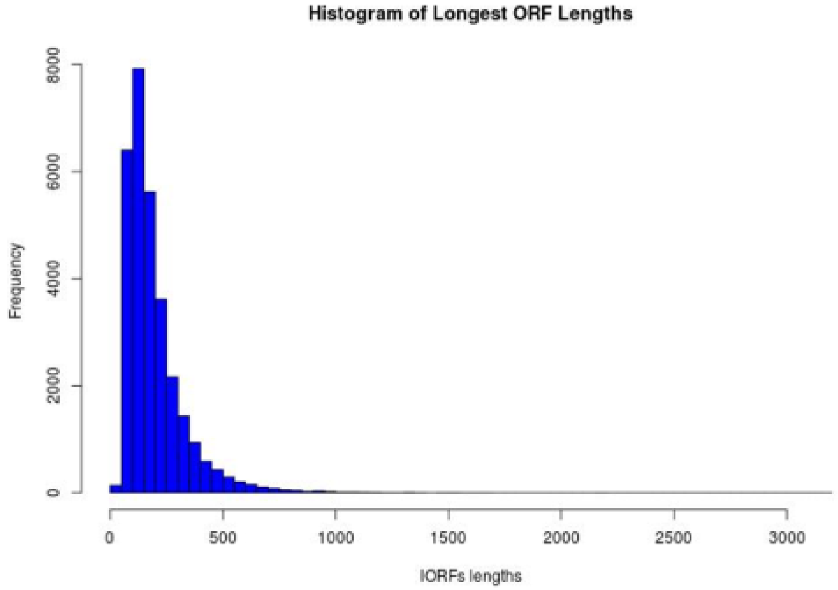
Results: graphical representation

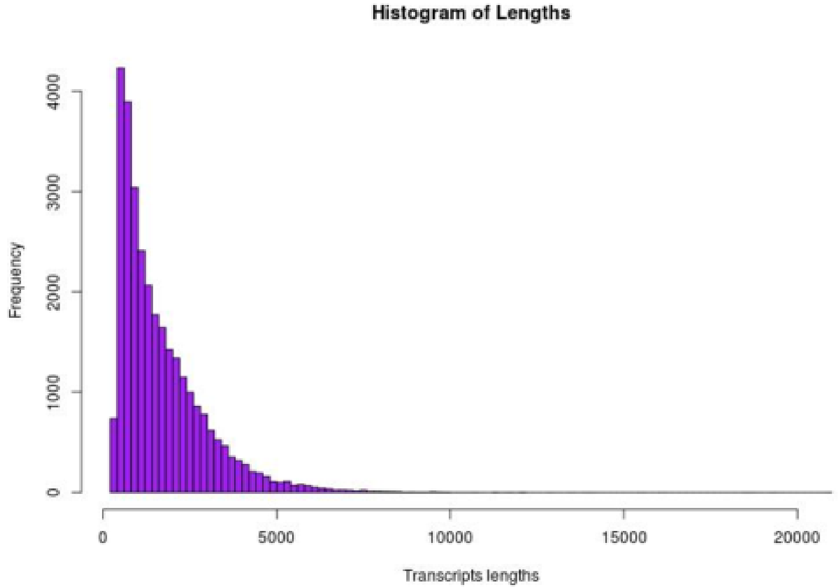
Annocript

Lengths and coverage


Distribution of Hit Coverages


Distribution of Lengths

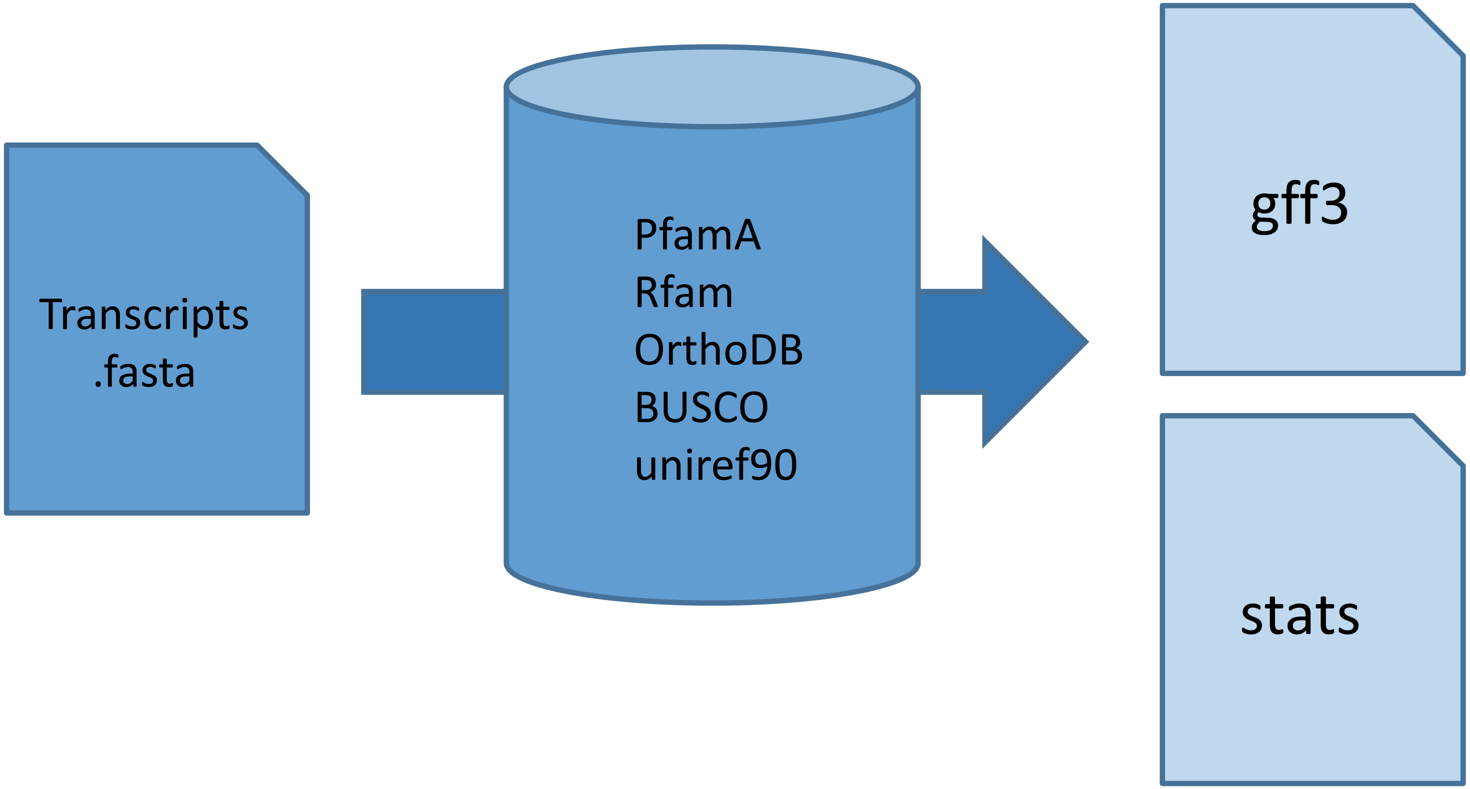

Histogram of Longest ORF Lengths


Histogram of Lengths

Dammit

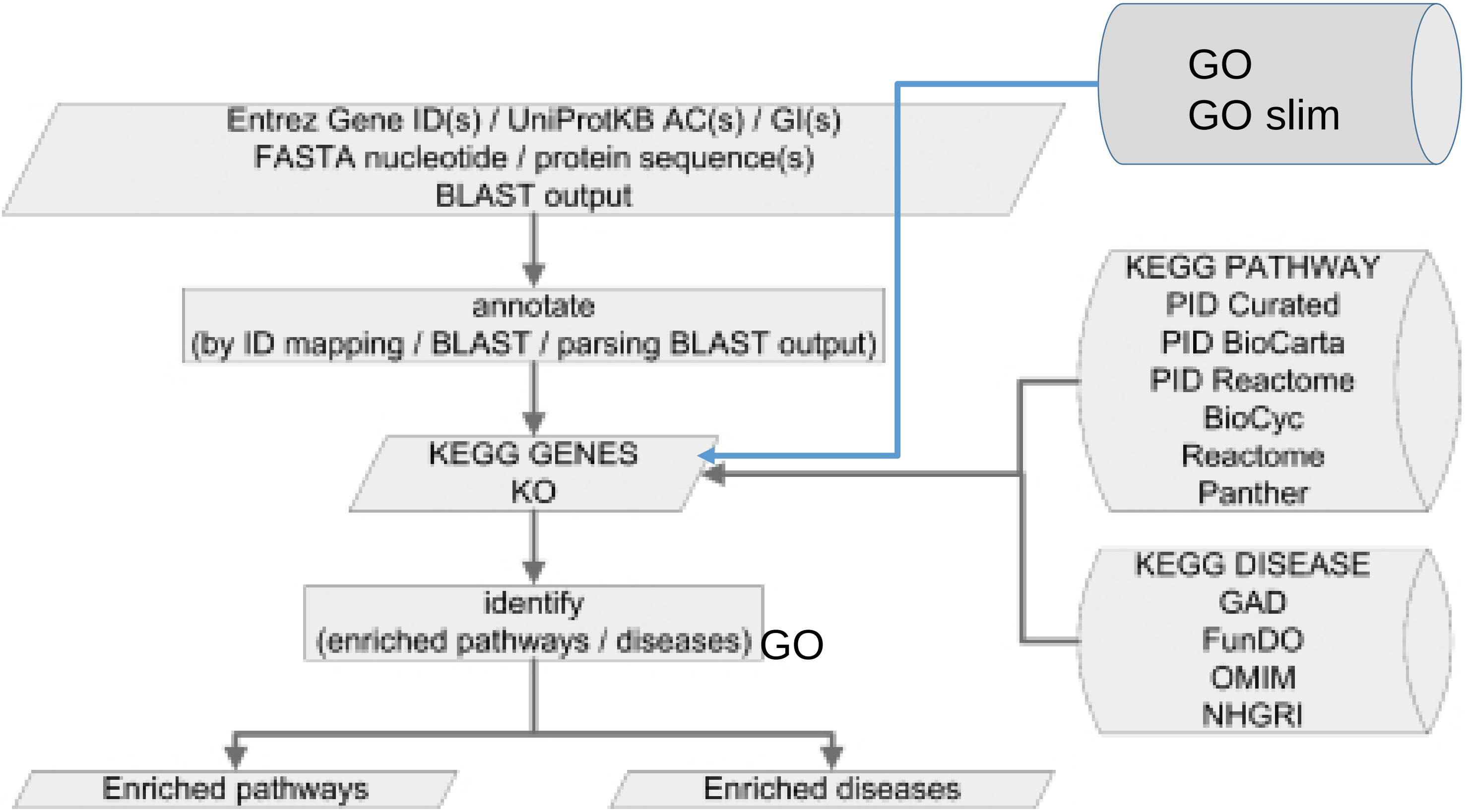
http://www.camillescott.org/dammit/

The *annotate* command runs the BUSCO assessment, assembly stats, and homology searches, aggregates the results, and outputs a GFF3 file and annotation report



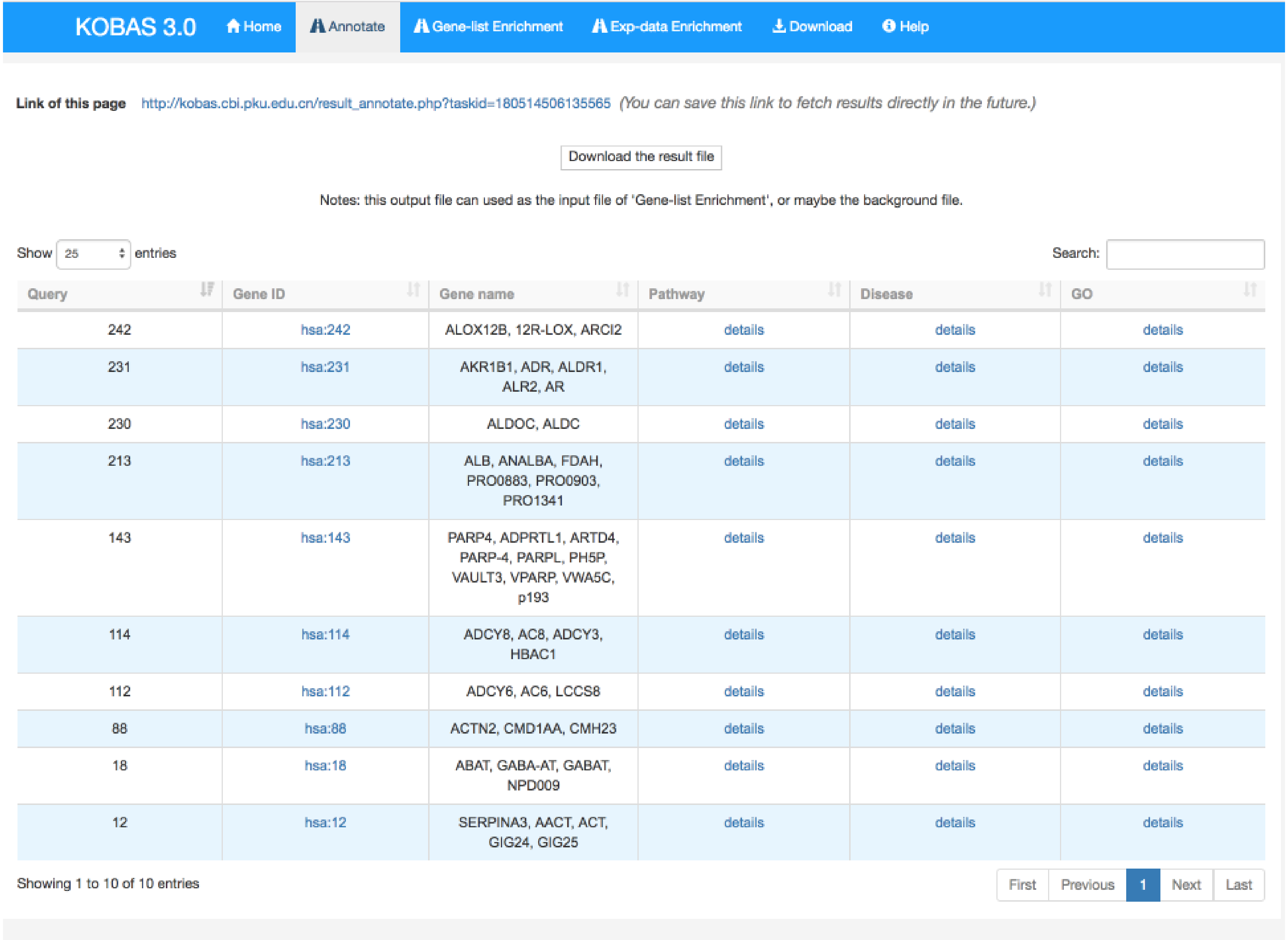
KOBAS : KO-Based Annotation System

KOBAS 3.0 : http://kobas.cbi.pku.edu.cn/



Nucleic Acids Res. 2011 Jul 1; 39(Web Server issue): W316–W322.
Published online 2011 Jun 27. doi: 10.1093/nar/gkr483

KOBAS : KO-Based Annotation System



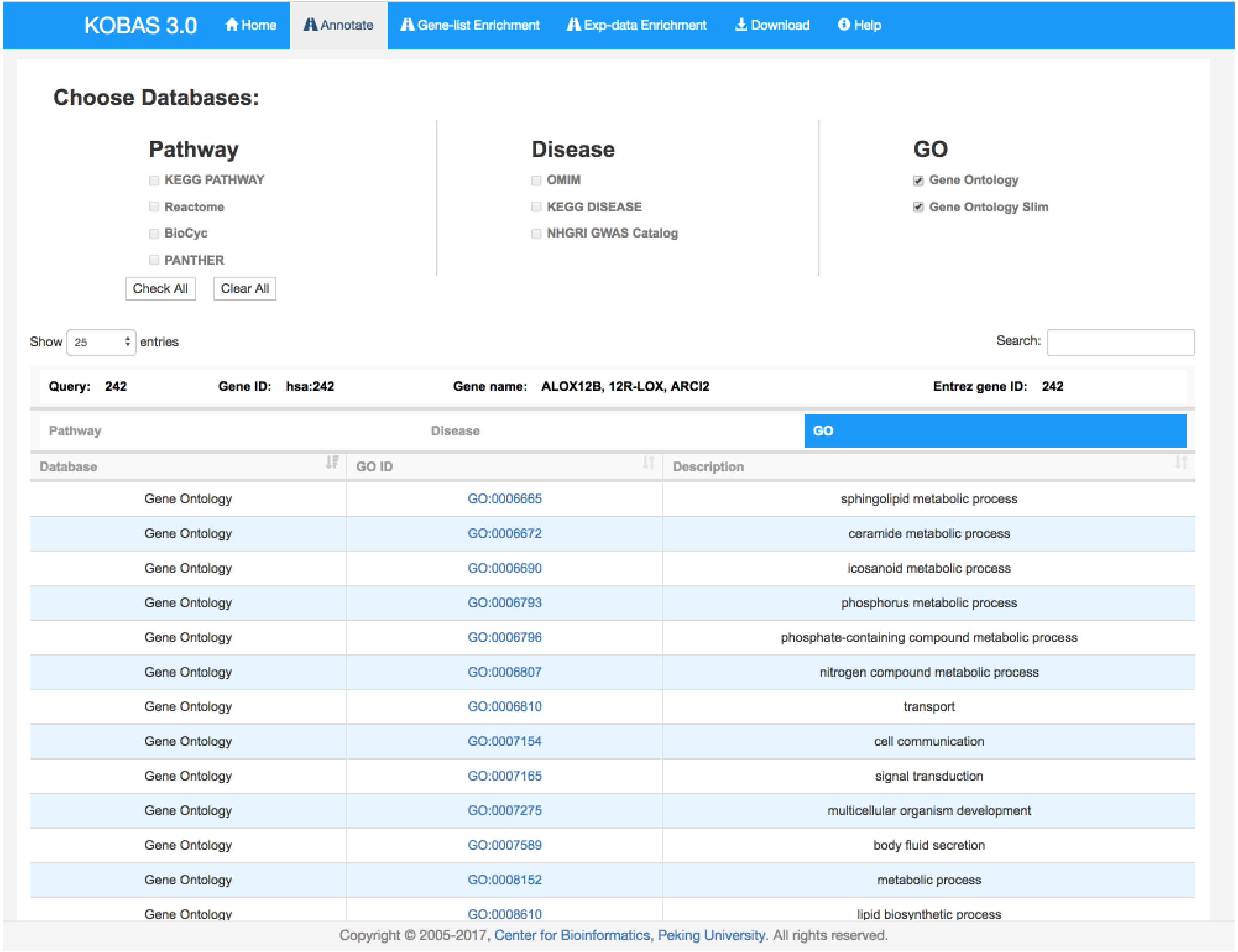
Link of this page http://kobas.cbi.pku.edu.cn/result_annotate.php?taskid=180514506135565 *(You can save this link to fetch results directly in the future.)*

Notes: this output file can used as the input file of 'Gene-list Enrichment', or maybe the background file.

Query	Gene ID	Gene name	Pathway	Disease	GO
242	hsa:242	ALOX12B, 12R-LOX, ARCI2	details	details	details
231	hsa:231	AKR1B1, ADR, ALDR1, ALR2, AR	details	details	details
230	hsa:230	ALDOC, ALDC	details	details	details
213	hsa:213	ALB, ANALBA, FDAH, PRO0883, PRO0903, PRO1341	details	details	details
143	hsa:143	PARP4, ADPRTL1, ARTD4, PARP-4, PARPL, PH5P, VAULT3, VPARP, VWA5C, p193	details	details	details
114	hsa:114	ADCY8, AC8, ADCY3, HBAC1	details	details	details
112	hsa:112	ADCY6, AC6, LCCS8	details	details	details
88	hsa:88	ACTN2, CMD1AA, CMH23	details	details	details
18	hsa:18	ABAT, GABA-AT, GABAT, NPD009	details	details	details
12	hsa:12	SERPINA3, AACT, ACT, GIG24, GIG25	details	details	details

Showing 1 to 10 of 10 entries

KOBAS : KO-Based Annotation System



Choose Databases:

Pathway

- □ KEGG PATHWAY
- □ Reactome
- □ BioCyc
- □ PANTHER

Check All | Clear All

Disease

- □ OMIM
- □ KEGG DISEASE
- □ NHGRI GWAS Catalog

GO

- ☑ Gene Ontology
- ☑ Gene Ontology Slim

Show 25 entries

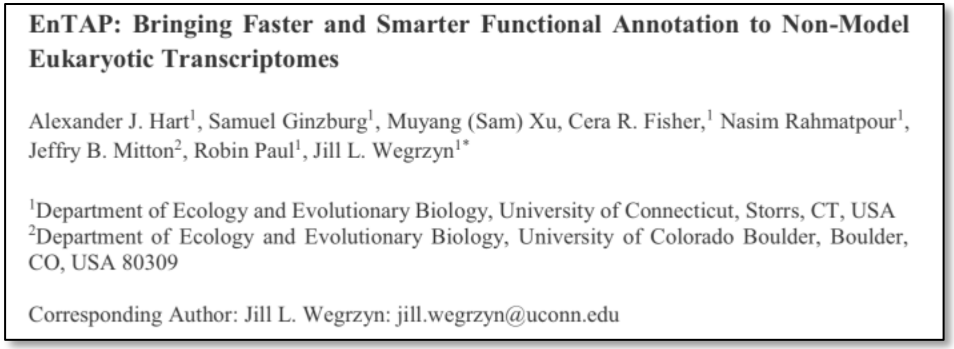
Search:

Query: 242 **Gene ID:** hsa:242 **Gene name:** ALOX12B, 12R-LOX, ARCI2 **Entrez gene ID:** 242

Database	GO ID	Description
Gene Ontology	GO:0006665	sphingolipid metabolic process
Gene Ontology	GO:0006672	ceramide metabolic process
Gene Ontology	GO:0006690	icosanoid metabolic process
Gene Ontology	GO:0006793	phosphorus metabolic process
Gene Ontology	GO:0006796	phosphate-containing compound metabolic process
Gene Ontology	GO:0006807	nitrogen compound metabolic process
Gene Ontology	GO:0006810	transport
Gene Ontology	GO:0007154	cell communication
Gene Ontology	GO:0007165	signal transduction
Gene Ontology	GO:0007275	multicellular organism development
Gene Ontology	GO:0007589	body fluid secretion
Gene Ontology	GO:0008152	metabolic process
Gene Ontology	GO:0008610	lipid biosynthetic process

Copyright © 2005-2017, Center for Bioinformatics, Peking University. All rights reserved.

EnTAP (Eukaryotic Non-Model Transcriptome Annotation Pipeline)



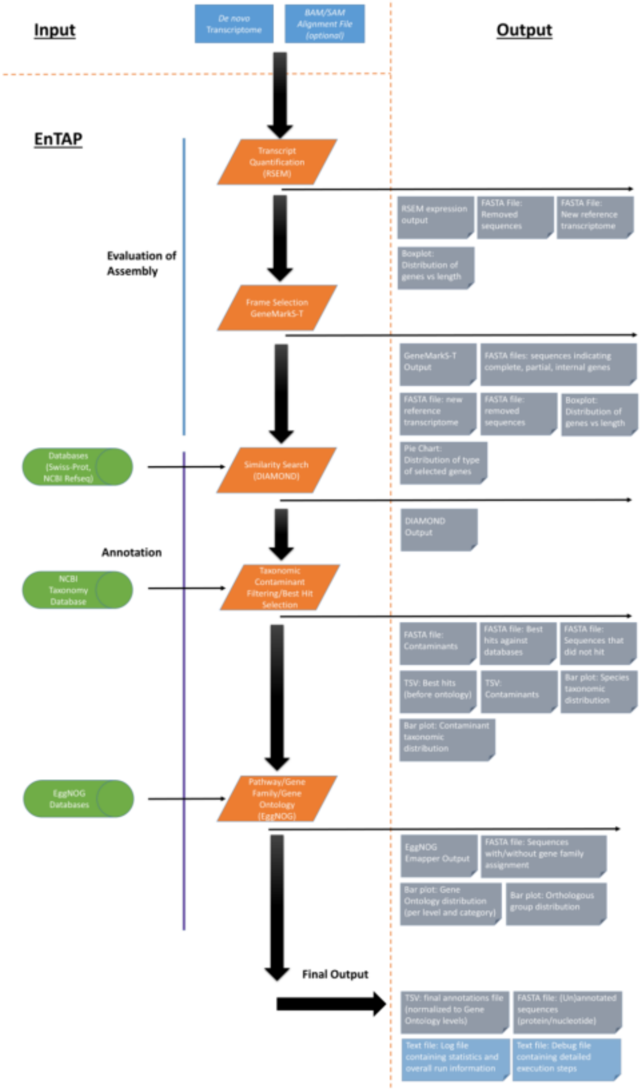
EnTAP: Bringing Faster and Smarter Functional Annotation to Non-Model Eukaryotic Transcriptomes

Alexander J. Hart[1], Samuel Ginzburg[1], Muyang (Sam) Xu, Cera R. Fisher,[1] Nasim Rahmatpour[1], Jeffry B. Mitton[2], Robin Paul[1], Jill L. Wegrzyn[1*]

[1]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA
[2]Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO, USA 80309

Corresponding Author: Jill L. Wegrzyn: jill.wegrzyn@uconn.edu

Hart et *al*. 2018 bioRxiv : http://dx.doi.org/10.1101/307868



Input

De novo Transcriptome

BAM/SAM Alignment File (optional)

Output

EnTAP

Transcript Quantification (RSEM)

RSEM expression output

FASTA File: Removed sequences

FASTA File: New reference transcriptome

Boxplot: Distribution of genes vs length

Evaluation of Assembly

Frame Selection GeneMarkS-T

GeneMarkS-T Output

FASTA files: sequences indicating complete, partial, internal genes

FASTA file: new reference transcriptome

FASTA file: removed sequences

Boxplot: Distribution of genes vs length

Pie Chart: Distribution of type of selected genes

Databases (Swiss-Prot, NCBI Refseq)

Similarity Search (DIAMOND)

DIAMOND Output

Annotation

NCBI Taxonomy Database

Taxonomic Contaminant Filtering/Best Hit Selection

FASTA file: Contaminants

FASTA file: Best hits against databases

FASTA file: Sequences that did not hit

TSV: Best hits (before ontology)

TSV: Contaminants

Bar plot: Species taxonomic distribution

Bar plot: Contaminant taxonomic distribution

EggNOG Databases

Pathway/Gene Family/Gene Ontology (EggNOG)

EggNOG Emapper Output

FASTA file: Sequences with/without gene family assignment

Bar plot: Gene Ontology distribution (per level and category)

Bar plot: Orthologous group distribution

Final Output

TSV: final annotations file (normalized to Gene Ontology levels)

FASTA file: (Un)annotated sequences (protein/nucleotide)

Text file: Log file containing statistics and overall run information

Text file: Debug file containing detailed execution steps

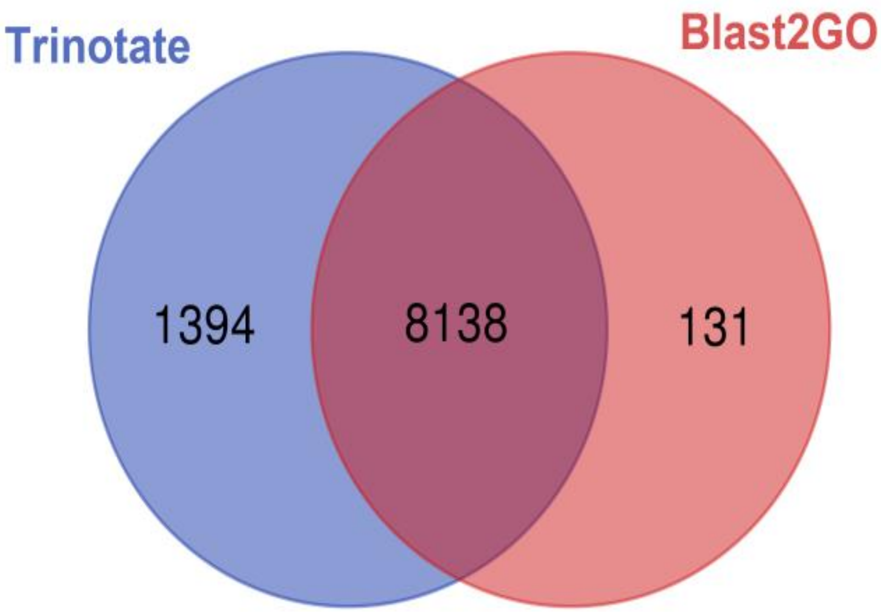
Transcriptome filtering :
RSEM

Transcriptome annotation
GeneMarkS-T (more complete genes than Transdecoder)

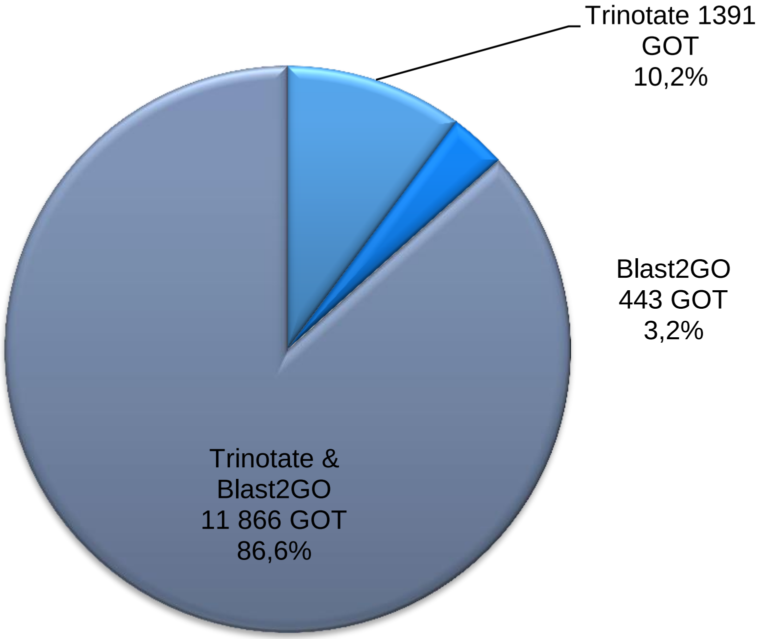
DIAMOND (Fast and Sensitive NCBI BLAST Alternative)
Combination of curated databases (at least 3)
Selection of Optimal Hit From Several Databases

Selection of Optimal Hit Based on Informativeness
Contaminant Identification and Filtering

Trinotate vs Blast2GO : Go Terms



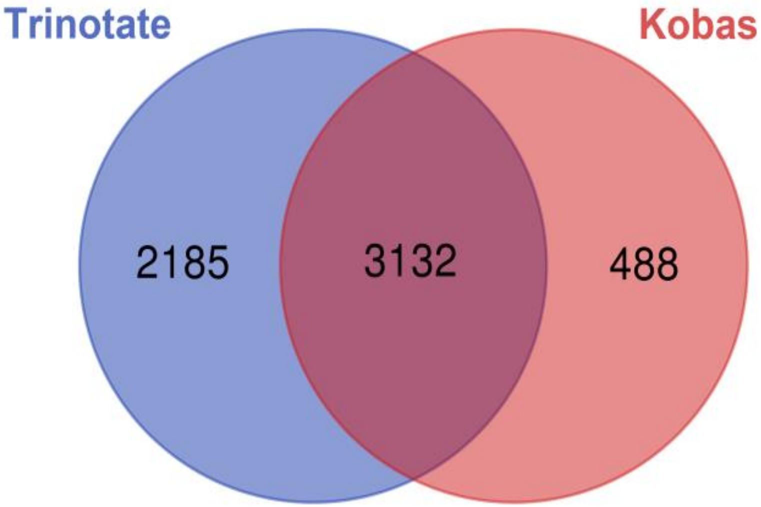
Number of sequence annotated with GO terms

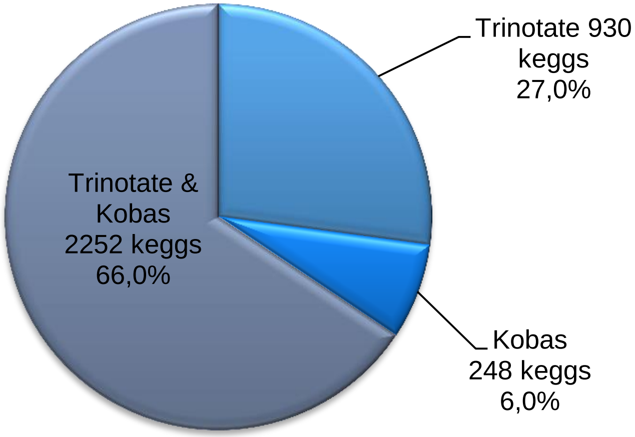


Number of GO terms

Saccharina japonica genome

Trinotate vs Kobas : Kegg

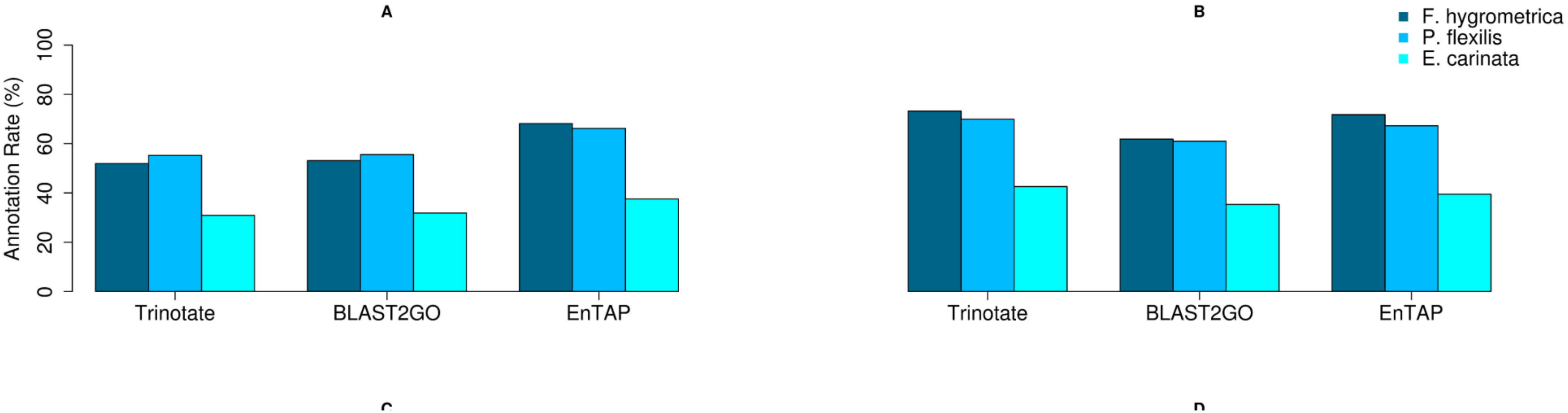




Number of sequence annotated with KEGG terms

Number of KEGG terms

Saccharina japonica genome



Overall Annotation Rate – UniProt Swiss-Prot (A) and NCBI RefSeq Complete (B)

Hart et *al*. 2018 bioRxiv : http://dx.doi.org/10.1101/307868.

Merci (s)

Membres du groupe PEPI annot : https://pepi-ibis.inra.fr/annotation-genomes

Xi Liu
Arnaud Meng
Guita Niang
Delphine Negre
Ehsan Kayal
Gildas Le Corguillé
Annie Le Breton
Eric Pelletier
Maxim Scheremetjew
Rob Finn

Jean Yves Toullec
Mark Cock
Jonas Colleen
Simon Dittami
Fabrice Not
Laur Guillou
Daniel Vaulot
Joel Henry
Celine Gaudin
Jean Yves Sire
Flavia Nunes
…