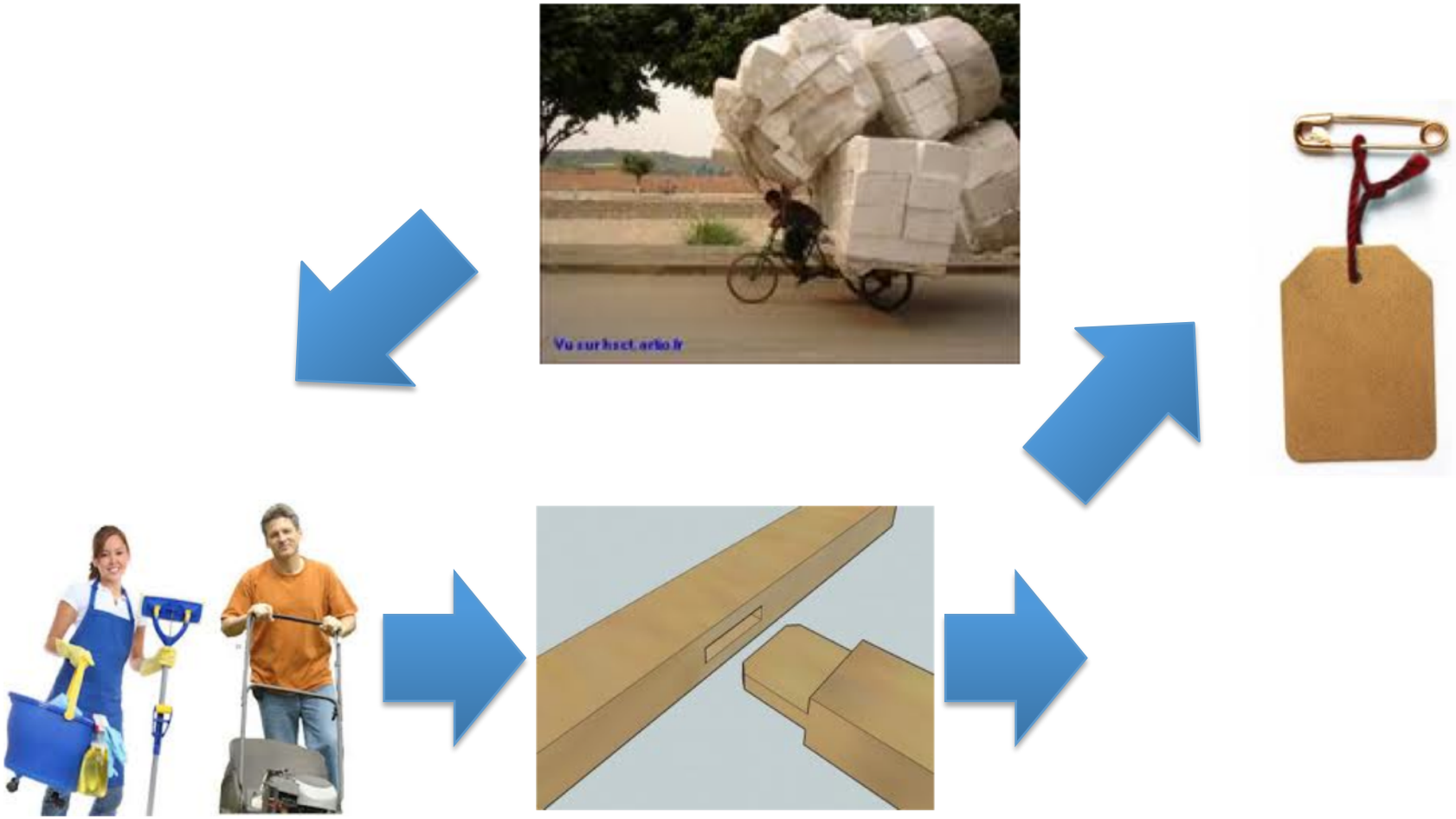
26/09/2019

RNA Seq analysis

Transcriptome annotation

ABiMS – Station Biologique Roscoff
South Green

RNA Seq analysis



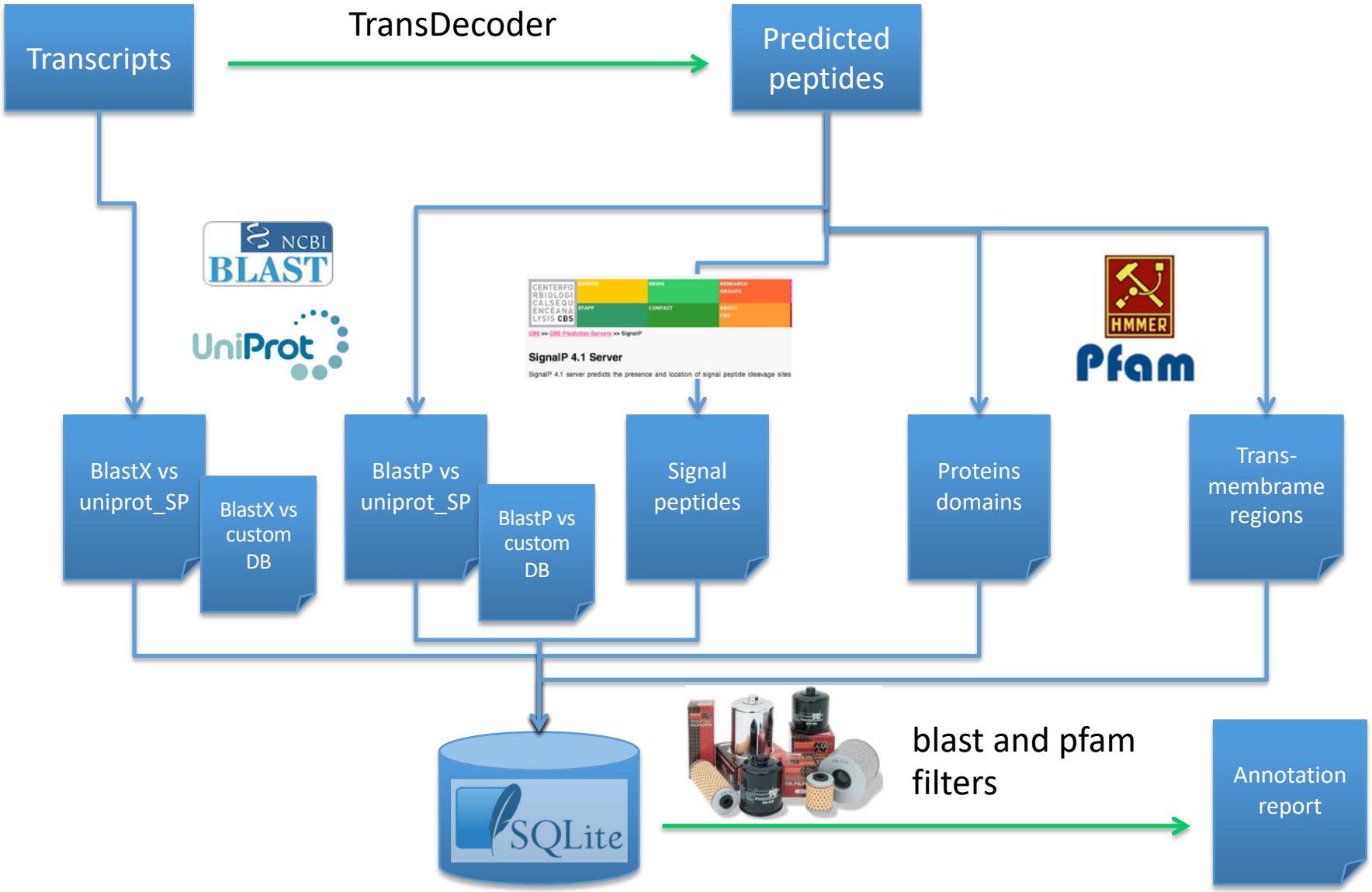
Transcriptome annotation

Pipeline	websites
Trinotate	David
Exemple Camera pipeline	Trapid
blast2Go	transcriptator
annoscript	
Damnit!	

Trinotate pipeline

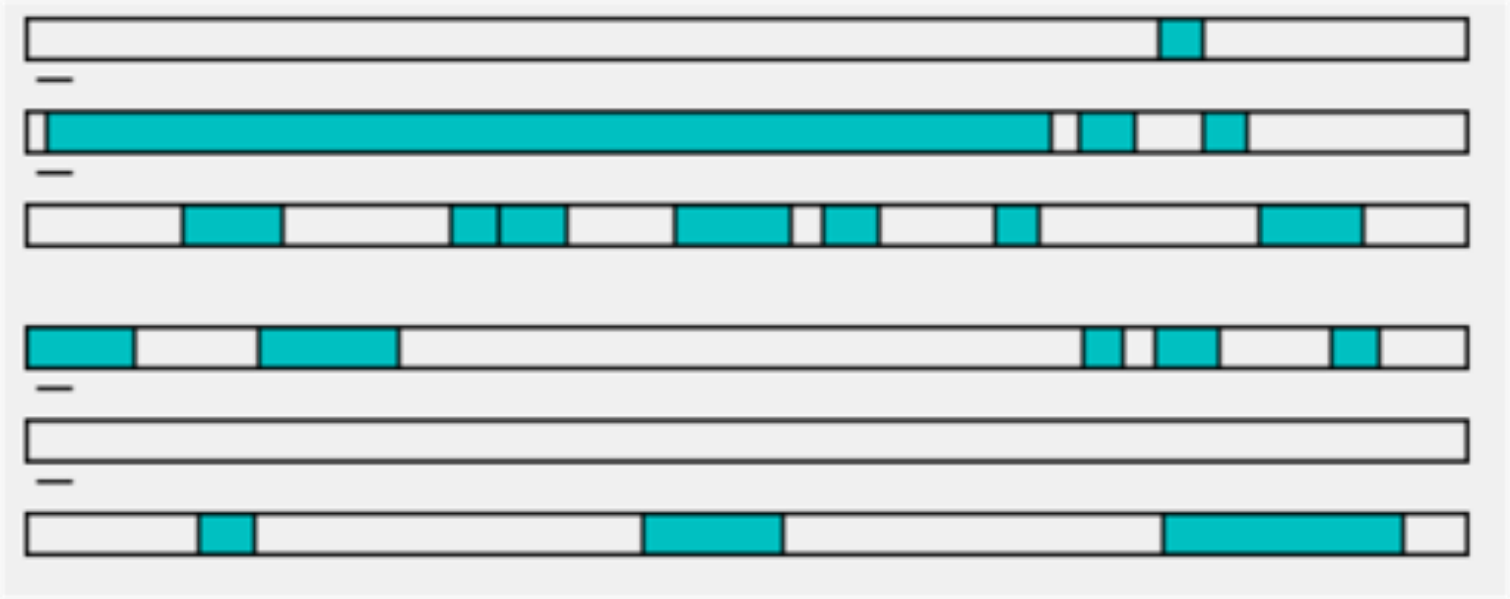
CNRS UPMC
Station Biologique
Roscoff
Abims

Trinotate pipeline



Trinotate pipeline

1. Find Likely Coding Regions(using TransDecoder)



- Find all 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.

Retrieving the databases

- 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

Retrieving the databases

```
TRINOTATE_HOME/admin/Build_Trinotate_Boilerplate_SQLite_db.pl Trinotate
```

- it will provide to you:
  - `Trinotate.sqlite`
  - `uniprot_sprot.pep`
  - `Pfam-A.hmm.gz`

- Prepare the protein database for blast searches :

```
makeblastdb -in uniprot_sprot.pep -dbtype prot
```

- Uncompress and prepare the Pfam database for use with 'hmmscan' like so:

```
gunzip Pfam-A.hmm.gz
hmmpress Pfam-A.hmm
```

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

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 ▼ Next Match ▲ Previous Match

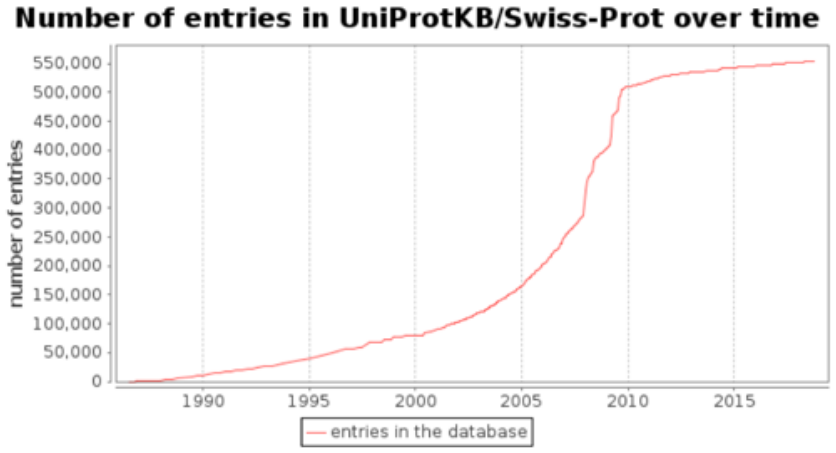
Score	Expect	Method	Identities	Positives	Gaps
99.9 bits(228)	4e-20	Compositional matrix adjust.	87/321(27%)	114/321(35%)	41/321(12%)

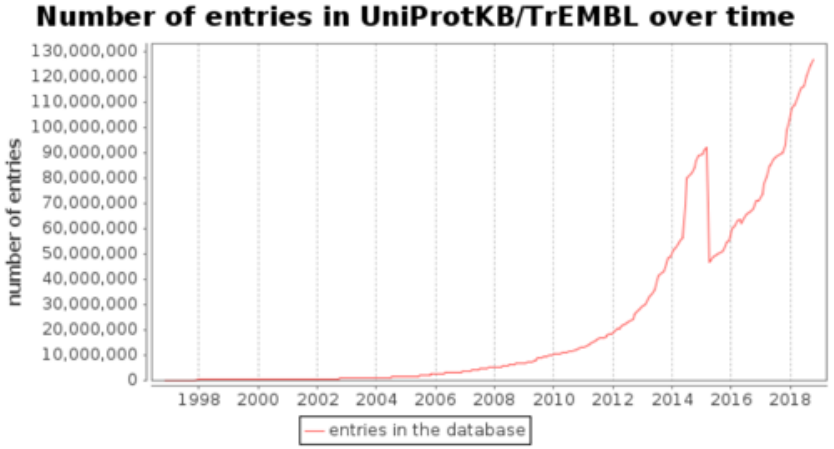
```
Query  8    SNVVGVHYRVGKKIGEGSFGMLFQGVNL--------INNQP-------IALKFESRKSEV  52
            +    + R+G  IG G FG  +               +  +P       + + F  R
Sbjct  40   TDLAKGQWRIGPSIGVGGFGEIYAACKVGEKNYDAVVKCEPHGNGPLFVEMHFYLRNAKL  99

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.





CNRS UPMC Station Biologique Roscoff

Abims

NCBI BLAST

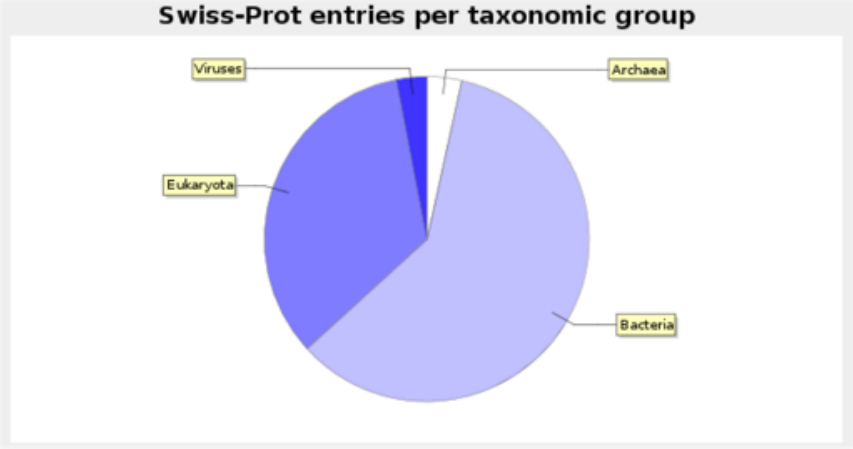
vs

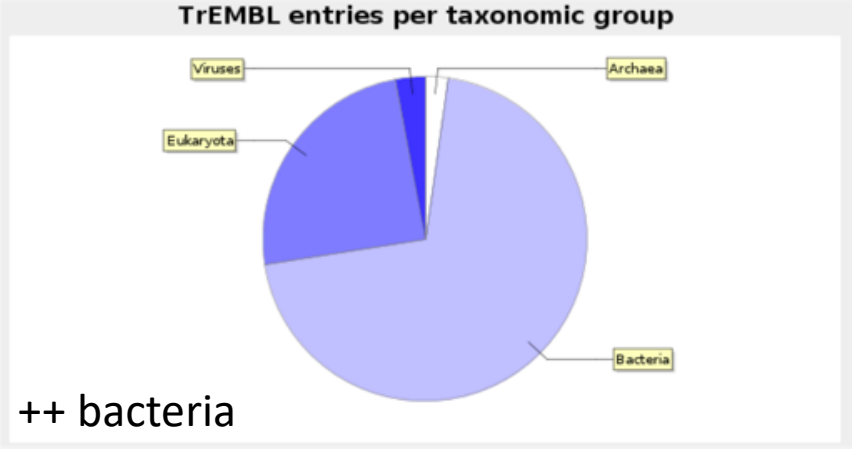
UniProt

swissprot SIB EBI

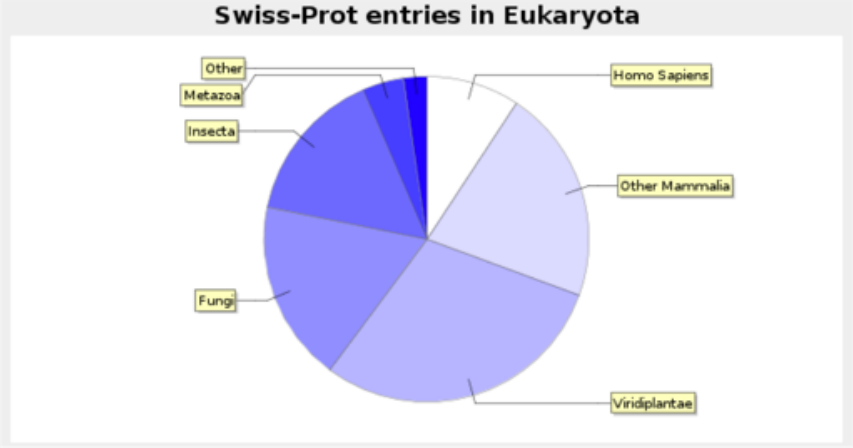
558 590 sequences

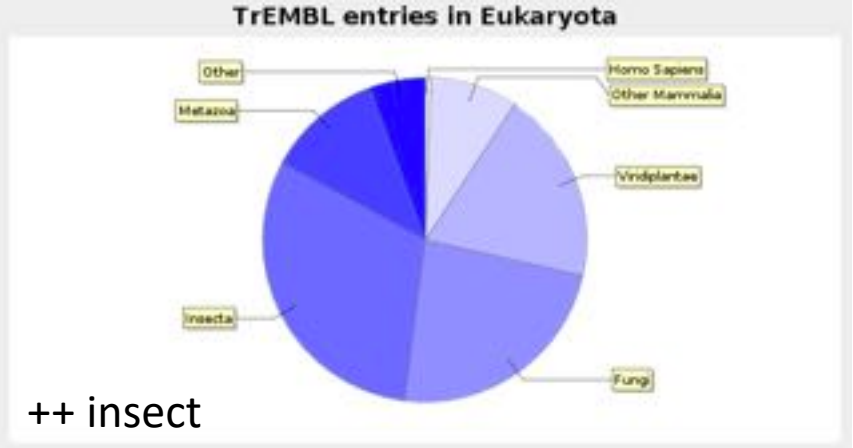
TrEMBL : 126 780 198 sequences





++ bacteria





++ 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

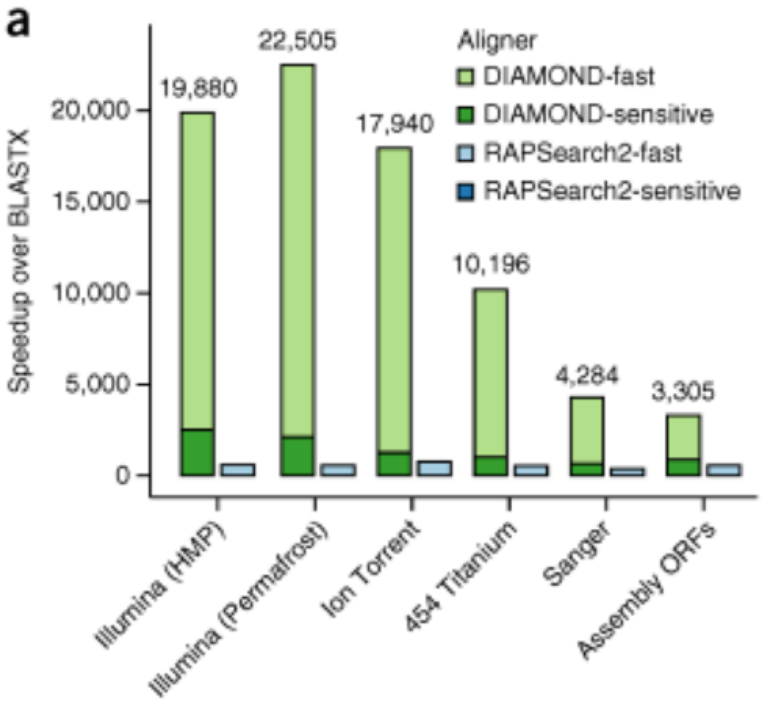
Specific databases …

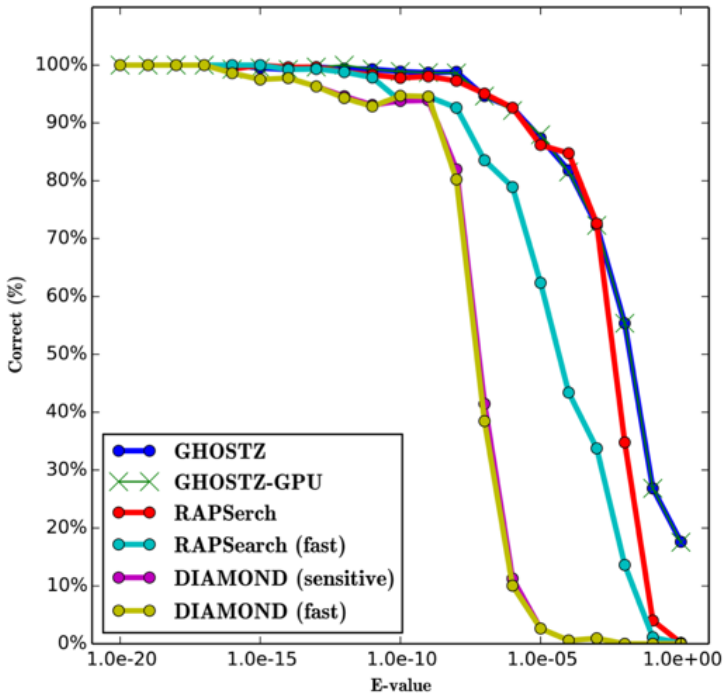
…

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

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 **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

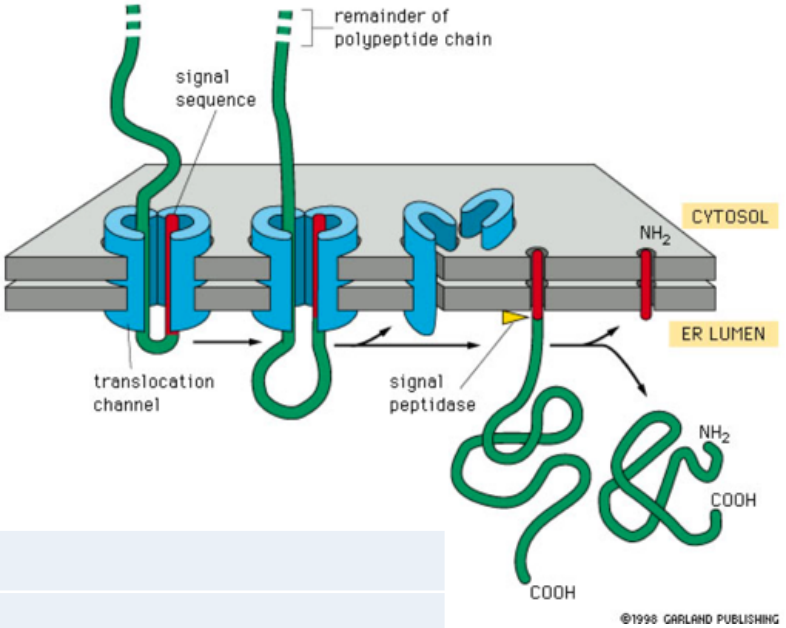
4. Running signalP to predict signal peptides

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

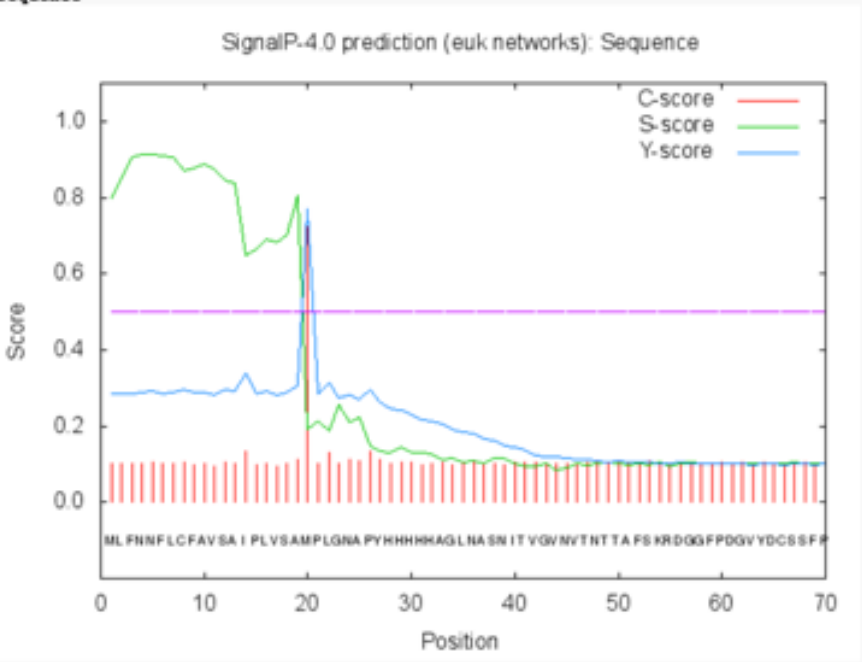


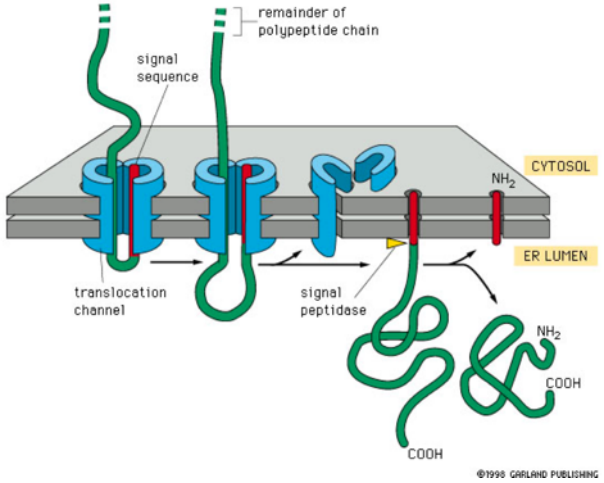
Typical Signal Peptides	
peptide function	Composition
Transport in cellular nucleus (NLS)	-Pro-Pro-Lys-Lys-Lys-Arg-Lys-Val-
Endoplasmic reticulum transport	H_2N-Met-Met-Ser-Phe-Val-Ser-Leu-Leu-Leu-Val-Gly-Ile-Leu-Phe-Trp-Ala-Thr-Glu-Ala-Glu-Gln-Leu-Thr-Lys-Cys-Glu-Val-Phe-Gln-
Endoplasmic reticulum retention	-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/

SignalP

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





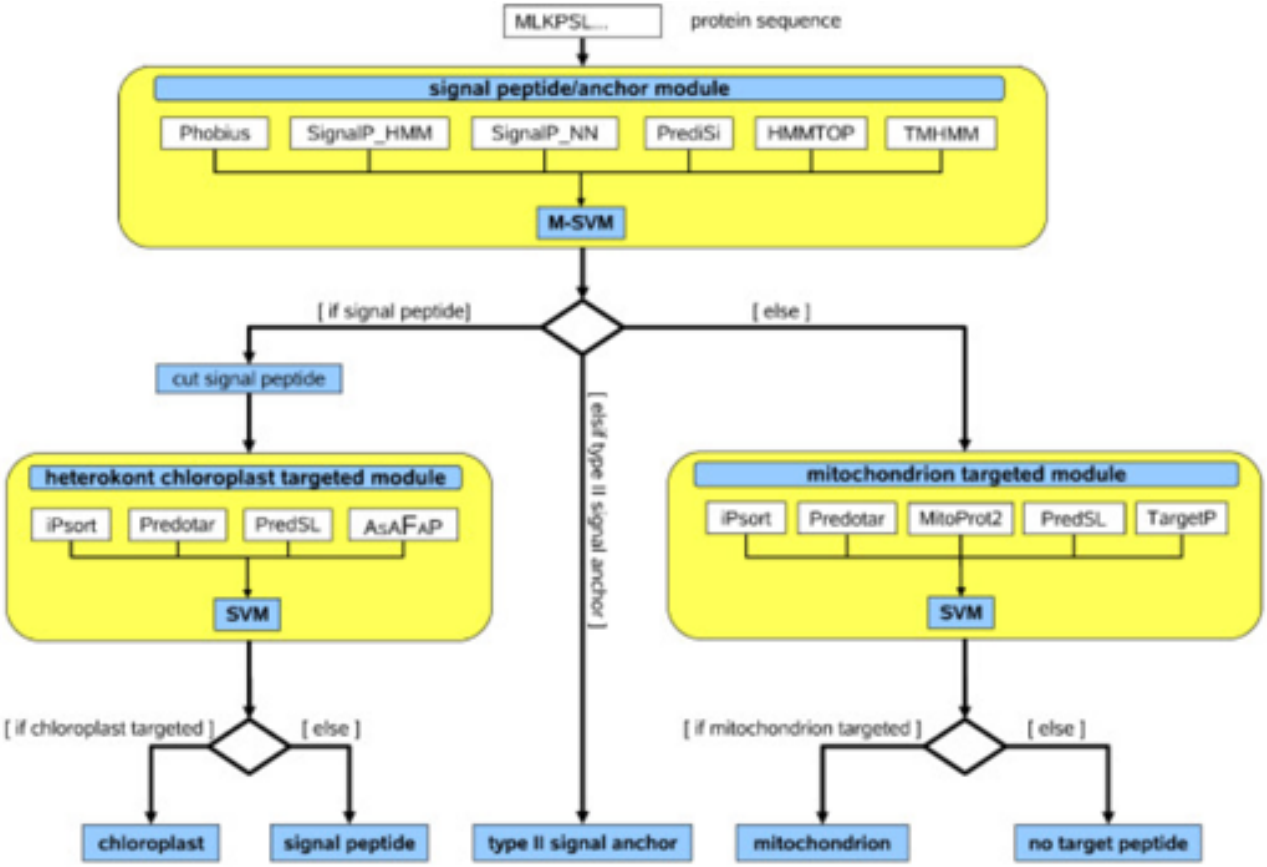
http://www.cbs.dtu.dk/services/SignalP/

```
# 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
```

```
##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

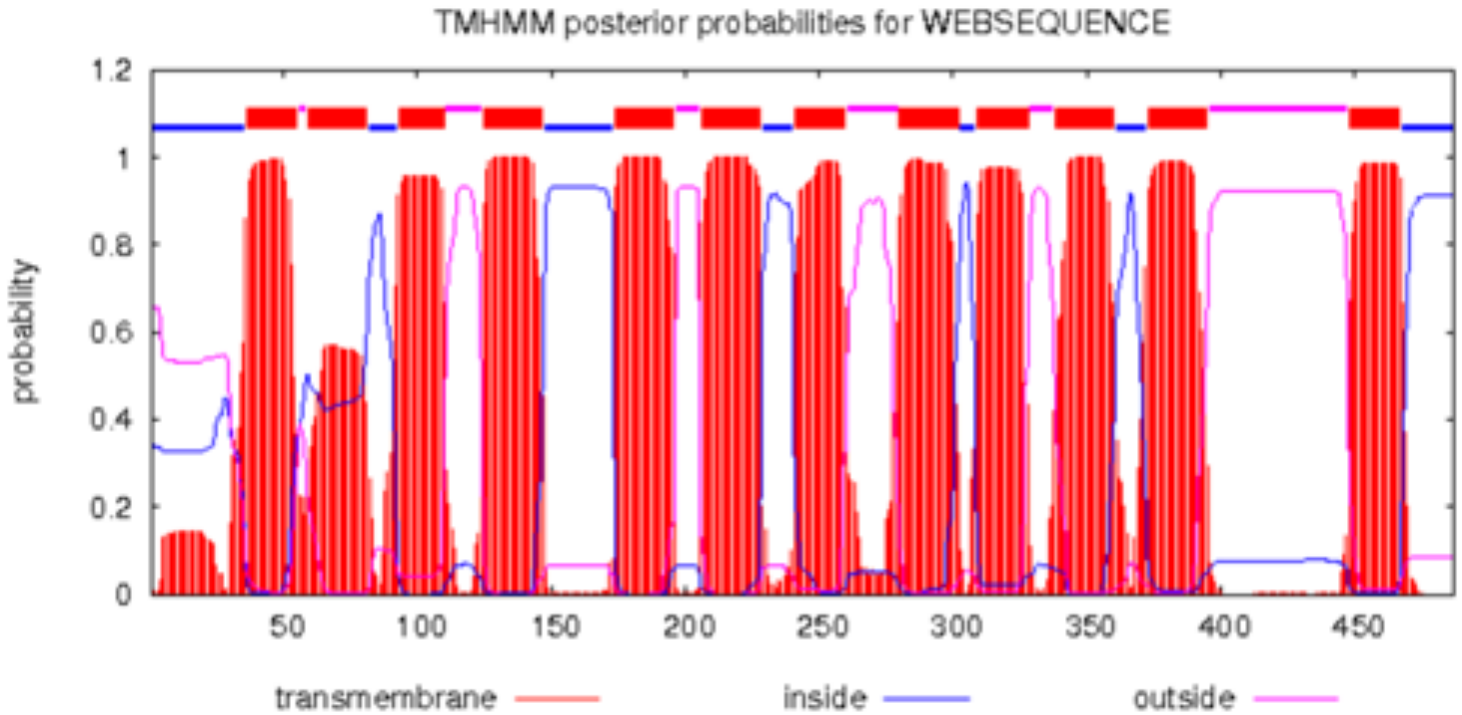
3. Running HMMER to identify protein domains

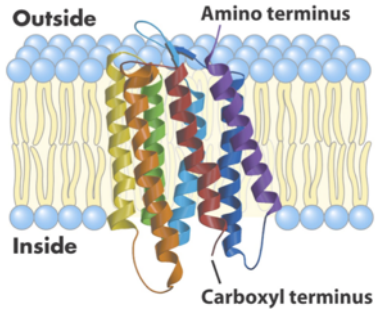
4. Running signalP to predict signal peptides

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

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_pombe_pep_BLASTX
8 custom_pombe_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

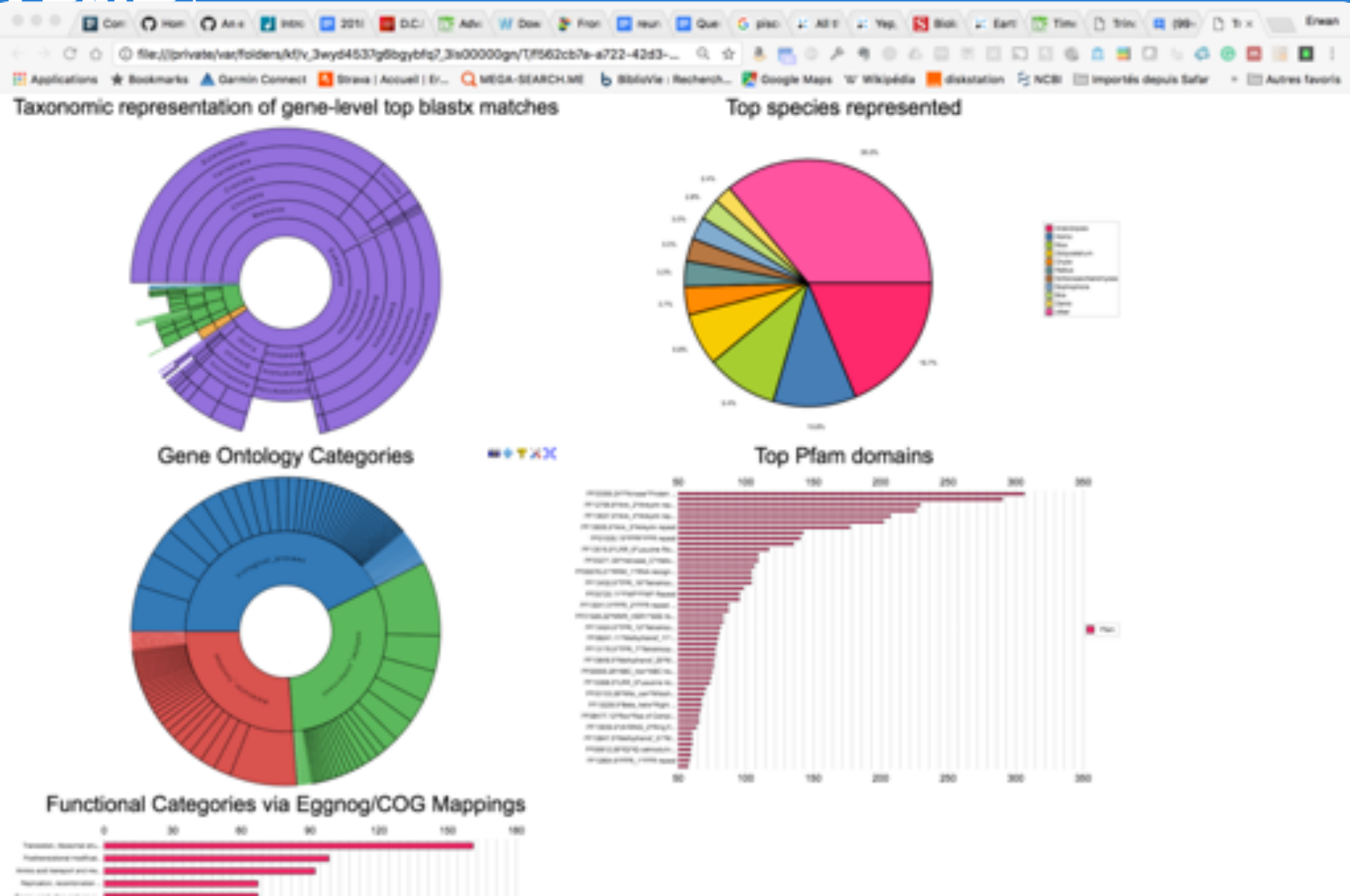
2 sprot_Top_BLASTX_hit GCS1_SCHPO^GCS1_SCHPO^Q:53-2476,H:1-808^100%ID^E:0^RecName: Full=Probable mannosyl-oligosaccharide glucosidase;^Eukaryota;
Fungi; Dikarya; Ascomycota; Taphrinomycotina; Schizosaccharomycetes; Schizosaccharomycetales; Schizosaccharomycetaceae; Schizosaccharomyces

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

CNRS UPMC
Station Biologique
Roscoff

New : trinotate_report_summary.pl



Automated Execution of Trinotate

TRINOTATE_HOME/auto/autoTrinotate.pl

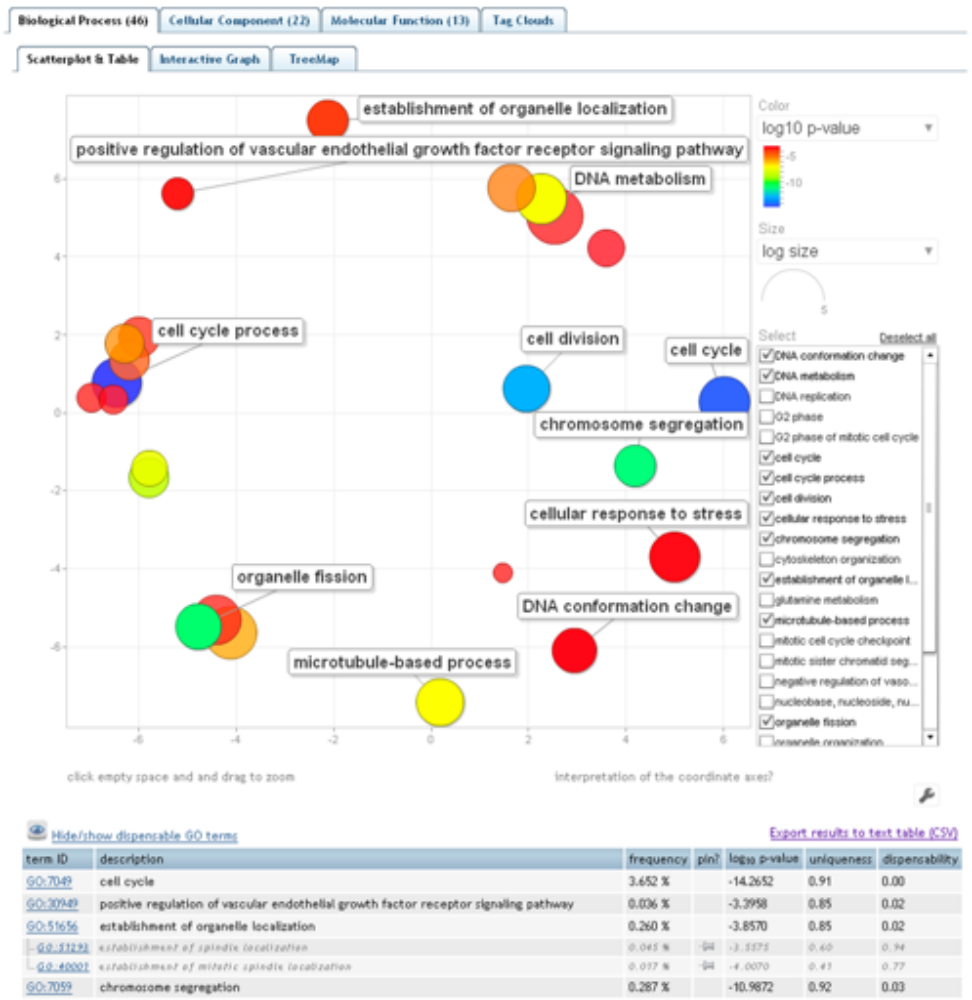
```
#########################################################################################
# Required:
#
#--Trinotate_sqlite <string> Trinotate.sqlite boilerplate database
#
#--transcripts <string> transcripts.fasta
#
#--gene_to_trans_map <string> gene-to-transcript mapping file
#
#--conf <string>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

REVIGO

reduce + visualize Gene Ontology

http://revigo.irb.hr/

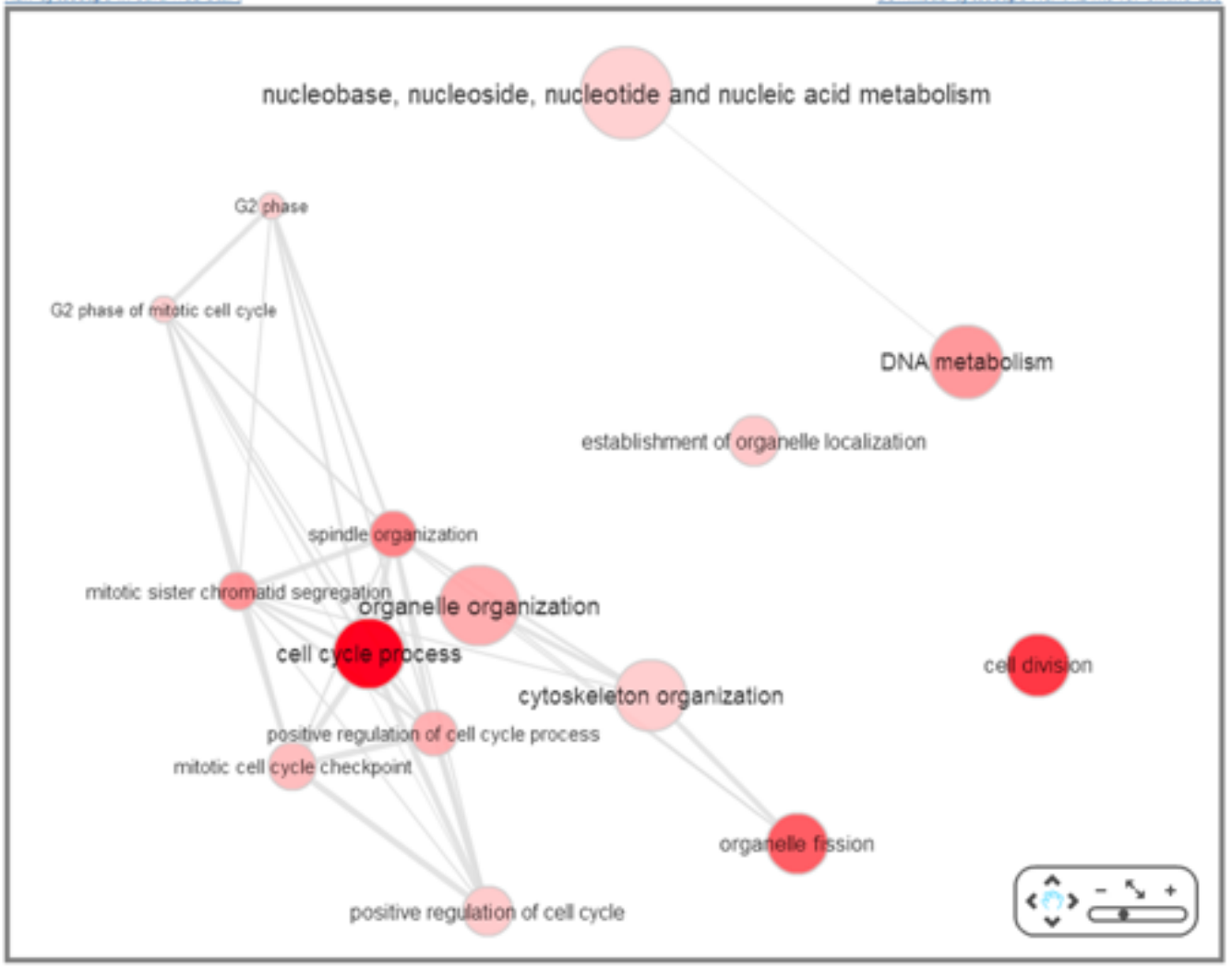
% GeneGroup	pValue
GO:0009268	1e-14
GO:0010447	1e-14
GO:0000027	1e-297
GO:0042255	1e-297
GO:0042257	1e-297
GO:0042273	1e-297
GO:0030880	1e-17
GO:0009775	1e-13
GO:0009853	1e-11
GO:0030255	1e-18
GO:0015797	1e-11
GO:0045158	1e-27
GO:0000786	1e-31
GO:0006334	1e-31
GO:0034728	1e-31
GO:0009539	1e-12



Scatterplot & Table | Interactive Graph | TreeMap

Run Cytoscape in Java web start

Download Cytoscape XGMML file for offline use



WEGO 2.0
Web Gene Ontology Annotation Plotting

http://wego.genomics.org.cn/

Condition 1

_c0_g1	GO:0000041,GO:0000287,GO:0003674,GO:0005215,GO:0005488,GO:0005575,GO:0005887,GO:0006461,GO:0
_c1_g1	GO:0003674,GO:0003824,GO:0005975,GO:0006464,GO:0006486,GO:0006493,GO:0007275,GO:0008150,GO:0
TRINITY_DN10197_c0_g1	GO:0003674,GO:0005488,GO:0005515
TRINITY_DN10203_c0_g1	GO:0003674,GO:0003824,GO:0004064,GO:0005575,GO:0005783,GO:0008150,GO:0008152,GO:0009058,GO:0

Condition 2

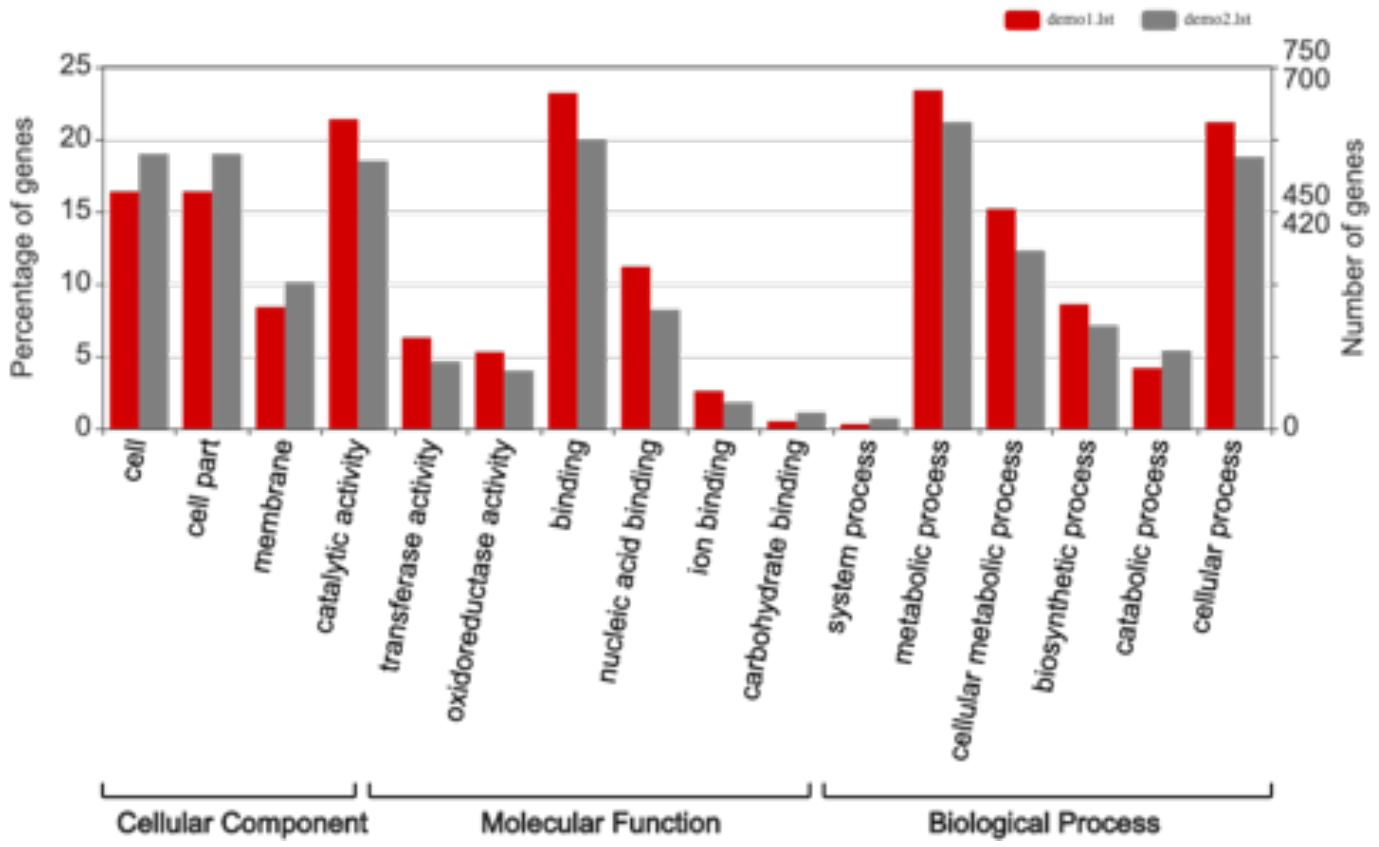
_c0_g1	GO:0000041,GO:0000287,GO:0003674,GO:0005215,GO:0005488,GO:0005575,GO:0005887,GO:0006461,GO
_c1_g1	GO:0003674,GO:0003824,GO:0005975,GO:0006464,GO:0006486,GO:0006493,GO:0007275,GO:0008150,GO
TRINITY_DN10197_c0_g1	GO:0003674,GO:0005488,GO:0005515
TRINITY_DN10203_c0_g1	GO:0003674,GO:0003824,GO:0004064,GO:0005575,GO:0005783,GO:0008150,GO:0008152,GO:0009058,GO
TRINITY_DN10208_c0_g1	GO:0003674,GO:0005488,GO:0005515
TRINITY_DN10212_c0_g1	GO:0000166,GO:0001882,GO:0001883,GO:0003674,GO:0003824,GO:0004672,GO:0004674,GO:0004871,GO
TRINITY_DN10236_c0_g1	GO:0000070,GO:0000819,GO:0003674,GO:0005488,GO:0005515,GO:0005575,GO:0005634,GO:0005815,GO
TRINITY_DN10236_c0_g2	GO:0003674,GO:0005488,GO:0005515,GO:0005575,GO:0005634,GO:0006464,GO:0007049,GO:0008150,GO
TRINITY_DN10258_c0_g1	GO:0003674,GO:0003676,GO:0003677,GO:0003824,GO:0003887,GO:0004518,GO:0004527,GO:0005488,GO
TRINITY_DN10275_c0_g1	GO:0000166,GO:0001882,GO:0001883,GO:0003674,GO:0003676,GO:0003723,GO:0003743,GO:0003824,GO
TRINITY_DN10294_c0_g1	GO:0003674,GO:0003824,GO:0008080,GO:0016407,GO:0016410,GO:0016740,GO:0016746,GO:0016747
TRINITY_DN10298_c0_g1	8,GO
TRINITY_DN10300_c0_g1	1,GO
TRINITY_DN10301_c0_g1	
TRINITY_DN10320_c0_g1	7,GO
TRINITY_DN10328_c0_g2	9,GO
TRINITY_DN10329_c0_g1	4
TRINITY_DN10337_c0_g1	

Summary

		Condition 1	Condition 2	Total
Gene		3000	2800	5800
Annotated Genes		1679	1551	3230
GO Terms	Biological	1164	1042	2206
	Cellular	532	571	1103
	Function	1462	1305	2767
	Total	3158	2918	6076

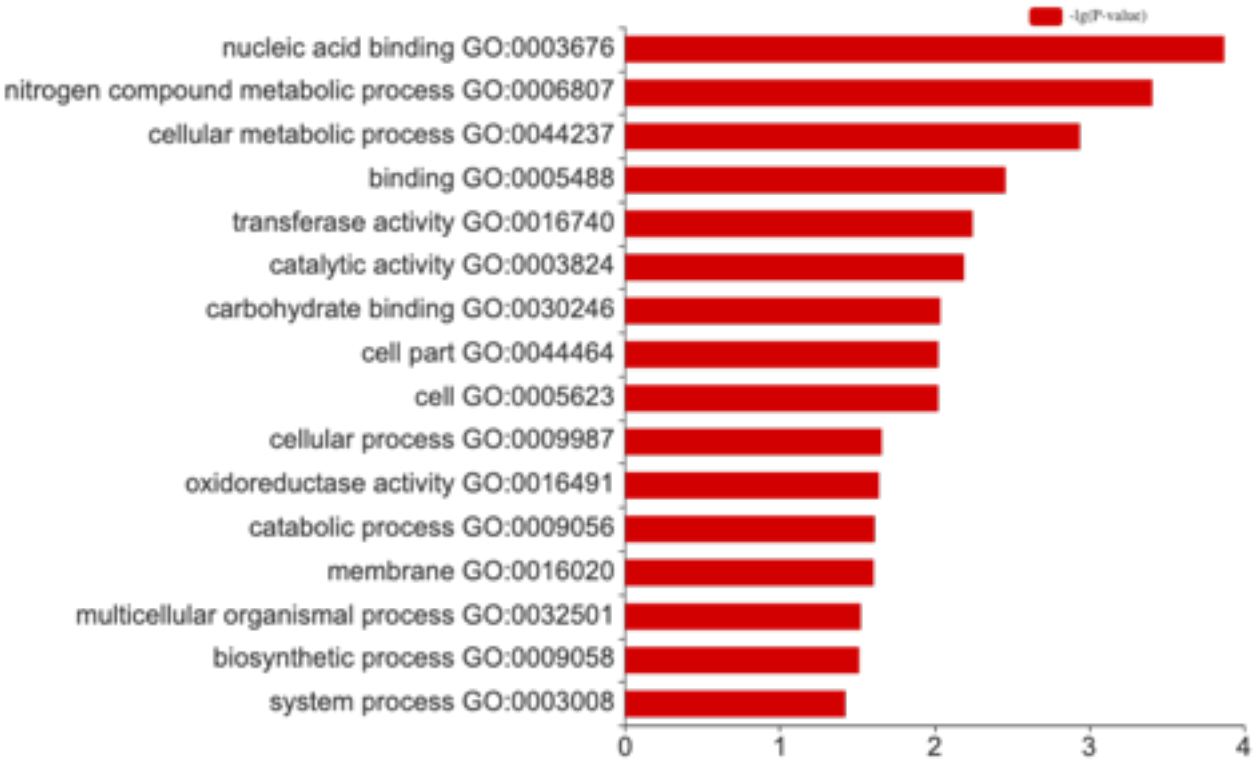
WEGO 2.0
Web Gene Ontology Annotation Plotting

http://wego.genomics.org.cn/



WEGO 2.0
Web Gene Ontology Annotation Plotting

http://wego.genomics.org.cn/



Scatterplot & Table | Interactive Graph | TreeMap

Trinotate web

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

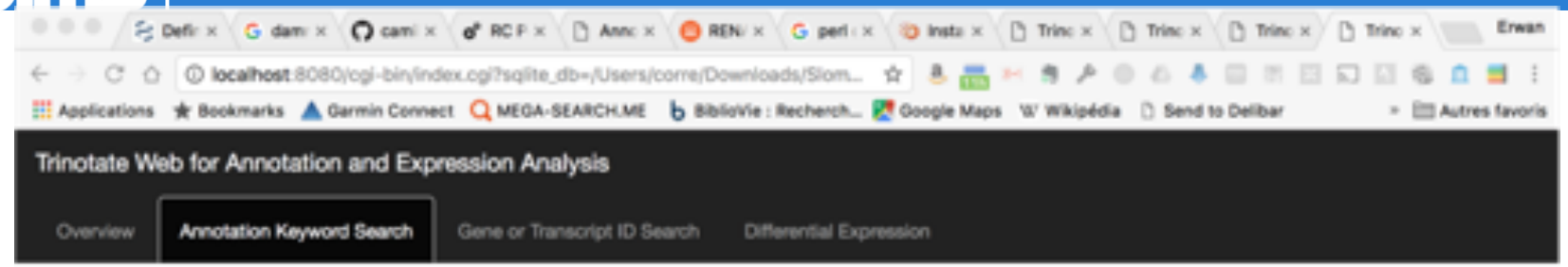
Note, Trinotate is not yet a full-featured application, but is instead in a very early state of development since 5-6 years .. :/

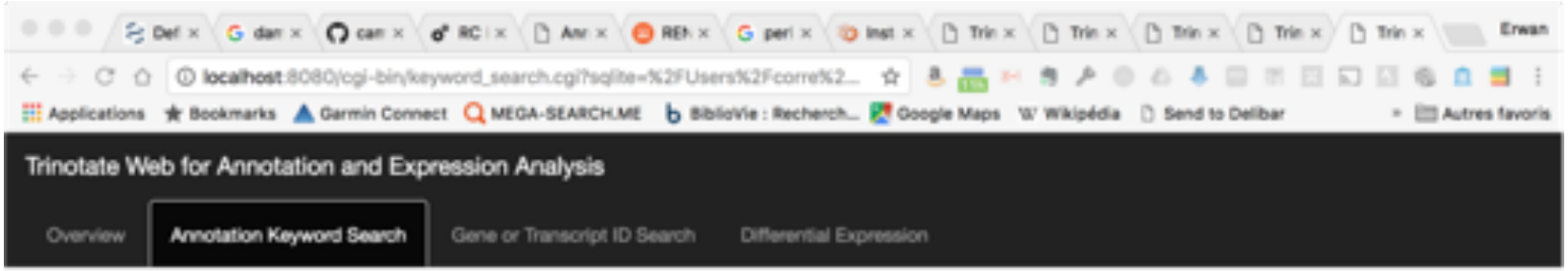
Dependancy
Lighttpd

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

CNRS UPMC
Station Biologique
Roscoff

Trinotate web





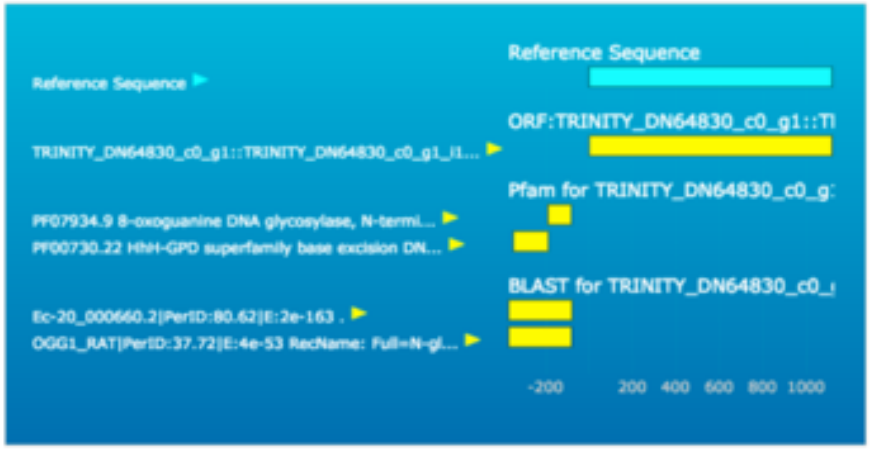
#	gene_id	transcript_id	annotation
1	TRINITY_DN583_c0_g2	TRINITY_DN583_c0_g2_i1	CYAA_STIAU^CYAA_STIAU^Q:669-448,H:334-409^32.89%ID^E:3e-06^RecName: Full=Adenylate cyclase 1;^Bacteria; Proteobacteria; Deltaproteobacteria; Myxococcales; Cystobacterineae; Cystobacteraceae; Stigmatella . TRINITY_DN583_c0_g2::TRINITY_DN583_c0_g2_i1::g.301::m.301 696-1[-] 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

Feature report for TRINITY_DN64830_c0_g1_i1

Expression Information

Transcript Annotations (Gene: TRINITY_DN64830_c0_g1, Transcript: TRINITY_DN64830_c0_g1_i1)



- 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:

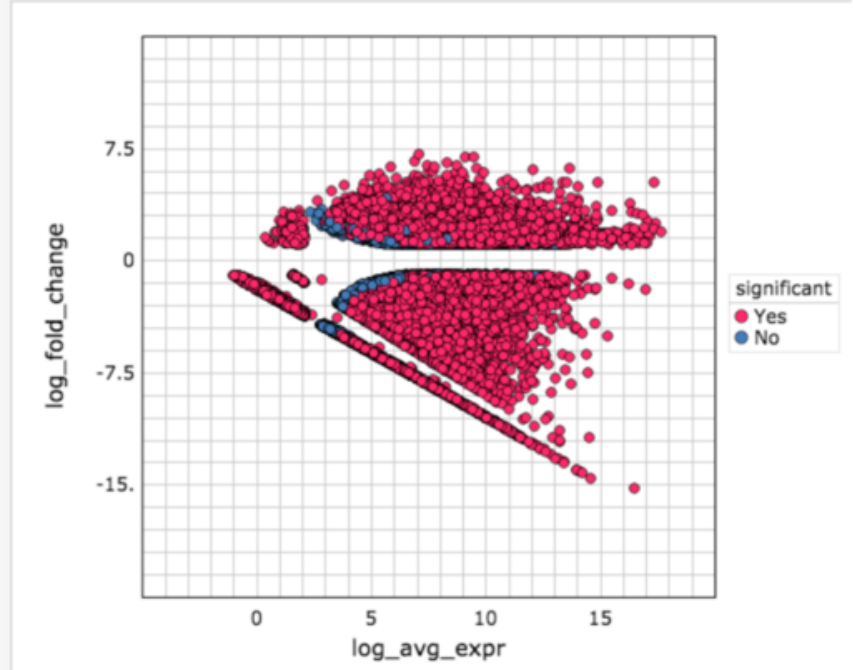
```
>TRINITY_DN64830_c0_g1_i1
GAATAGATCCCCGACACGGGCGTACACGGTAGGCGTCAACGACTTGCAGTCCAGCAAGCT
TGGGTCGTAATCTCGACACGCGATCCTCCAAACATGTACGTCCACAGGGATGGTGGAAGC
TTGATCCAGAGAAAAGAGCGCAATGCAGTCCGCCACCTTCGGACCTACGCCACACAAGGT
AATCAGCTGGTTTCGAACTTCGTCTCTCTCCTTGTTCCTCATTTCCAGCGCCCACGTCTC
CCCGCCGTTGGCGTGCATTGCCCTTGCGCTTTCCACTATGTACTTGGCACGATAGCCGAA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

- peptide sequences:

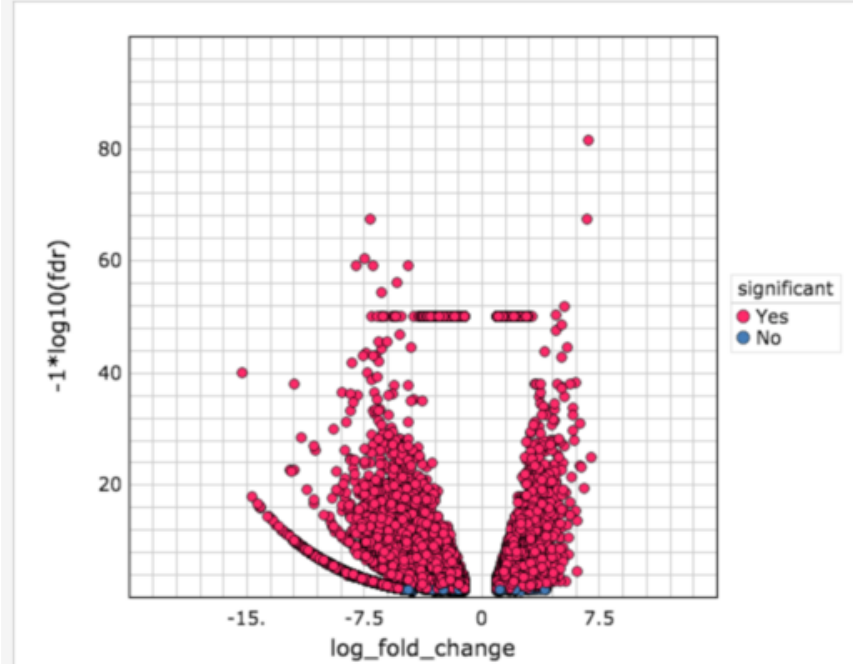
```
>TRINITY_DN64830_c0_g1::TRINITY_DN64830_c0_g1_i1::g.53680::
LDSPLFTPNDWLNKTRICAYCATRRKGVVRAACRAFVHIINMSPQRSAAWGTALSSRRR
CFELSFSSTATSFVFALVFICSANTGPDCWVGVLGREVIAIRQTPDTTLFRSLSVAKKT
HEDVKMATASHVAADGTATAALAATLREYFFLSVPLAPLYRRWSEGDARMAAVAASIPG
RVVRQTPVECIFSFICSSNNNIPRITGMLDKLRTTYGELLLSVGKGGLAATGALGDMKE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

Trinotate web

MA plot: Slom_GA vs. Slom_SP



Volcano plot: Slom_GA vs. Slom_SP



Trinotate Web for Annotation and Expression Analysis

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

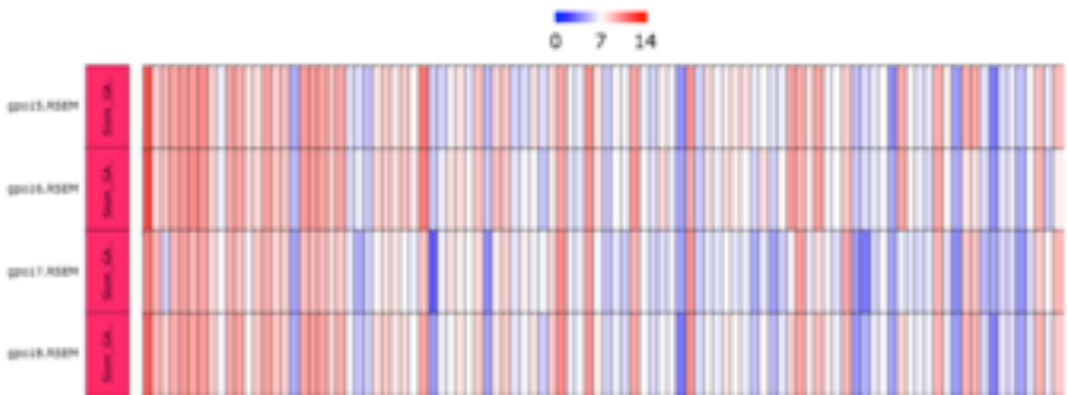
Restrict to top-most expressed in any given sample.

Max genes to show: 100

Valider

(Only 100 of 4609 randomly selected features are shown)

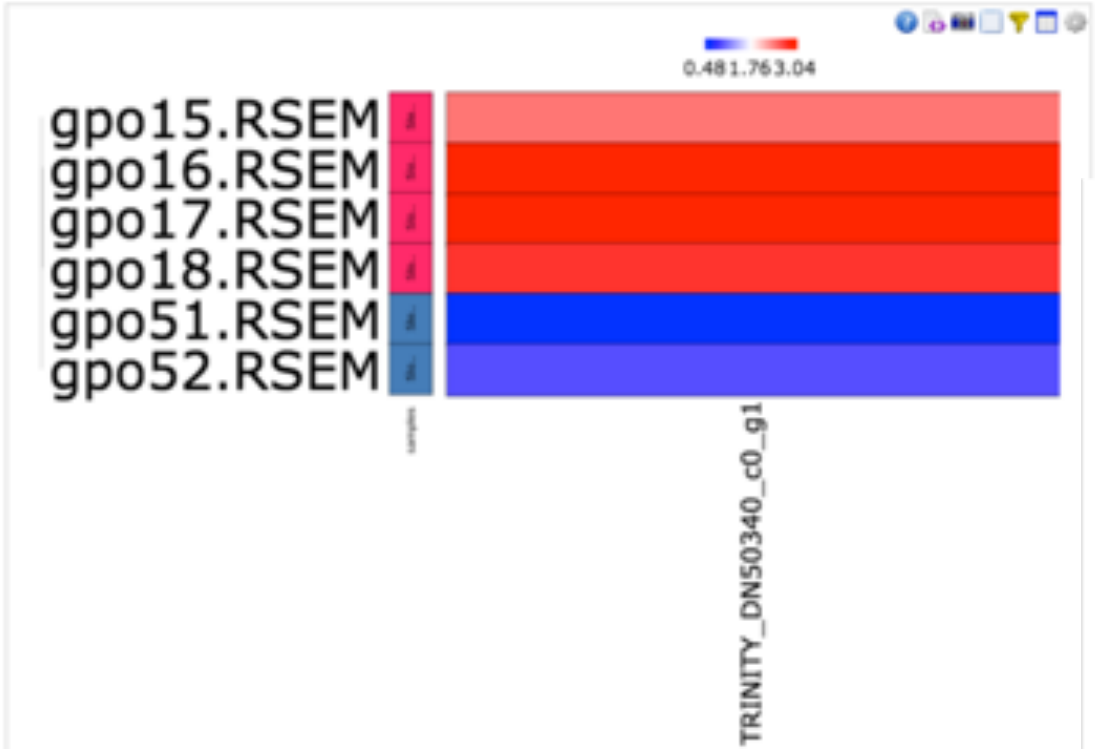
Found 100 features.

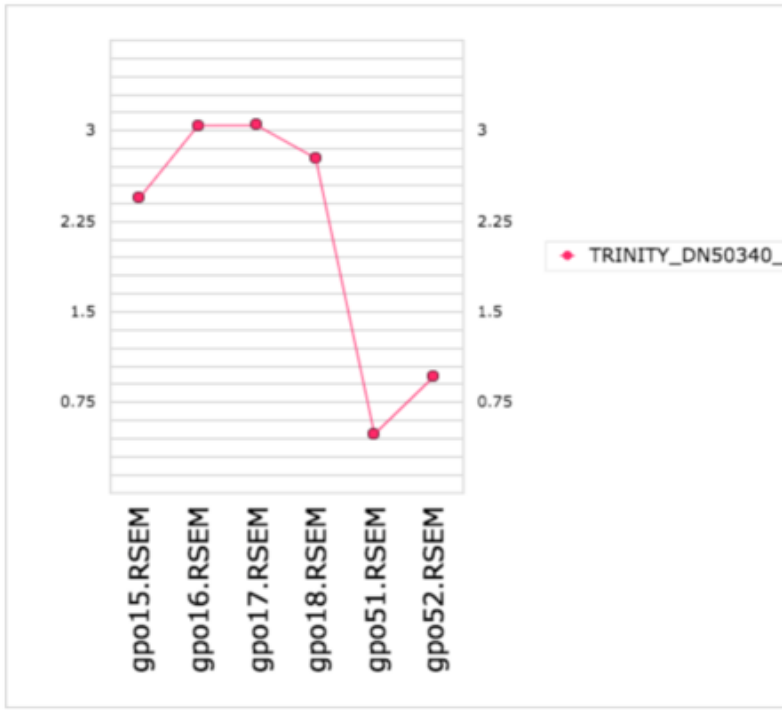


Trinotate Web for Annotation and Expression Analysis

Feature report for TRINITY_DN50340_c0_g1

Expression Information



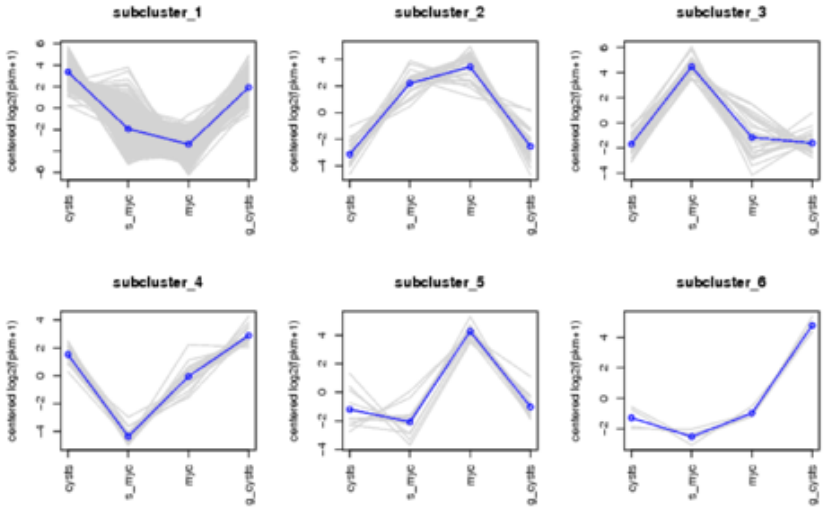


Transcript Annotations (Gene: TRINITY_DN50340_c0_g1, Transcript: TRINI

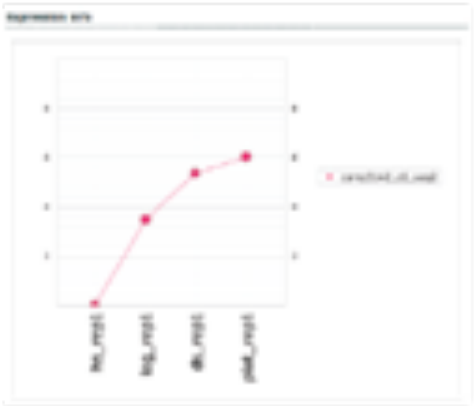
CNRS UPMC
Station Biologique
Roscoff

Abims 4

Clustered Expression Profiles



Transcript/Protein Annotation Report
Blast Hits, Pfam Domains, etc.



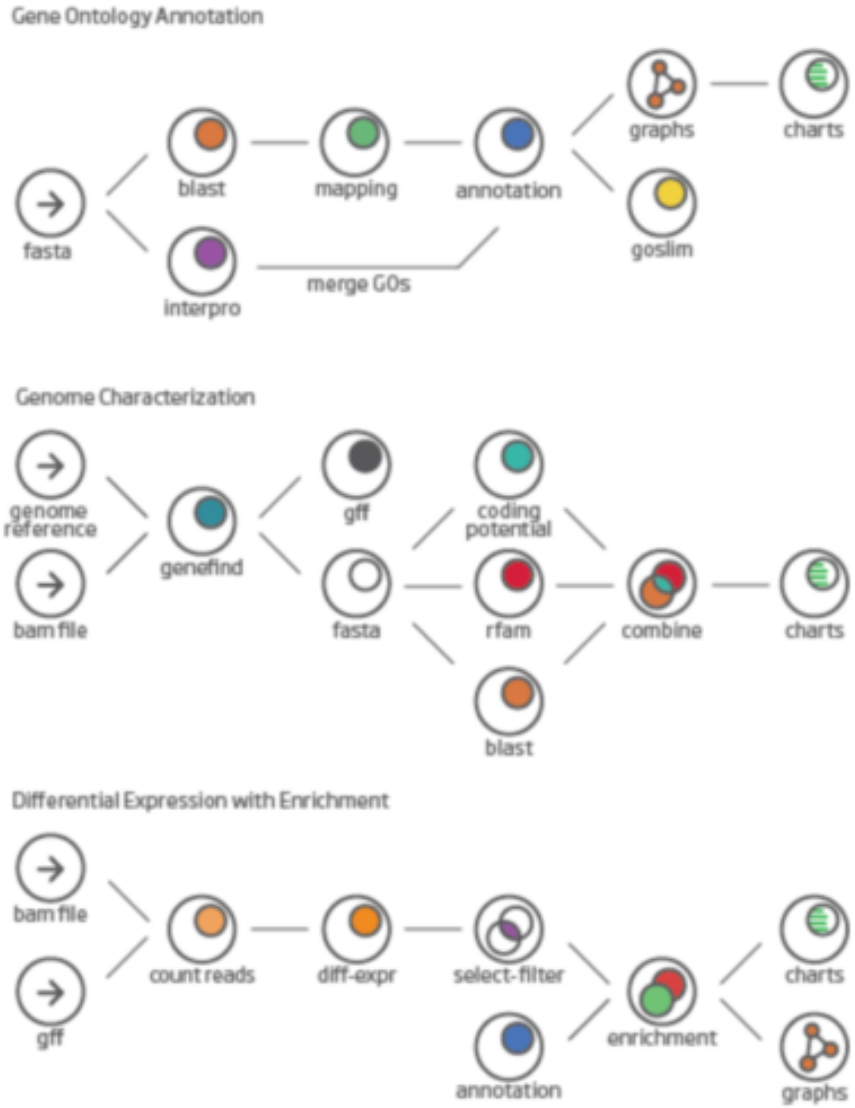
Individual Transcript Expression Profiles

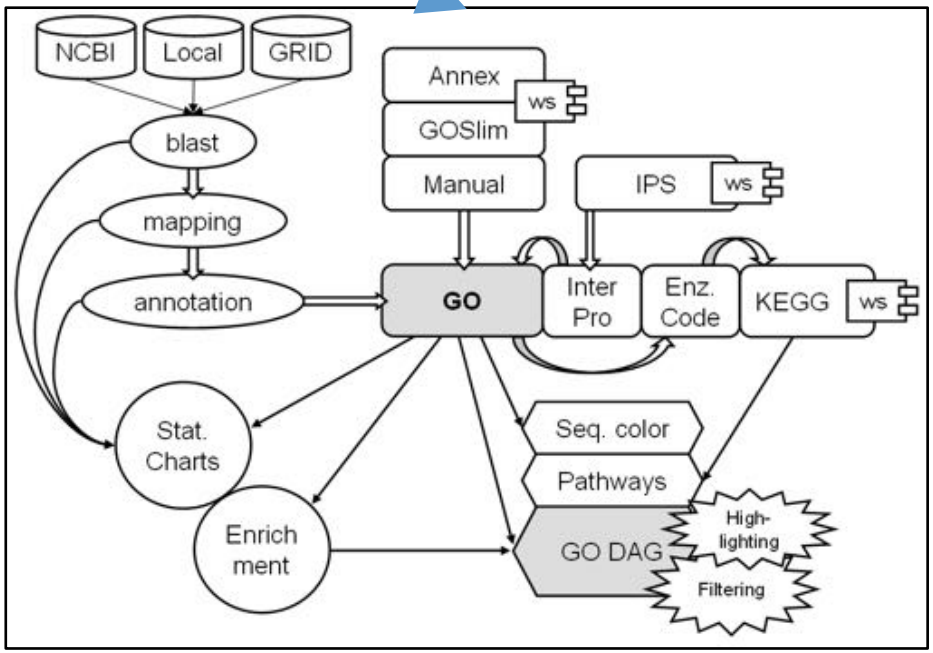
Transcript and Protein Sequence

Alternatives

- Blast2Go
- FunctionAnnotator
- Annoscript
- Dammit
- KOBAS
- Others

Blast2GO Schema





Blast2GO Annotation Rule

Lowest term satisfying the requirements

Similarity requirement

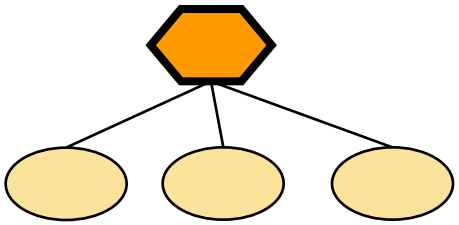
$$sim = \frac{\sum positives_{hsp}}{\sum alignmentlength_{hsp}}$$

Quality of source annotation

EC	weight
IC	1
TAS	1
IDA	1
IMP	0.9
IGI	0.9
IPI	0.9
ISS	0.8
IEP	0.8
NAS	0.7
IEA	0.7
ND	0.5
NR	0.5
RCA	0.5

Evidence Codes

Possibility of abstraction



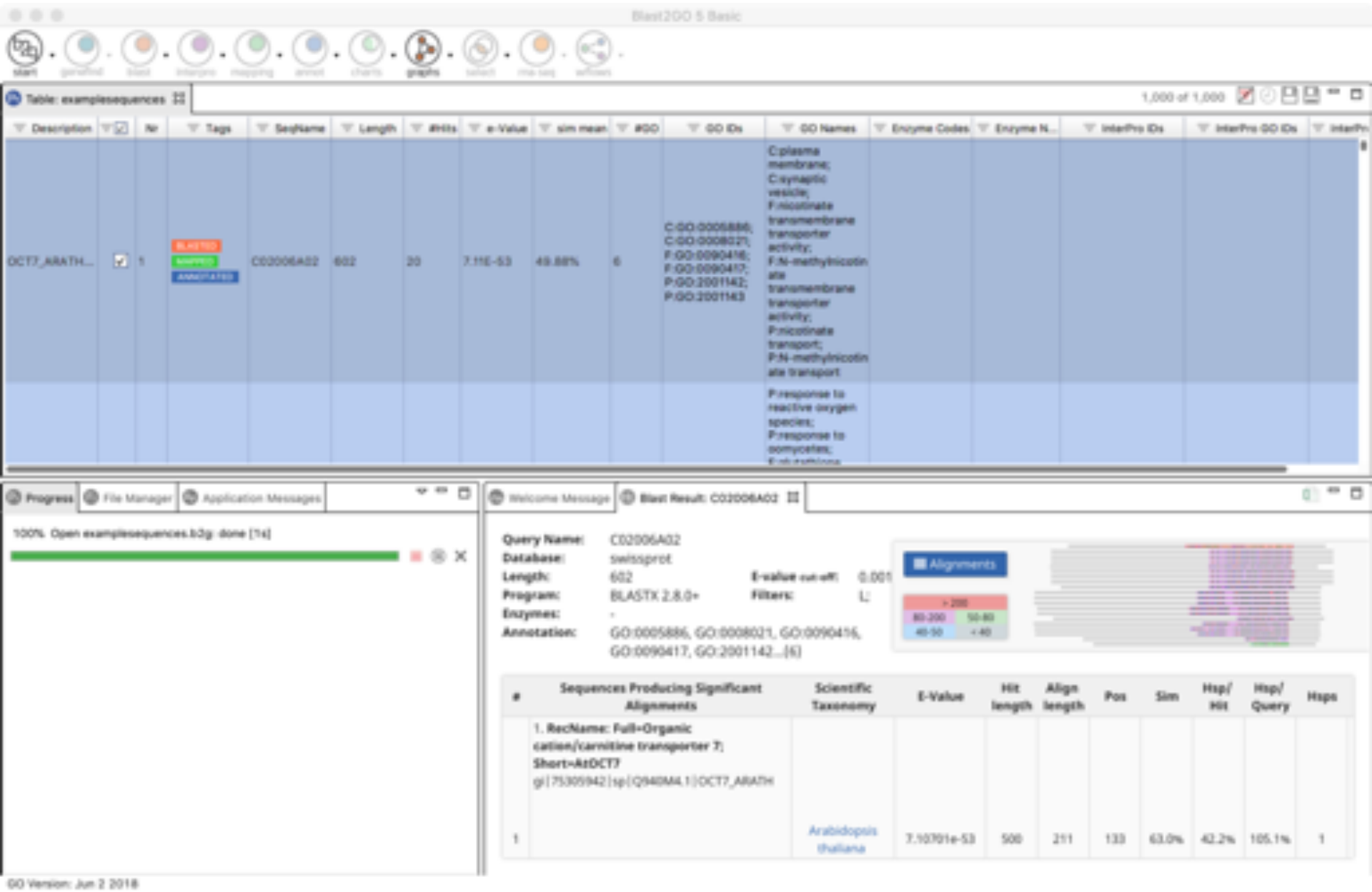
Recall vs. Precision

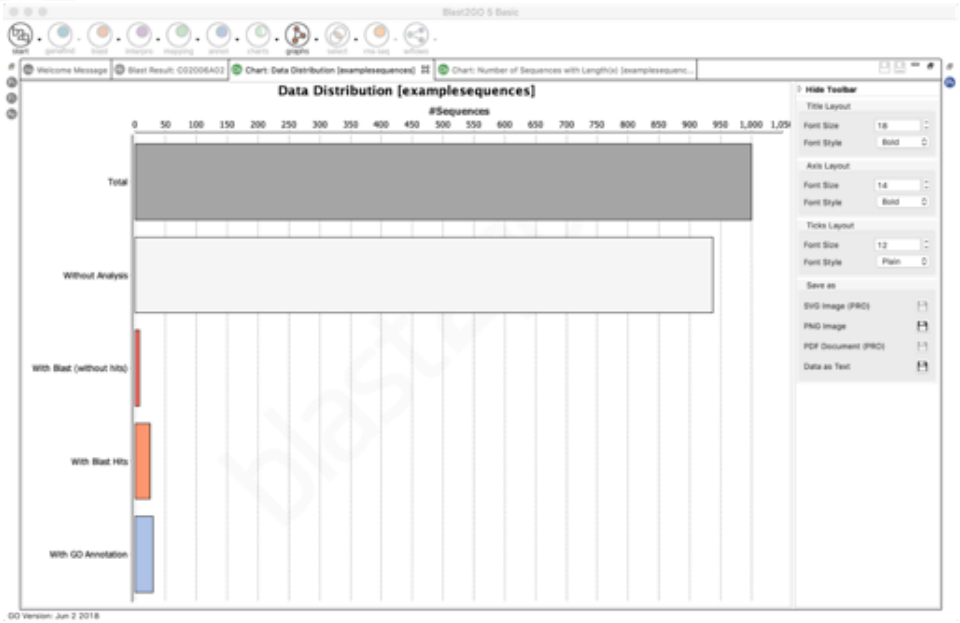
$$lowest.node[(\max .sim \times ECw) + (\#GO \times GOw)] \leq threshold$$

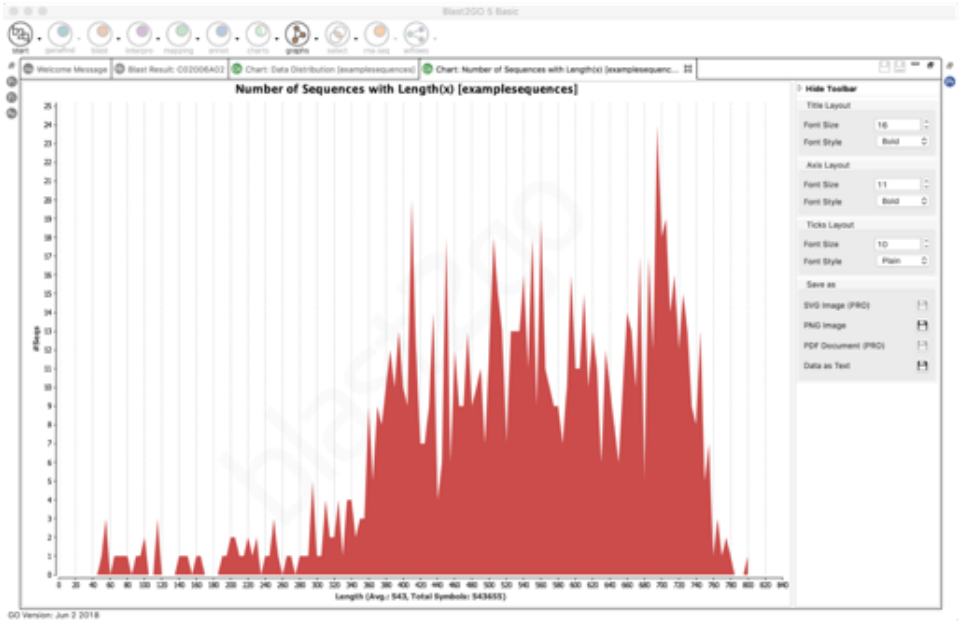
Annotation Rule

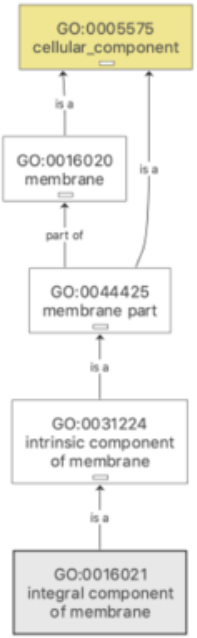
CNRS UPMC
Station Biologique
Roscoff

Abims 4



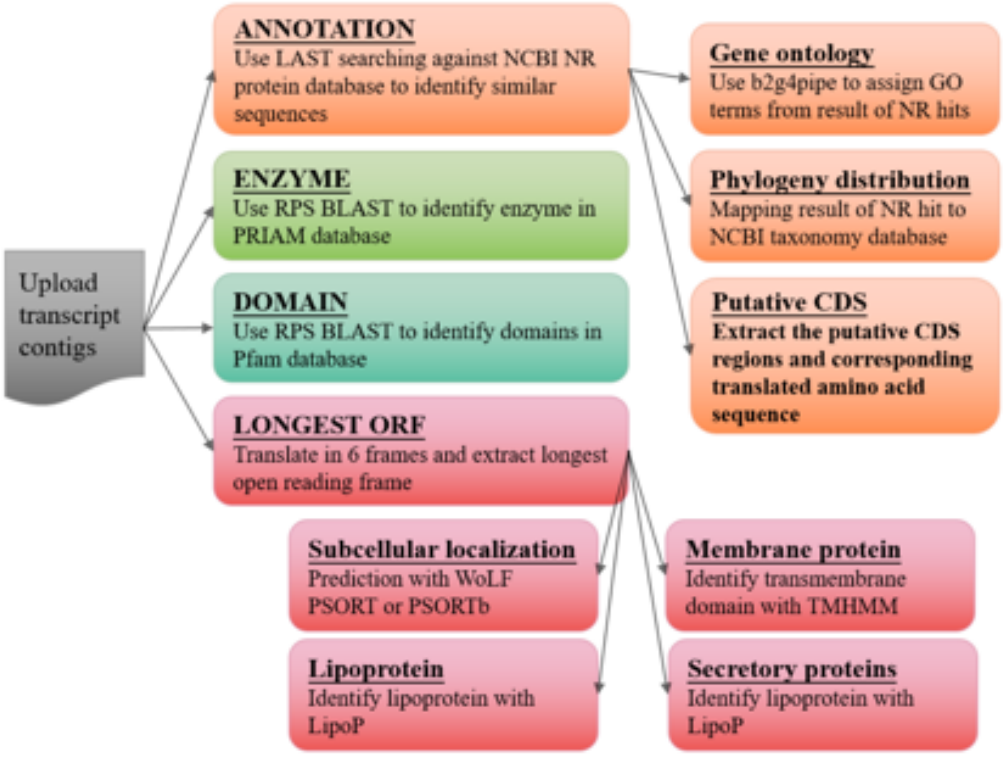

Data Distribution [examplesequences]
#Sequences
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950 1,000 1,05
Total
Without Analysis
With Blast (without hits)
With Blast Hits
With GO Annotation


Number of Sequences with Length(x) [examplesequences]
#Seqs
Length (Avg.: 543, Total Symbols: 543655)


GO:0005575
cellular_component
is a
GO:0016020
membrane
is a
part of
GO:0044425
membrane part
is a
GO:0031224
intrinsic component
of membrane
is a
GO:0016021
integral component
of membrane

CNRS UPMC
Station Biologique
Roscoff

*Function*Annotator



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

FunctionAnnotator

Home | Analysis | Tutorial | Demo & Benchmark

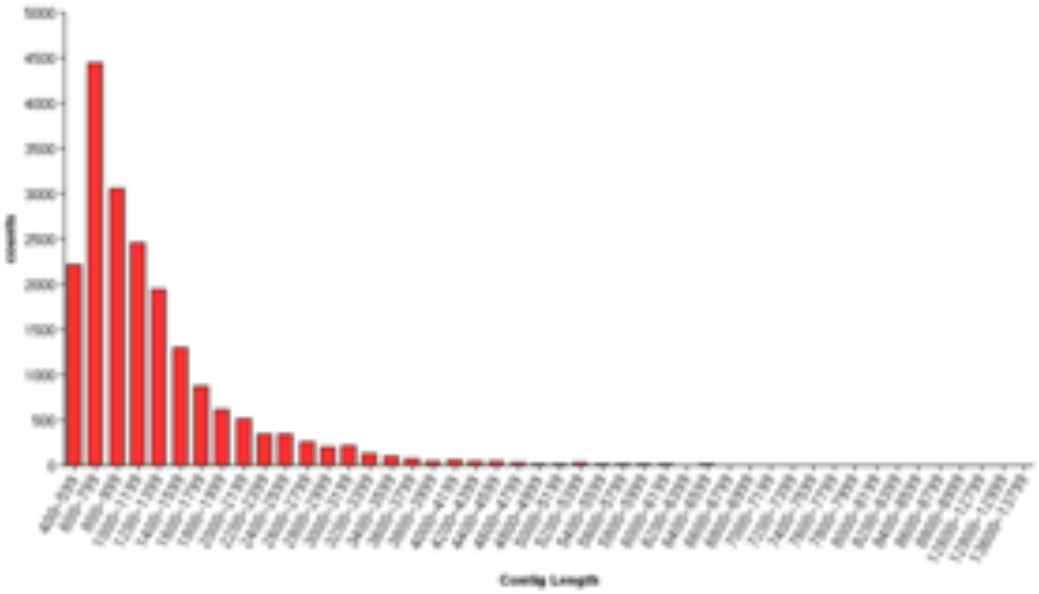
Job ID	1483622587857
Fasta file	TI_RNAseq_Contig.fa
File size	25,388,148 bytes
Number of Entries	19,415 entries
Uploaded on	Thu, 05 Jan 17 21:23:17 +0800

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
TriBase	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

Annocript

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

Annocript

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)

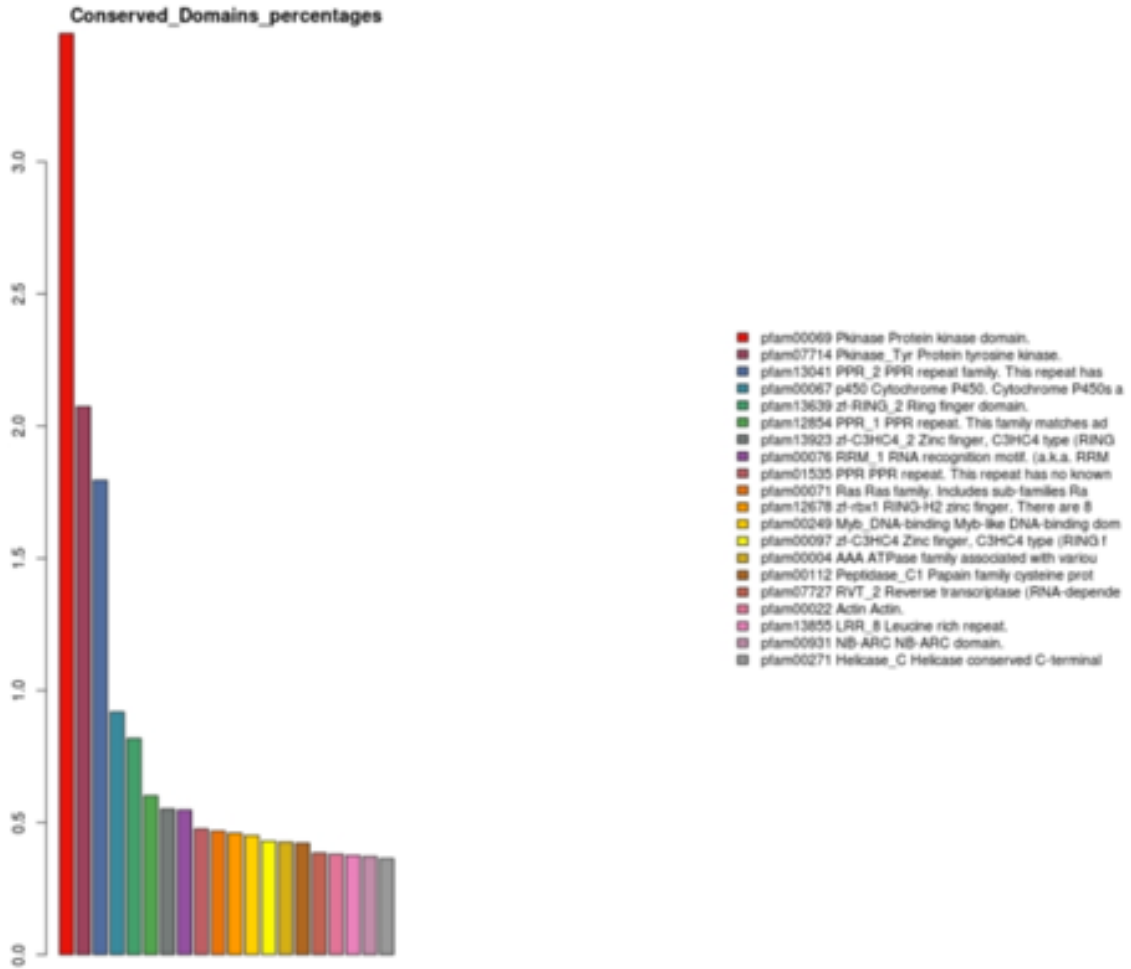
Statistics for transcriptome | Homology statistics | Lengths and coverage |

Fri Jan 16 18:16:05 2015

Results: graphical representation

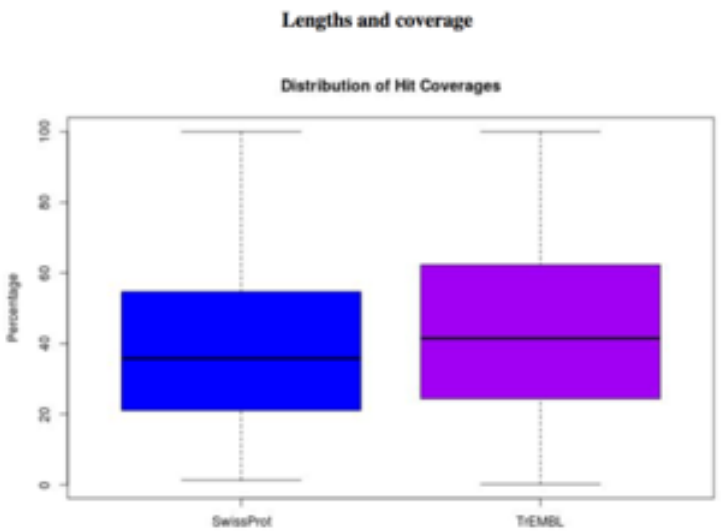
Annocript

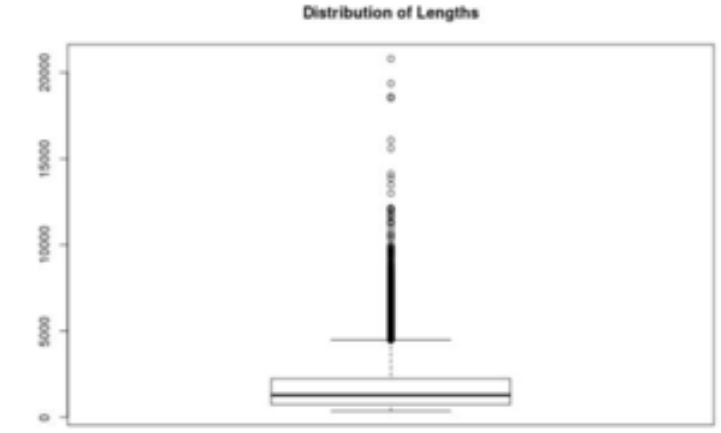
Homology statistics

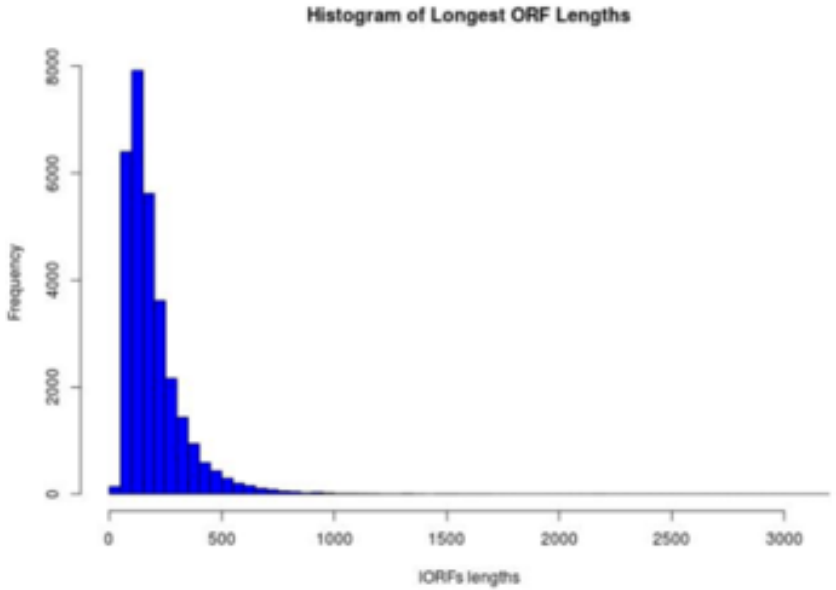


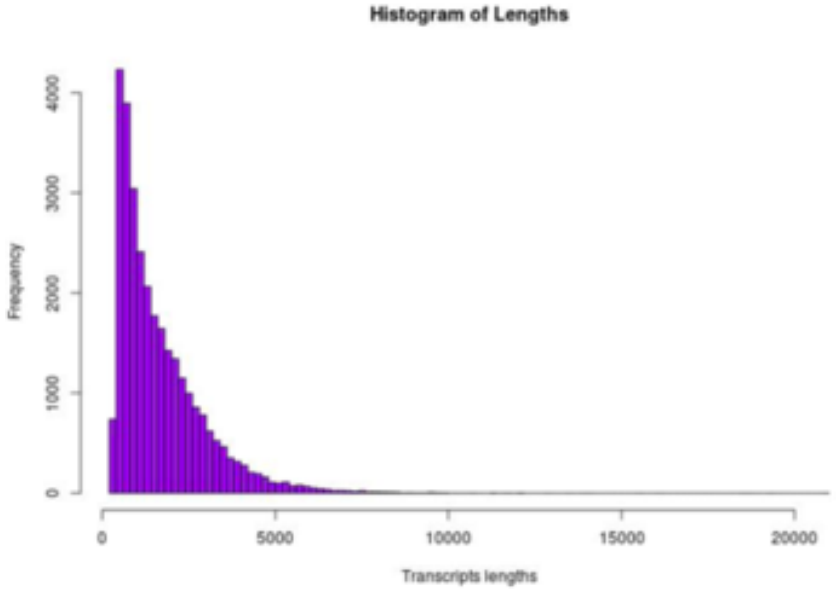
Results: graphical representation

Annocript





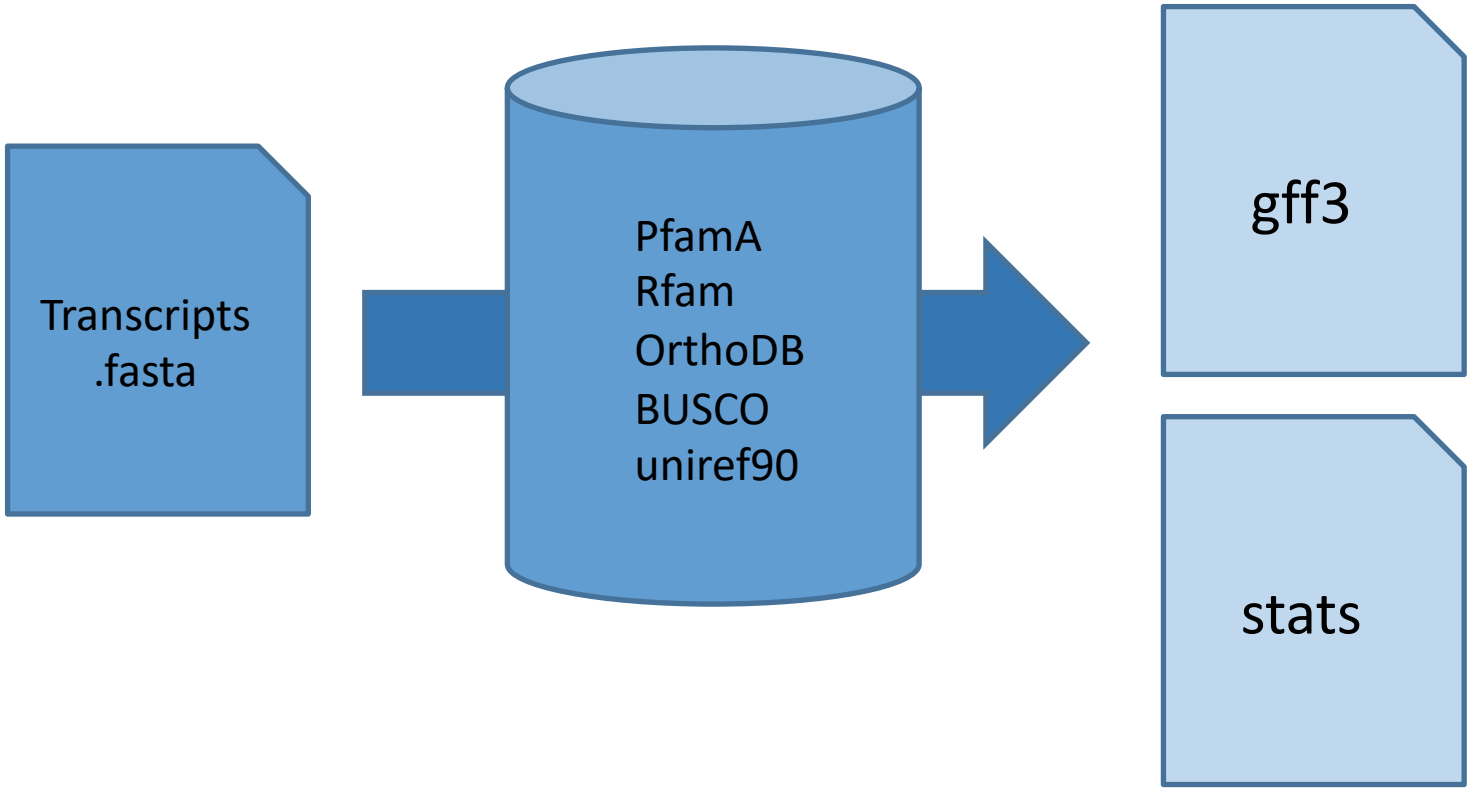




Dammit

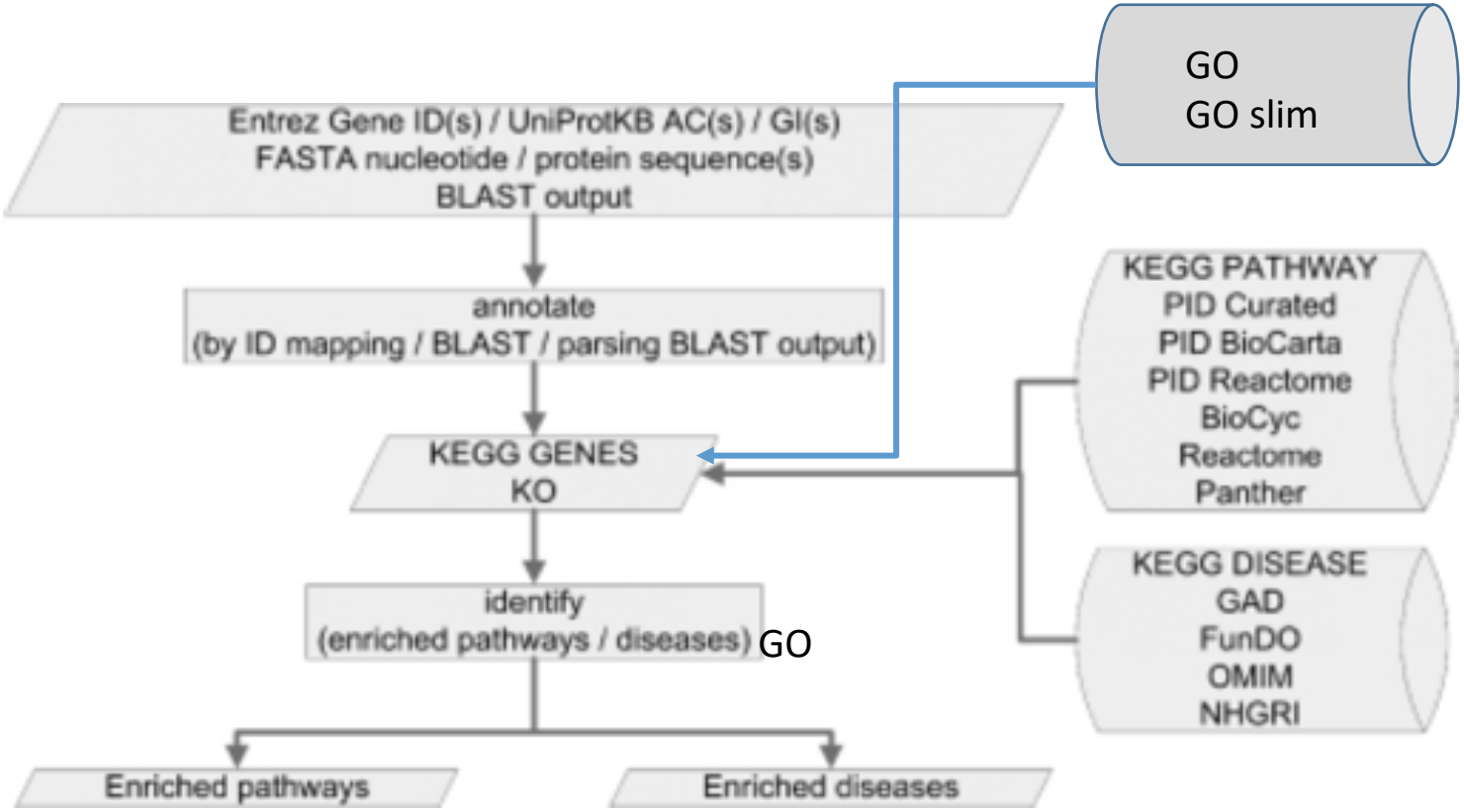
http://www.camillescott.org/dammit/

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

KOBAS 3.0 | Home | Annotate | Gene-list Enrichment | Exp-data Enrichment | Download | Help

Link of this page http://kobas.cbi.pku.edu.cn/result_annotate.php?taskid=1605145061355565 *(You can save this link to fetch results directly in the future.)*

Download the result file

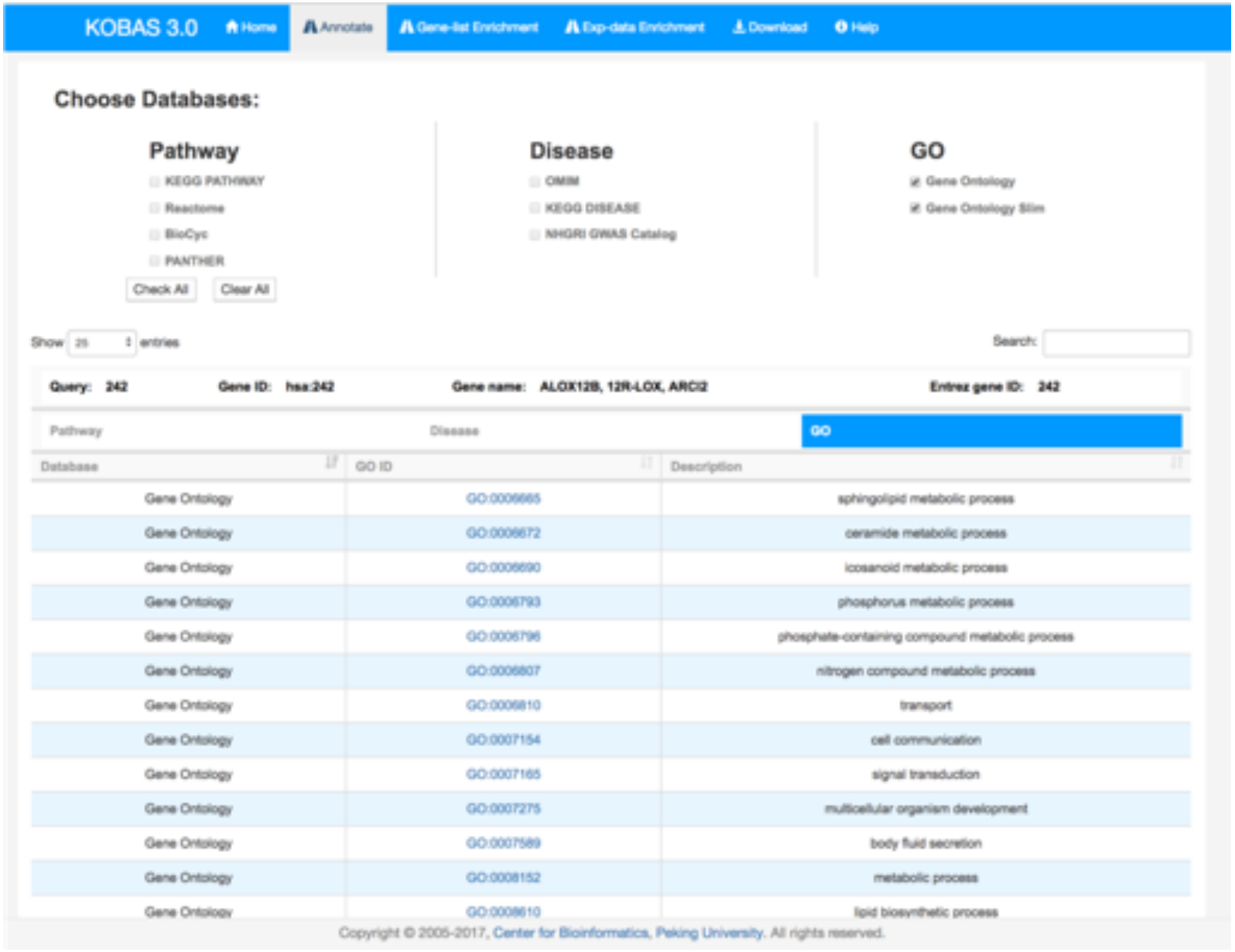
Notes: this output file can used as the input file of 'Gene-list Enrichment', or maybe the background file.

Show 25 entries Search:

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 First Previous 1 Next Last

KOBAS : KO-Based Annotation System



EnTAP (Eukaryotic Non-Model Transcriptome Annotation Pipeline)

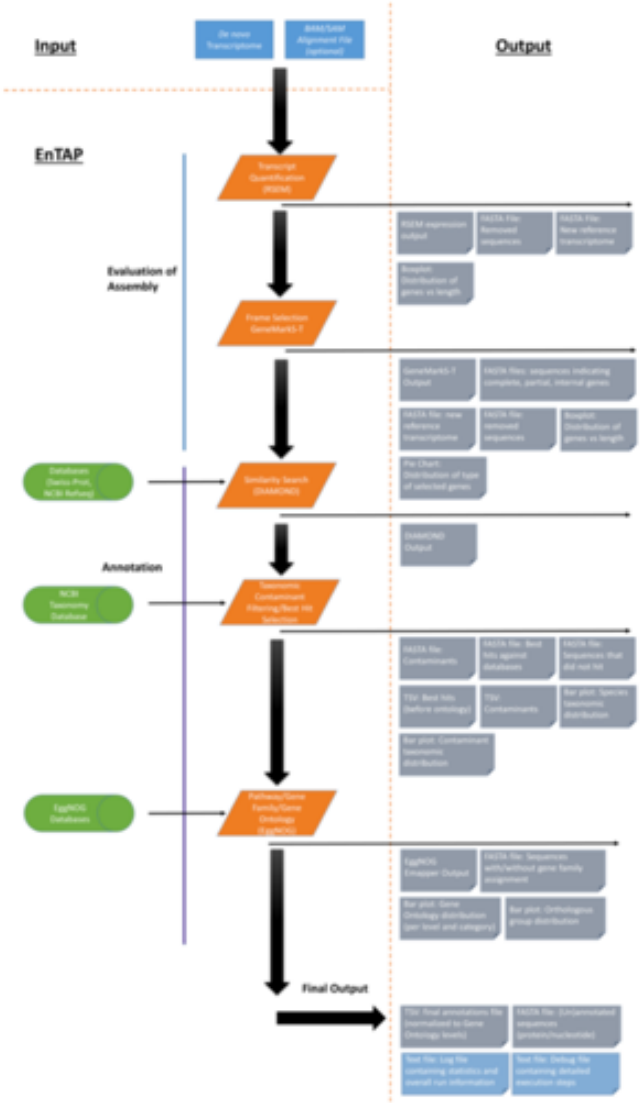
EnTAP: Bringing Faster and Smarter Functional Annotation to Non-Model Eukaryotic Transcriptomes

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



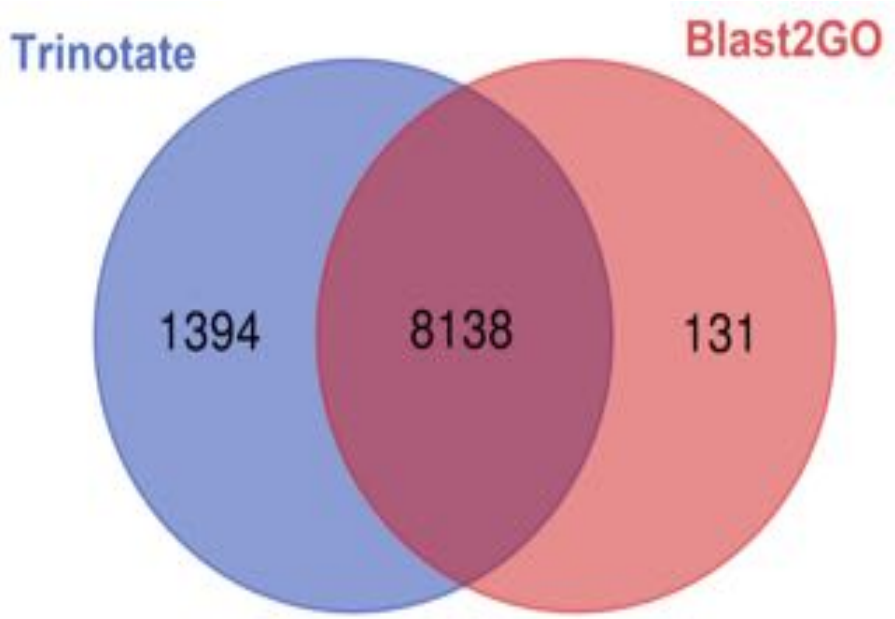
Transcriptome filtering :
RSEM

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

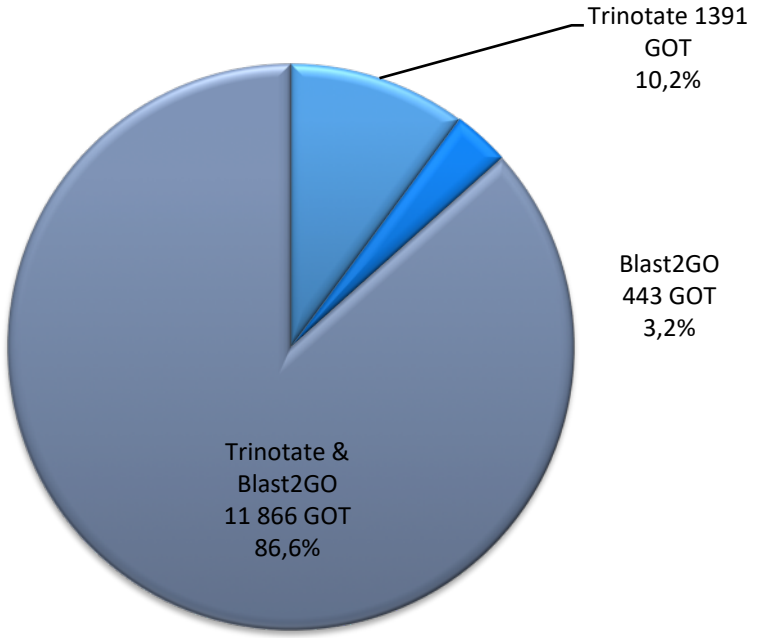
DIAMOND (Fast and Sensitive NCBI BLAST Alternative)
Combination of curated databases (at least 3)
Selection of Optimal Hit From Several Databases

Selection of Optimal Hit Based on Informativeness
Contaminant Identification and Filtering

Trinotate vs Blast2GO : Go Terms



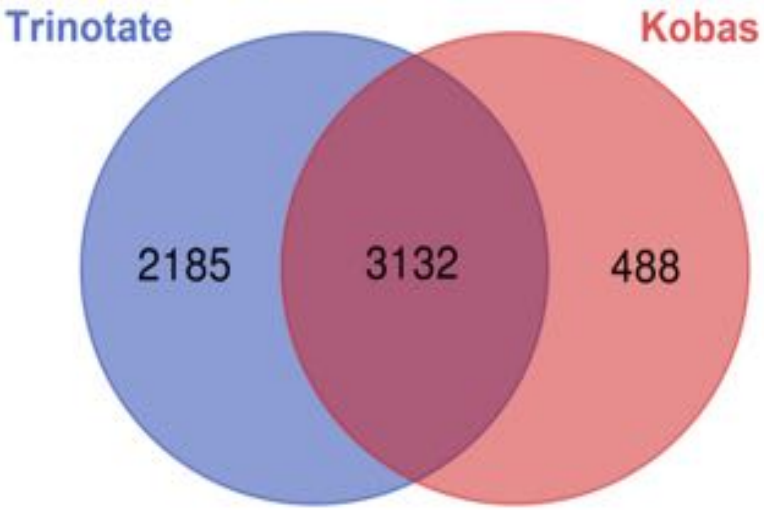
Number of sequence annotated with GO terms



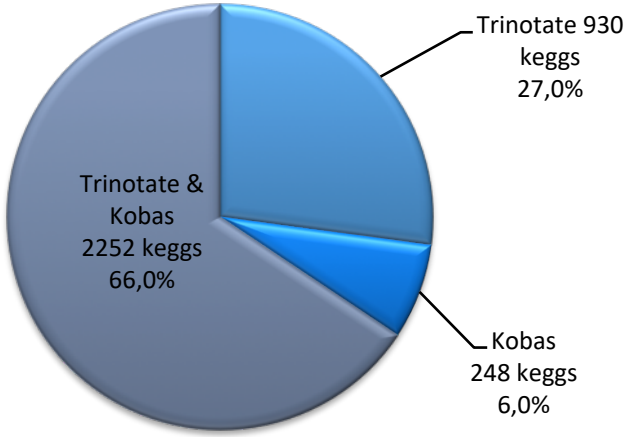
Number of GO terms

Saccharina japonica genome

Trinotate vs Kobas : Kegg



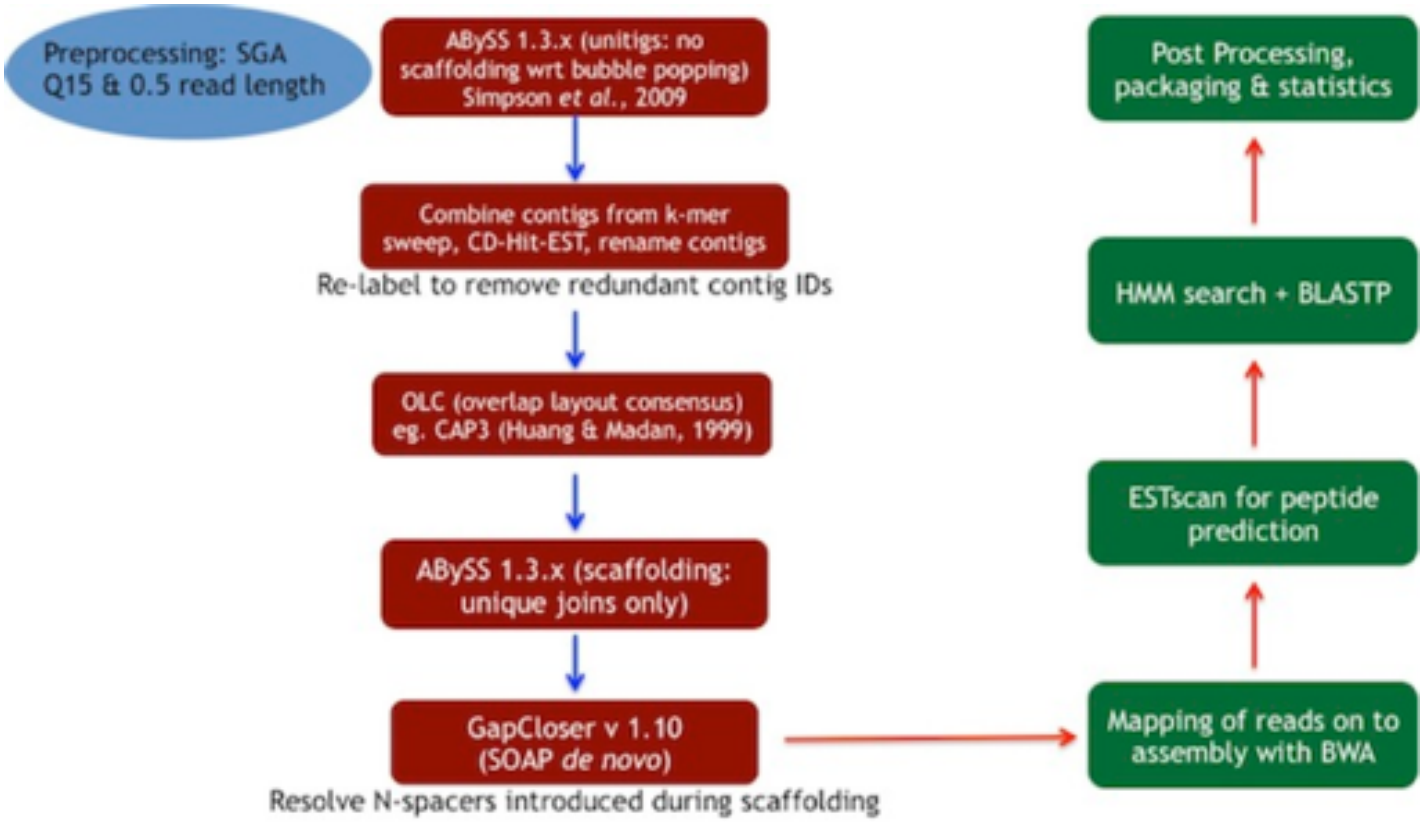
Number of sequence annotated with KEGG terms



Number of KEGG terms

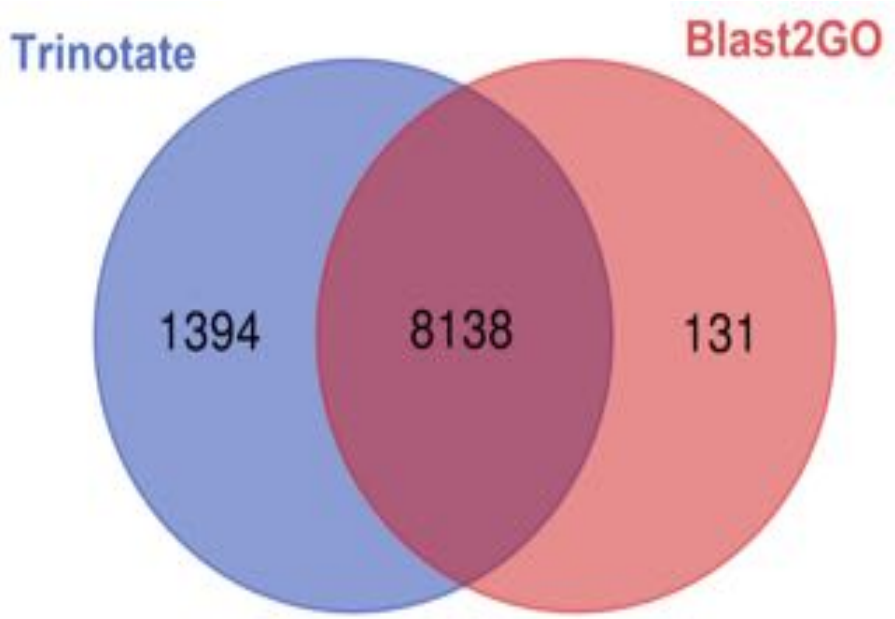
Saccharina japonica genome

CAMERA (NCGR : national center for genome ressources) Annotation process

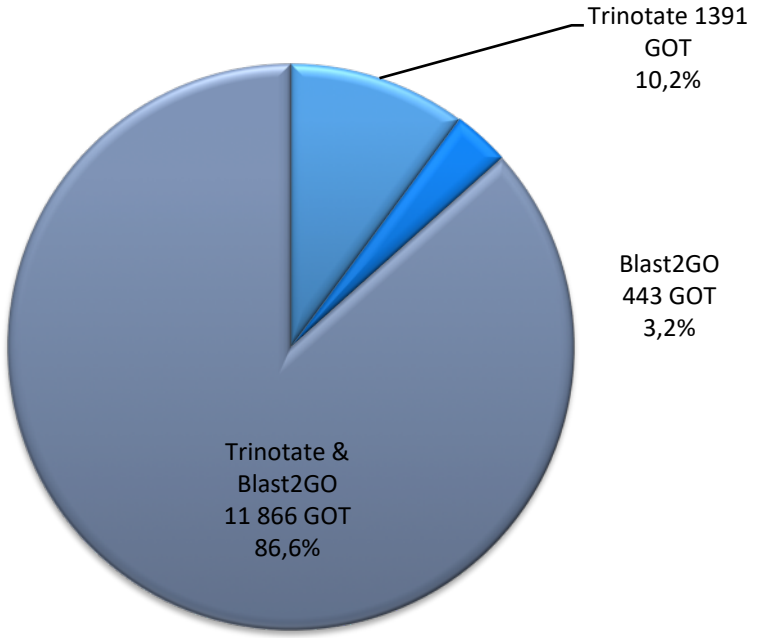


Annotation files based on hits to Swiss-Prot, Pfam-A, and TIGRFAMs include InterPro associations in the Ontology term attribute

Trinotate vs Blast2GO : Go Terms



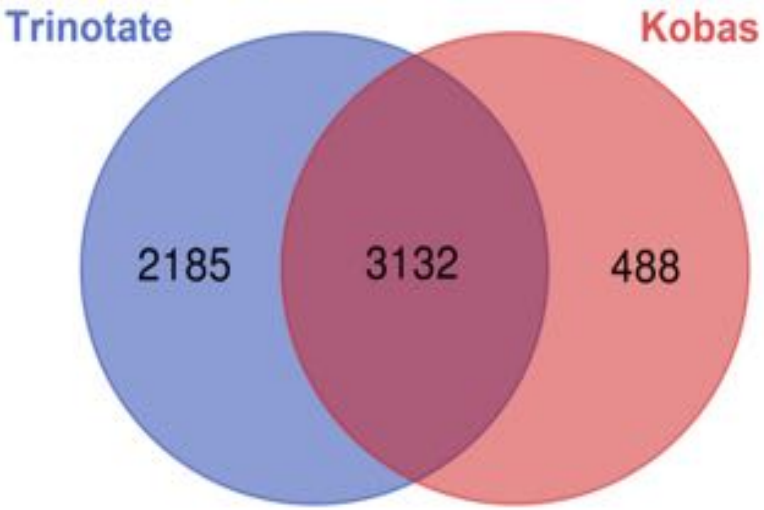
Number of sequence annotated with GO terms



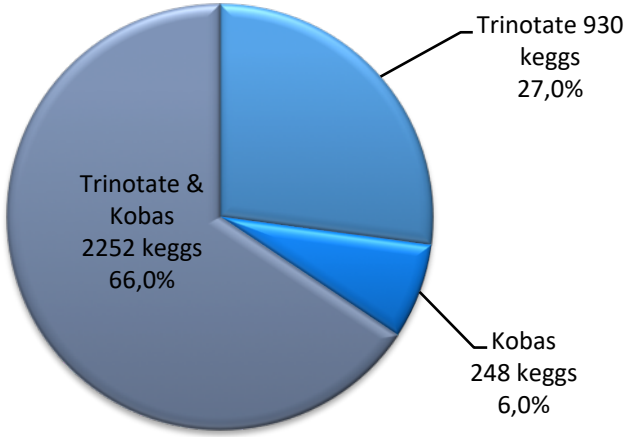
Number of GO terms

Saccharina japonica genome

Trinotate vs Kobas : Kegg

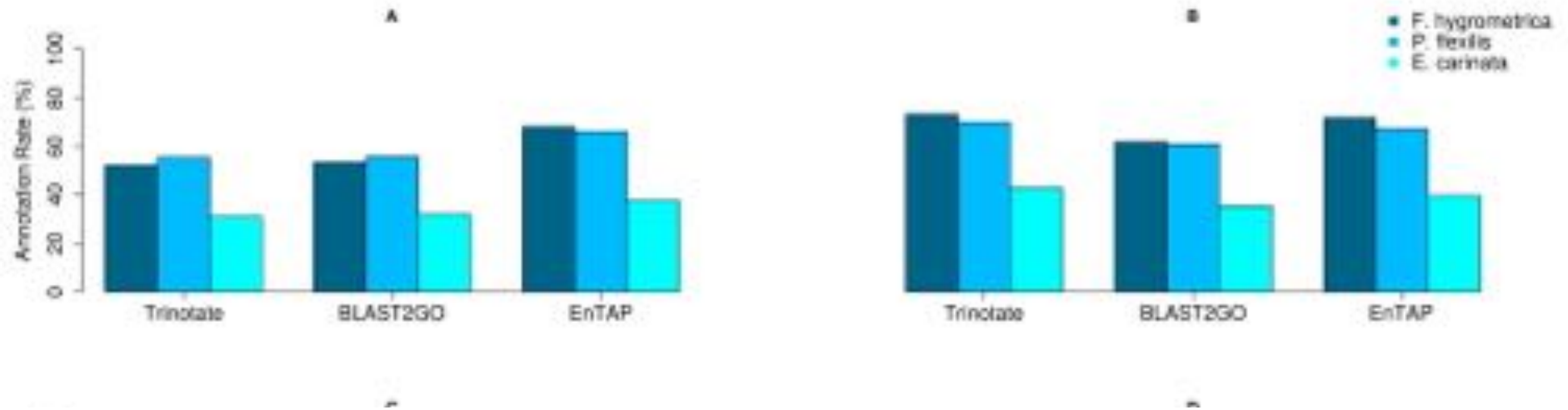


Number of sequence annotated with KEGG terms



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.

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.

TRAPID: Rapid Analysis of Transcriptome Data

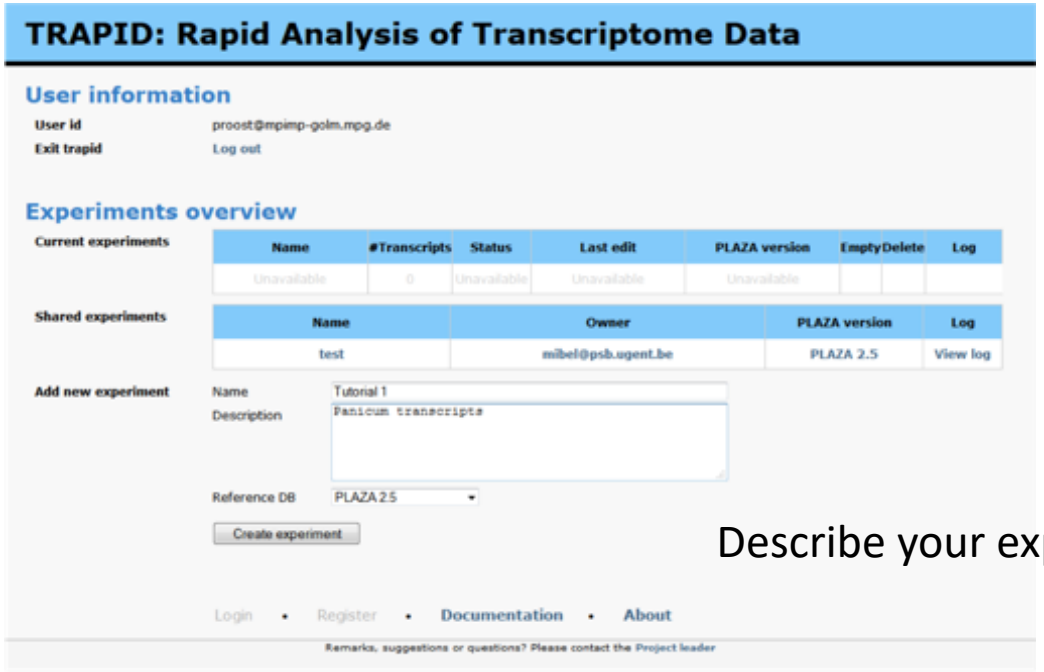
http://bioinformatics.psb.ugent.be/webtools/trapid/

TRAPID system offers functional and comparative analyses for transcriptome data sets

Two reference databases:

- for plants and green algae PLAZA 2.5,
- for Alveolata, Amoebozoa, Euglenozoa, Fungi, Metazoa and prokaryotes (Bacteria and Archaea) OrthoMCL-DB version 5 is available.

- ORF detection,
- frameshift correction
- includes a functional, comparative and phylogenetic toolbox

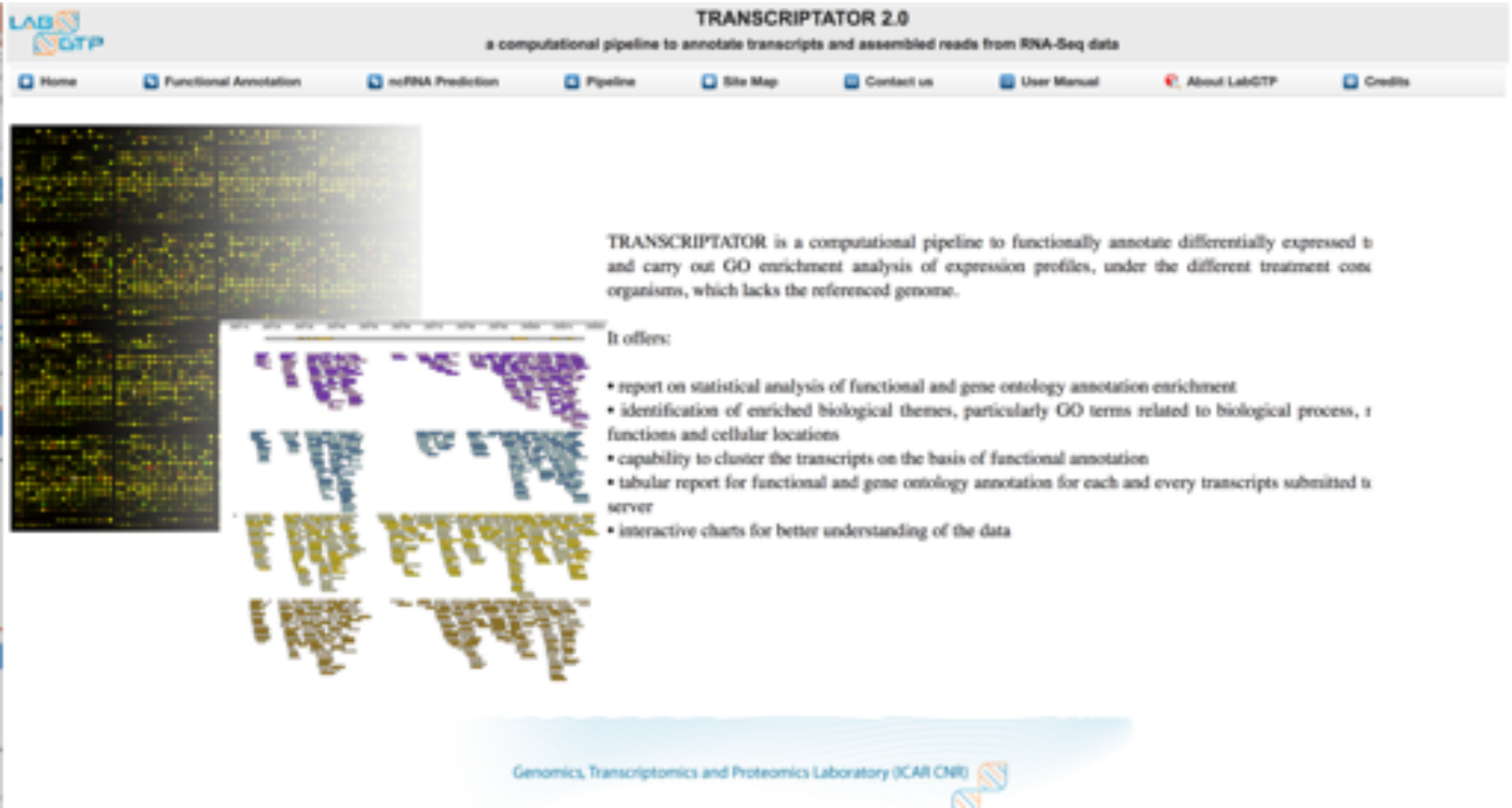


Describe your experiment

Transcriptator

Transcriptator: An Automated Computational Pipeline to Annotate Assembled Reads and Identify Non Coding RNA

http://www-labgtp.na.icar.cnr.it/Transcriptator/



Transcriptom visualisation tools

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

RNAbrowse :
Mariette J, Noirot C, Nabihoudine I, Bardou P, Hoede C, et al. (2014) RNAbrowse: RNA-Seq De Novo Assembly Results Browser. PLoS ONE 9(5), e96821.

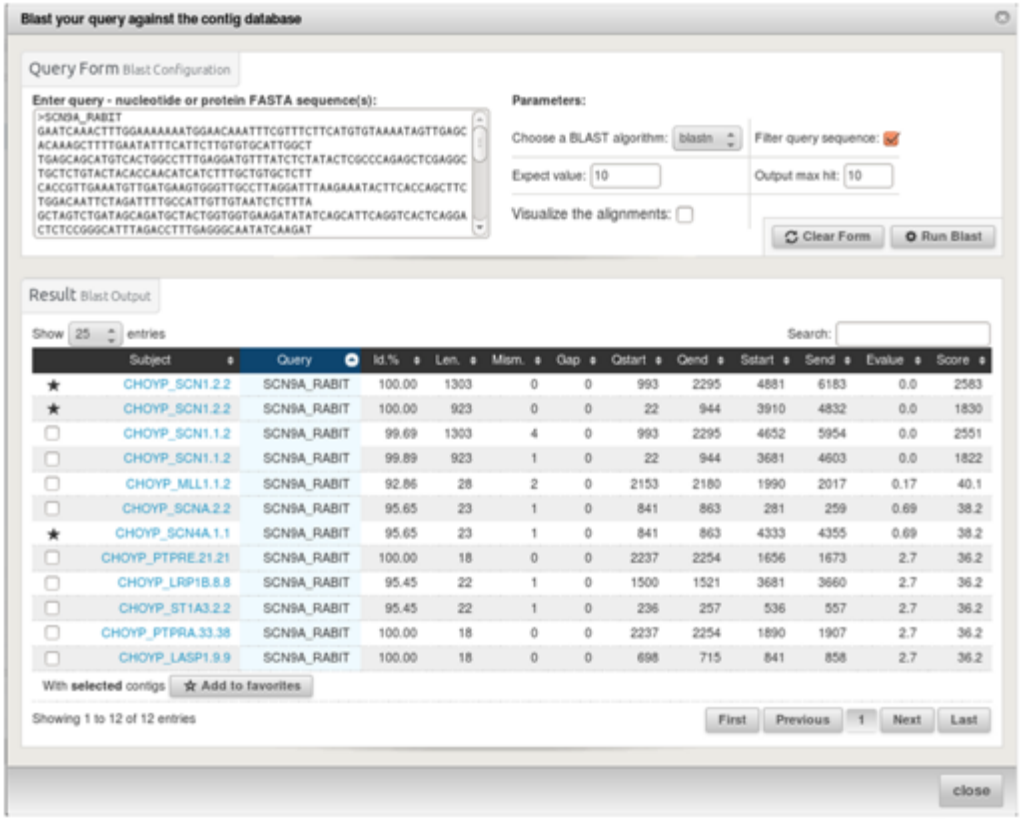
RNAseqViewer :
Rogé X , and Zhang X Bioinformatics 2014;30:891-892

TraV :
Dietrich S, Wiegand S, Liesegang H (2014) TraV: A Genome Context Sensitive Transcriptome Browser. PLoS ONE 9(4)

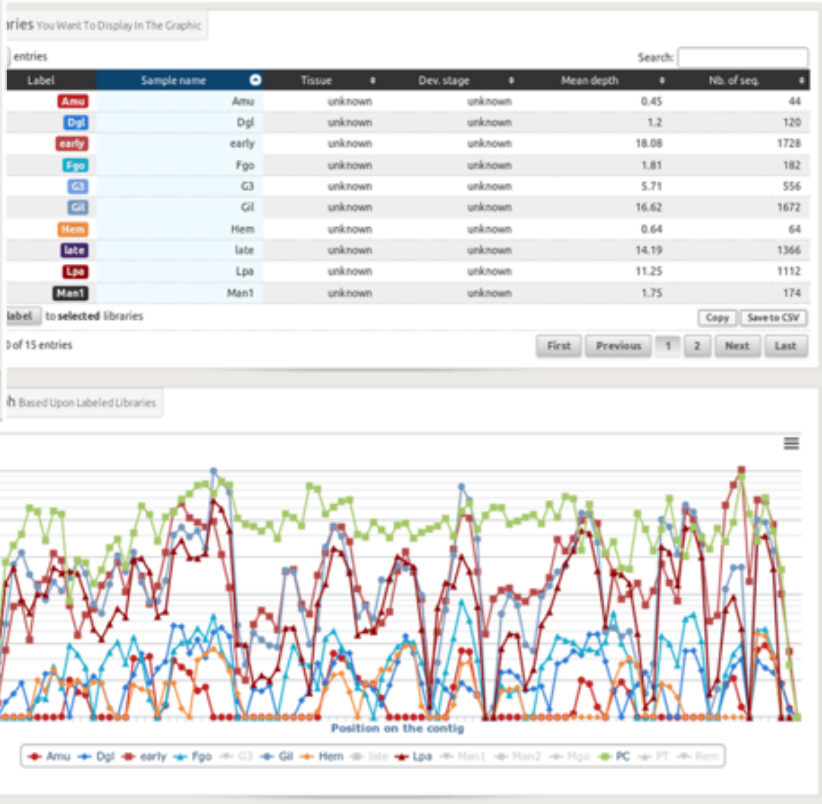
RNASeqExpressionBrowser :
Nussbaumer, T., Kugler, K. G., Bader, K. C., Sharma, S., Seidel, M., & Mayer, K. F. X. (2014). RNASeqExpressionBrowser - A web interface to browse and visualize high-throughput expression data. Bioinformatics.doi:10.1093/bioinformatics/btu334

RNAbrowse

http://bioinfo.genotoul.fr/RNAbrowse.

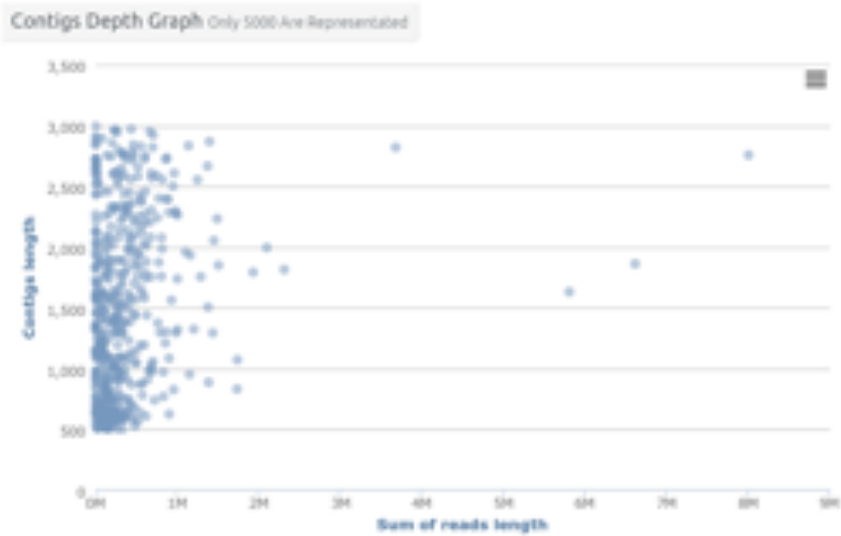


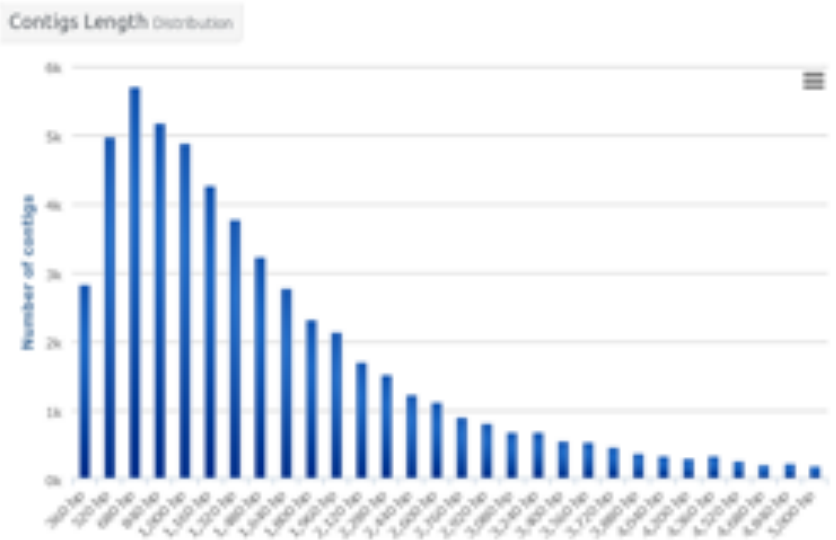
Blast interface

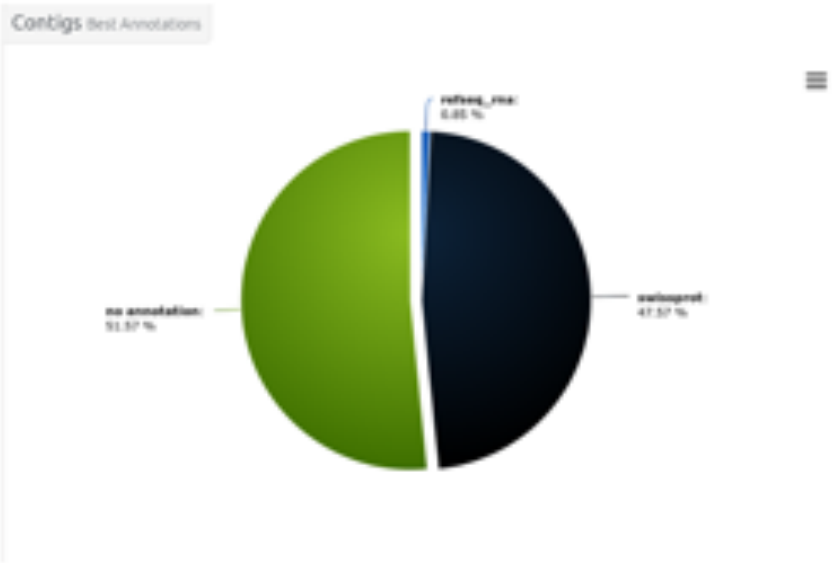


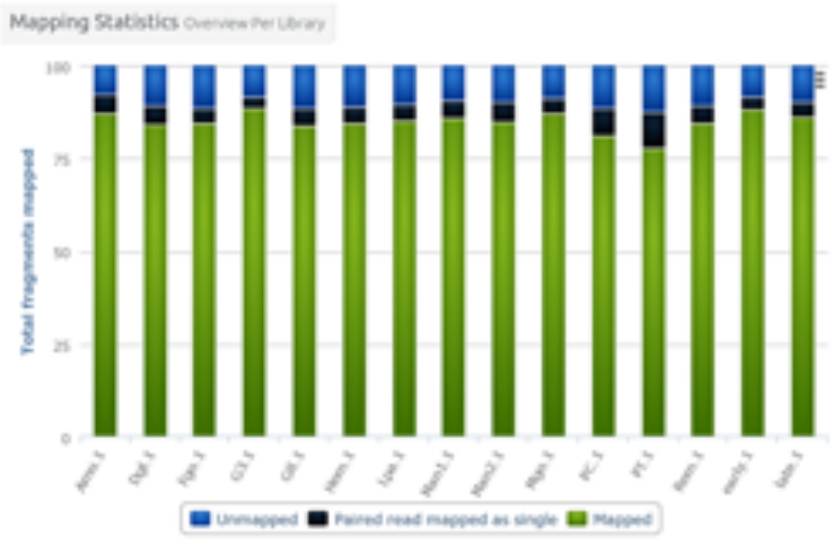
The contig depth view enables to visualise the coverage of the reads of the different libraries

RNAbrowse





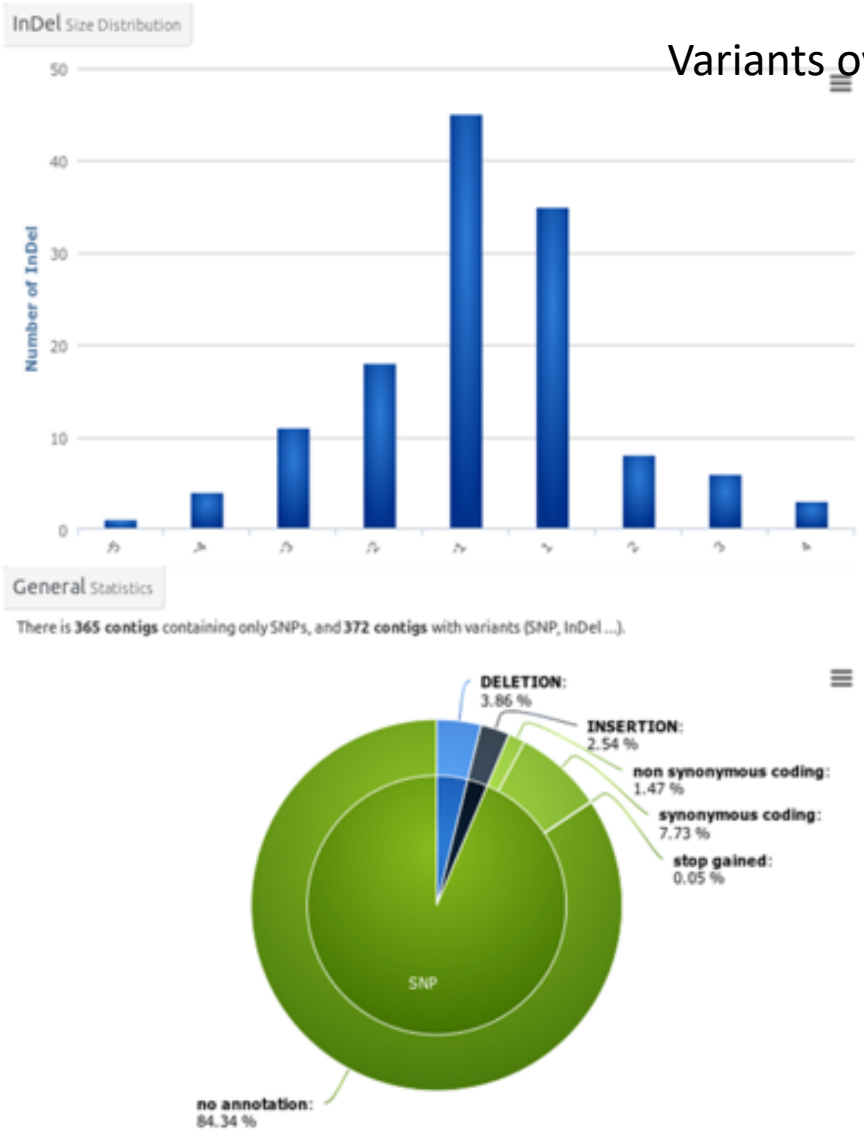


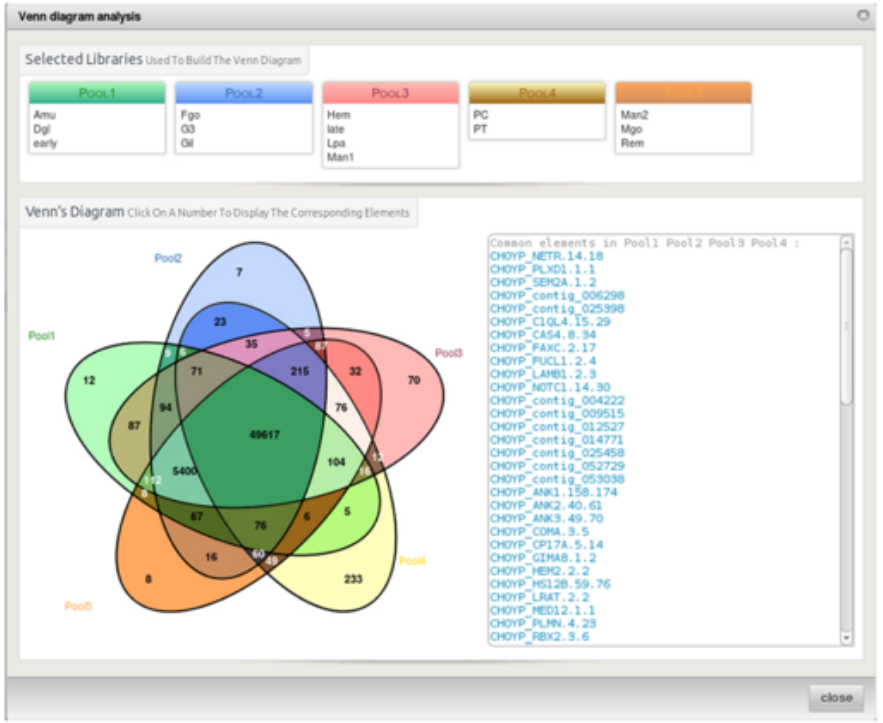


Contigs overview figures

RNAbrowse

Variants overview figures





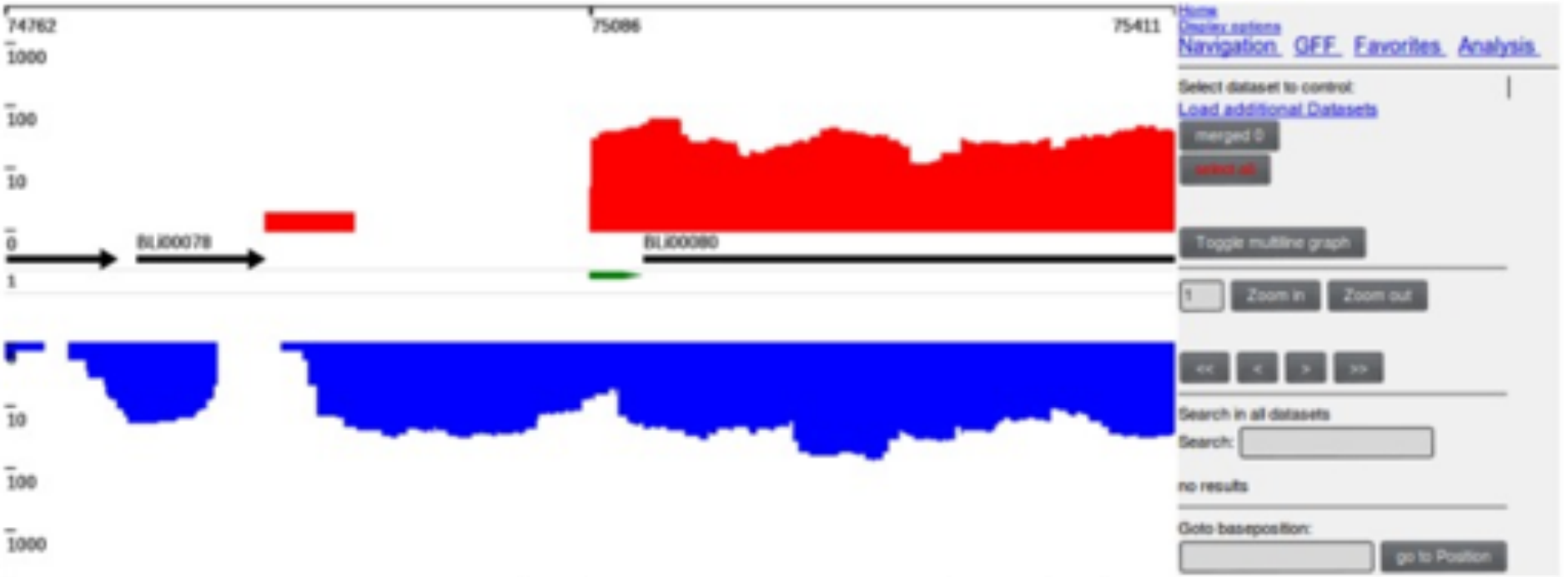
The Venn diagram shows the number of contigs shared between libraries

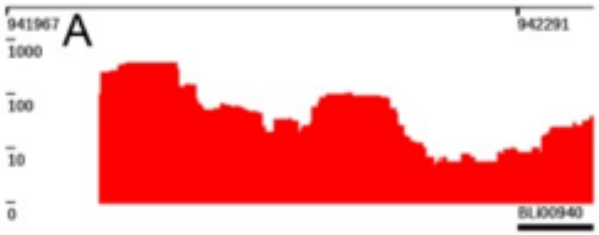
RNAseqViewer.

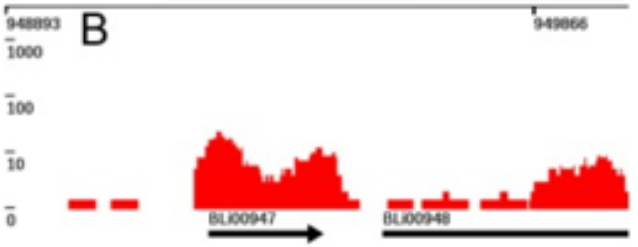
Rogé X , and Zhang X Bioinformatics 2014;30:891-892

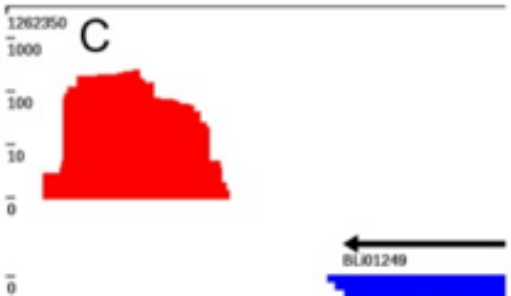
CNRS UPMC
Station Biologique
Roscoff

TraV





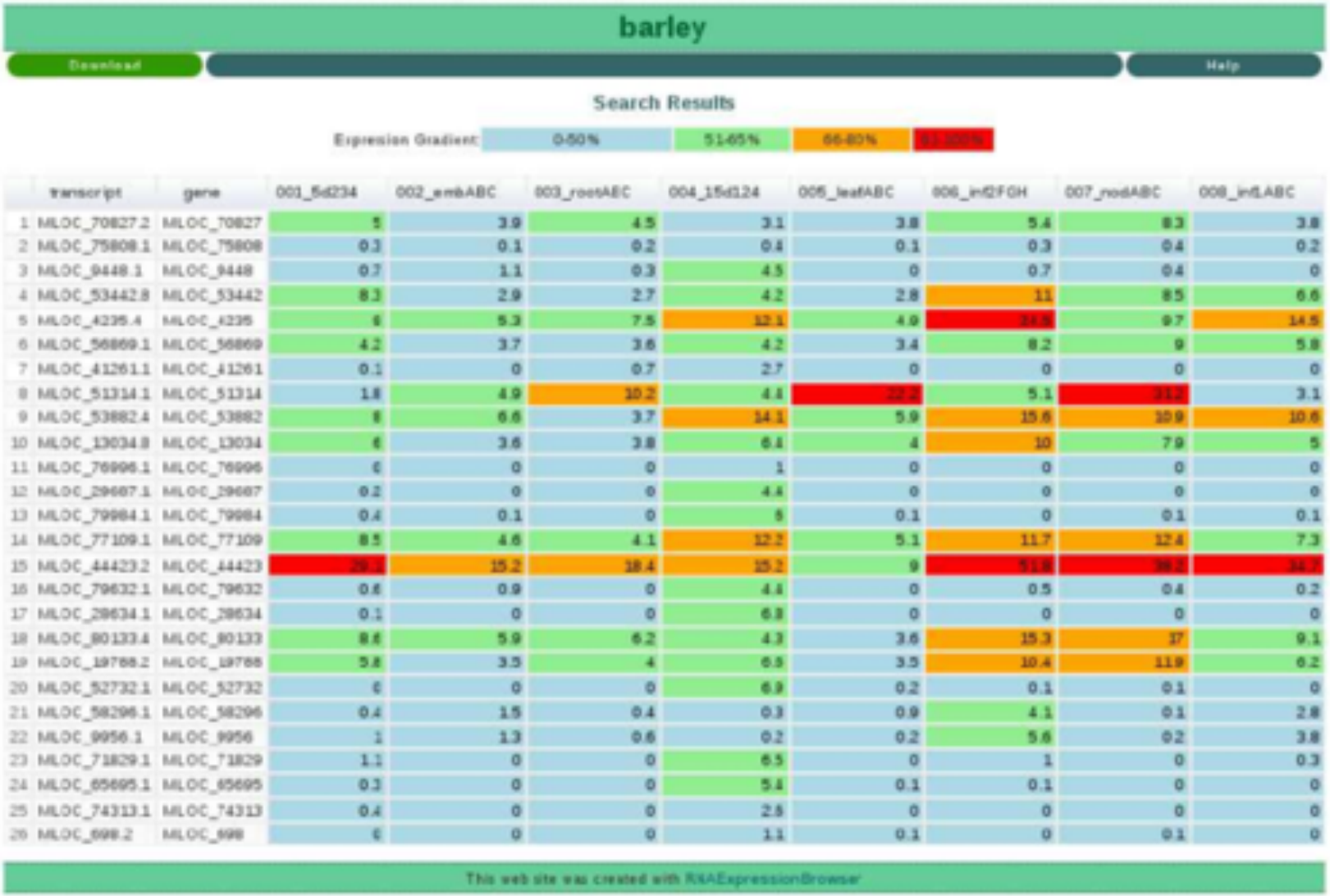




Dietrich S, Wiegand S, Liesegang H (2014) TraV: A Genome Context Sensitive Transcriptome Browser. PLoS ONE 9(4)

RNASeqExpressionBrowser

Nussbaumer, T., Kugler, K. G., Bader, K. C., Sharma, S., Seidel, M., & Mayer, K. F. X. (2014). RNASeqExpressionBrowser - A web interface to browse and visualize high-throughput expression data. Bioinformatics.doi:10.1093/bioinformatics/btu334



barley

Download Help

Search Results

Expression Gradient: 0-50% 51-65% 66-80% 81-100%

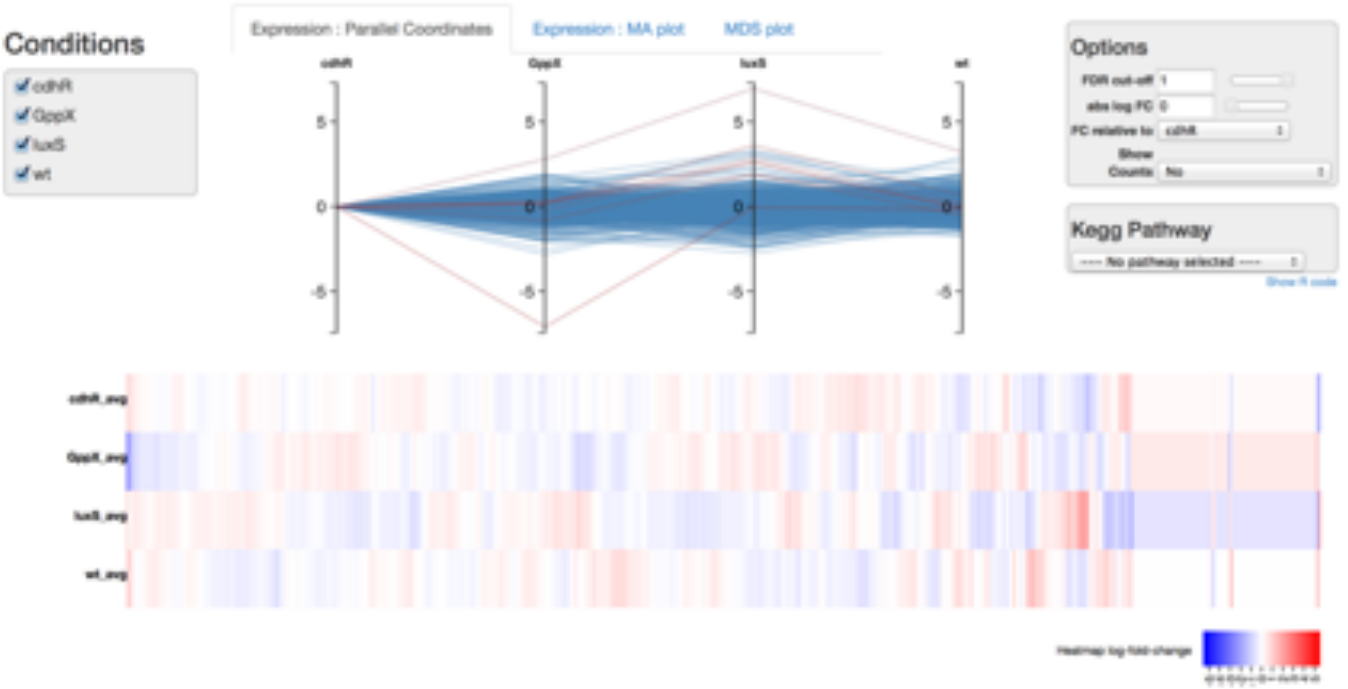
	transcript	gene	001_5d234	002_embABC	003_rootABC	004_15d124	005_leafABC	006_inf2FGH	007_rootABC	008_inf1ABC
1	MLOC_70827.2	MLOC_70827	5	3.9	4.5	3.1	3.8	5.4	8.3	3.8
2	MLOC_75808.1	MLOC_75808	0.3	0.1	0.2	0.4	0.1	0.3	0.4	0.2
3	MLOC_9448.1	MLOC_9448	0.7	1.1	0.3	4.5	0	0.7	0.4	0
4	MLOC_53442.8	MLOC_53442	8.3	2.9	2.7	4.2	2.8	11	8.5	6.6
5	MLOC_4235.4	MLOC_4235	6	5.3	7.5	12.1	4.9	34.5	9.7	14.5
6	MLOC_56869.1	MLOC_56869	4.2	3.7	3.6	4.2	3.4	8.2	9	5.8
7	MLOC_41261.1	MLOC_41261	0.1	0	0.7	2.7	0	0	0	0
8	MLOC_51314.1	MLOC_51314	1.8	4.9	10.2	4.4	22.2	5.1	31.2	3.1
9	MLOC_53882.4	MLOC_53882	8	6.6	3.7	14.1	5.9	15.6	10.9	10.6
10	MLOC_13034.8	MLOC_13034	6	3.6	3.8	6.4	4	10	7.9	5
11	MLOC_76996.1	MLOC_76996	6	0	0	1	0	0	0	0
12	MLOC_29667.1	MLOC_29667	0.2	0	0	4.4	0	0	0	0
13	MLOC_79984.1	MLOC_79984	0.4	0.1	0	5	0.1	0	0.1	0.1
14	MLOC_77109.1	MLOC_77109	8.5	4.6	4.1	12.2	5.1	11.7	12.4	7.3
15	MLOC_44423.2	MLOC_44423	29.1	15.2	18.4	15.2	9	51.8	38.2	34.7
16	MLOC_79632.1	MLOC_79632	0.8	0.9	0	4.4	0	0.5	0.4	0.2
17	MLOC_28634.1	MLOC_28634	0.1	0	0	6.8	0	0	0	0
18	MLOC_80133.4	MLOC_80133	8.6	5.9	6.2	4.3	3.6	15.3	17	9.1
19	MLOC_19768.2	MLOC_19768	5.8	3.5	4	6.5	3.5	10.4	11.9	6.2
20	MLOC_52732.1	MLOC_52732	6	0	0	6.9	0.2	0.1	0.1	0
21	MLOC_58296.1	MLOC_58296	0.4	1.5	0.4	0.3	0.9	4.1	0.1	2.8
22	MLOC_9956.1	MLOC_9956	1	1.3	0.6	0.2	0.2	5.6	0.2	3.8
23	MLOC_71829.1	MLOC_71829	1.1	0	0	6.5	0	1	0	0.3
24	MLOC_65695.1	MLOC_65695	0.3	0	0	5.4	0.1	0.1	0	0
25	MLOC_74313.1	MLOC_74313	0.4	0	0	2.5	0	0	0	0
26	MLOC_698.2	MLOC_698	6	0	0	1.1	0.1	0	0.1	0

This web site was created with RNAExpressionBrowser

CNRS UPMC
Station Biologique
Roscoff

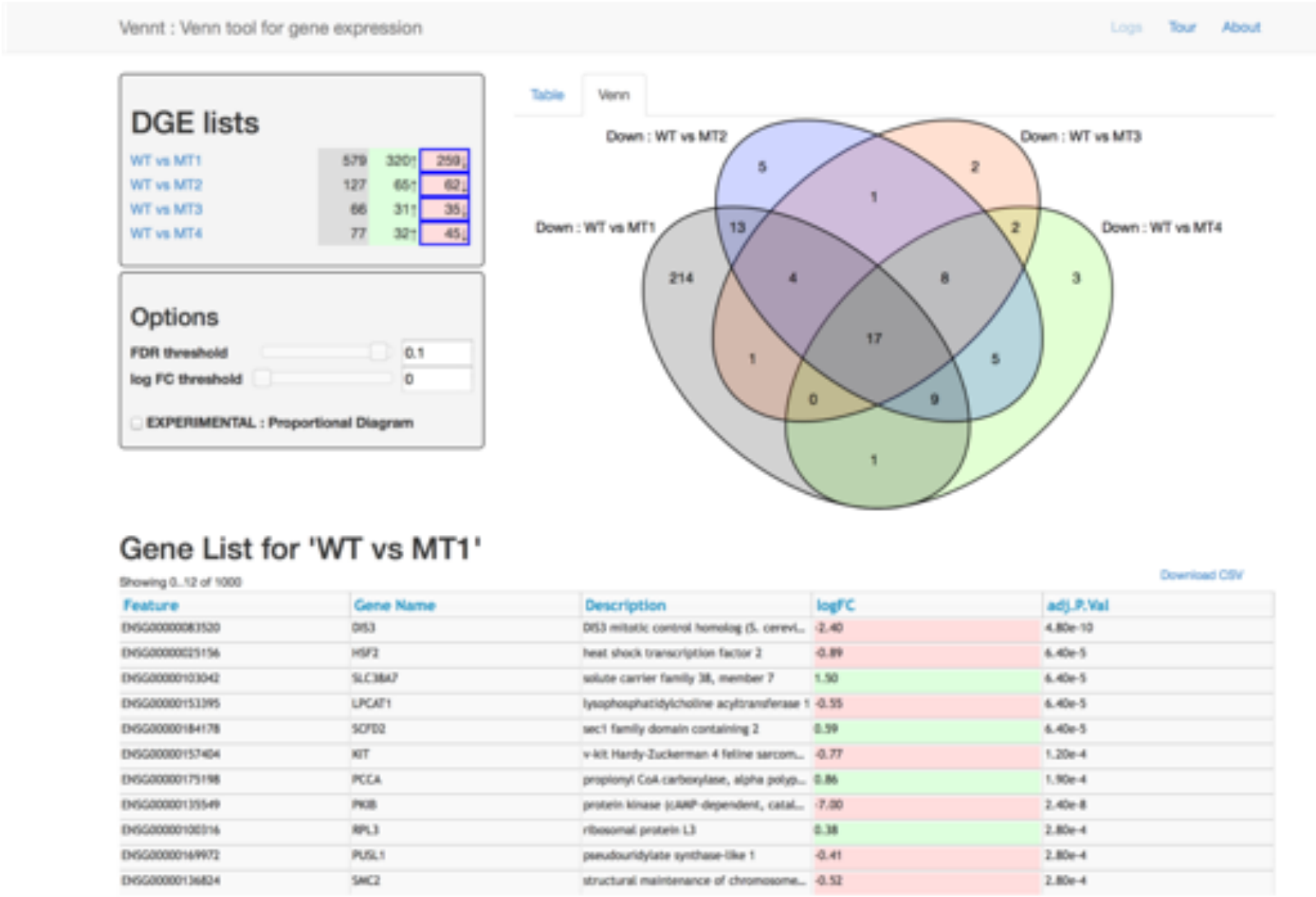
Degust (formerly DGE-Vis)

An interactive web tool for visualising Differential Gene Expression data
http://victorian-bioinformatics-consortium.github.io/degust/



Vennt : Dynamic Venn diagrams for Differential Gene Expression

A web-tool to generate dynamic Venn diagrams for differential gene expression.
http://drpowell.github.io/vennt/



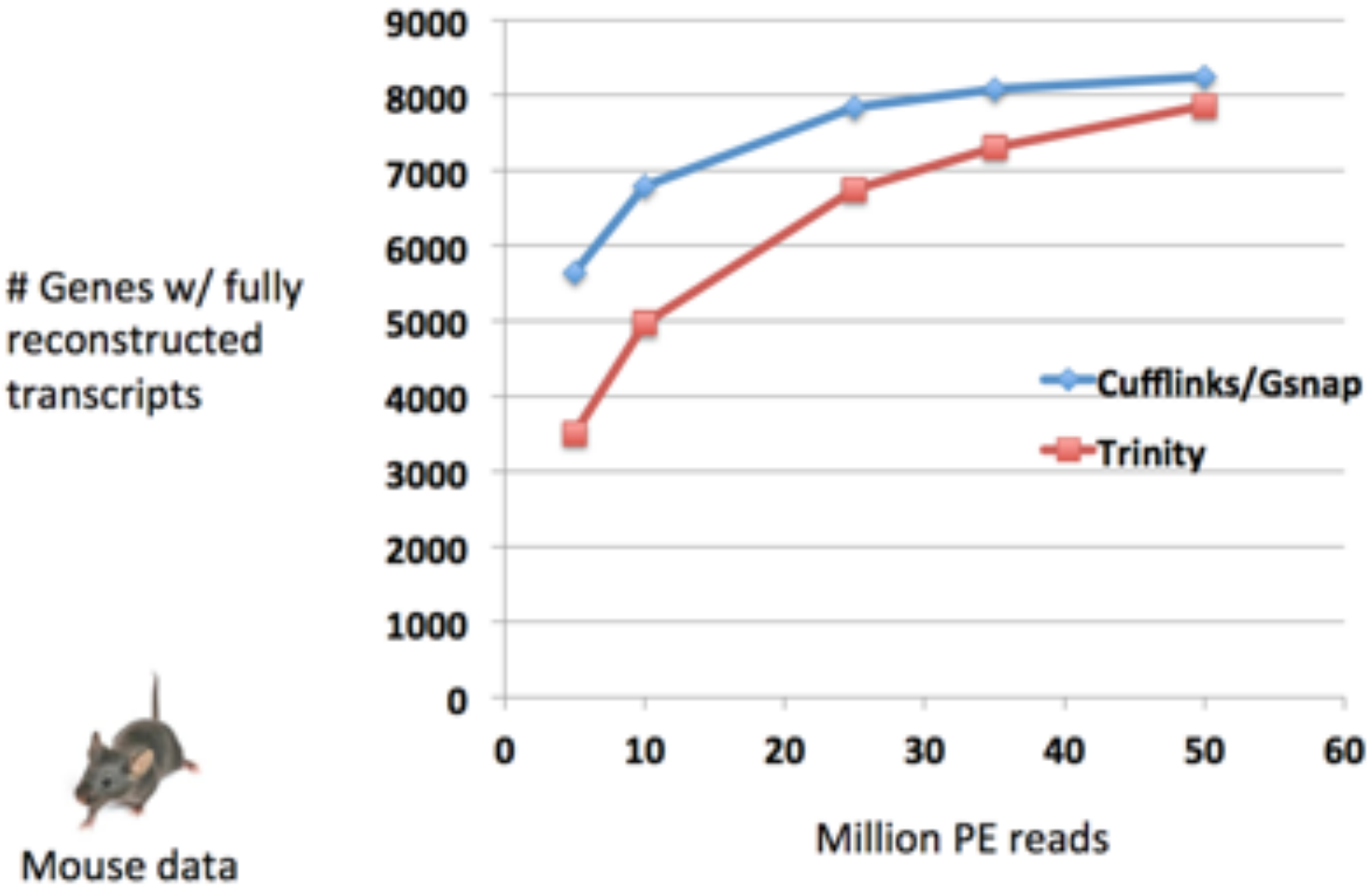
shinyheatmap

http://shinyheatmap.com/

Khomtchouk BB, Hennessy JR, Wahlestedt C. (2017) **shinyheatmap: Ultra fast low memory heatmap web interface for big data genomics**. *PLoS One* 12(5):e0176334.

With ref. vs de novo

Improved reconstruction with deeper sequencing depth and Genome-based reconstruction is more sensitive than de novo methods



The PASA Pipeline for Genome Annotation

PASA: Program to Assemble Spliced Alignments

5654–5666 Nucleic Acids Research, 2003, Vol. 31, No. 19
DOI: 10.1093/nar/gkg770

Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies

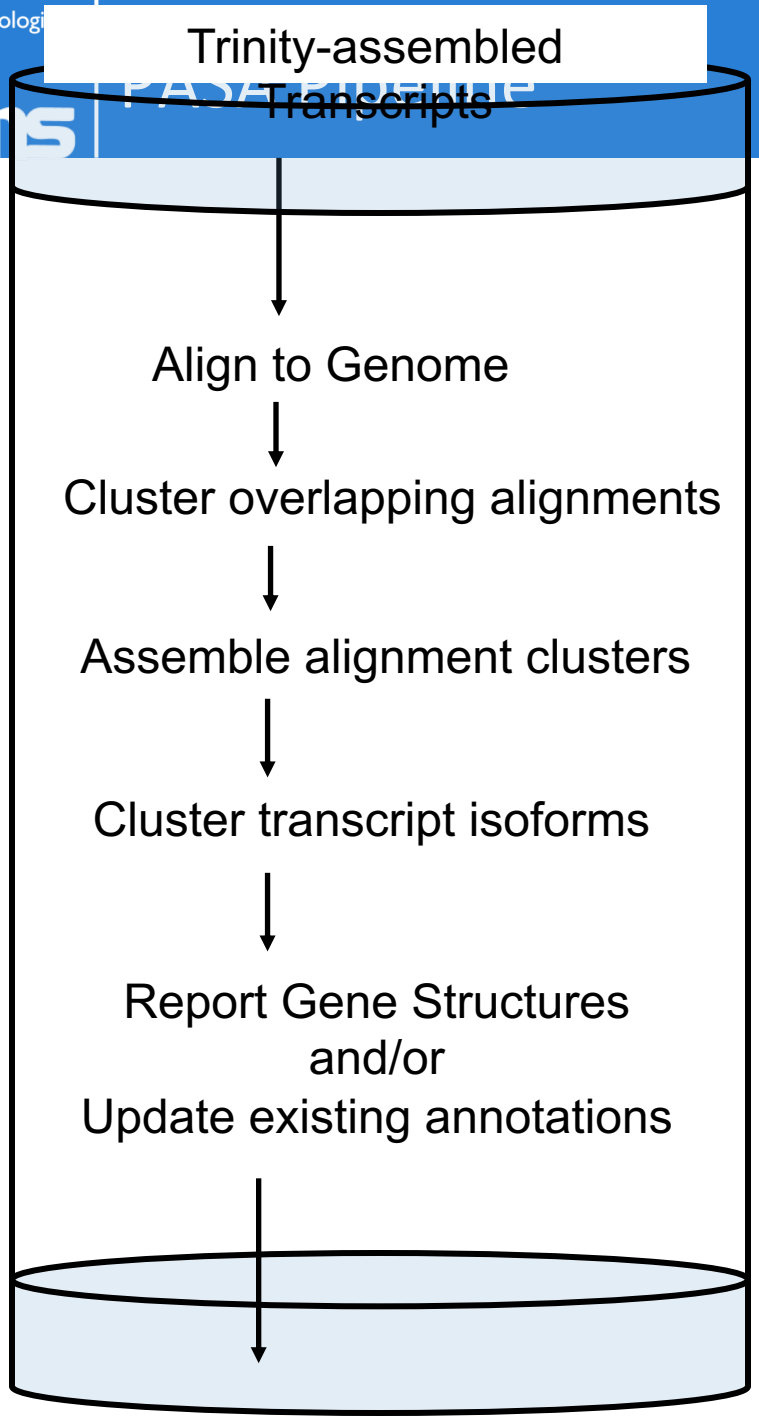
Brian J. Haas*, Arthur L. Delcher, Stephen M. Mount[1], Jennifer R. Wortman, Roger K. Smith Jr, Linda I. Hannick, Rama Maiti, Catherine M. Ronning, Douglas B. Rusch[2], Christopher D. Town, Steven L. Salzberg and Owen White

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, [1]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA and [2]The Center for Advancement of Genomics, 1901 Research Boulevard, Rockville, MD 20850, USA

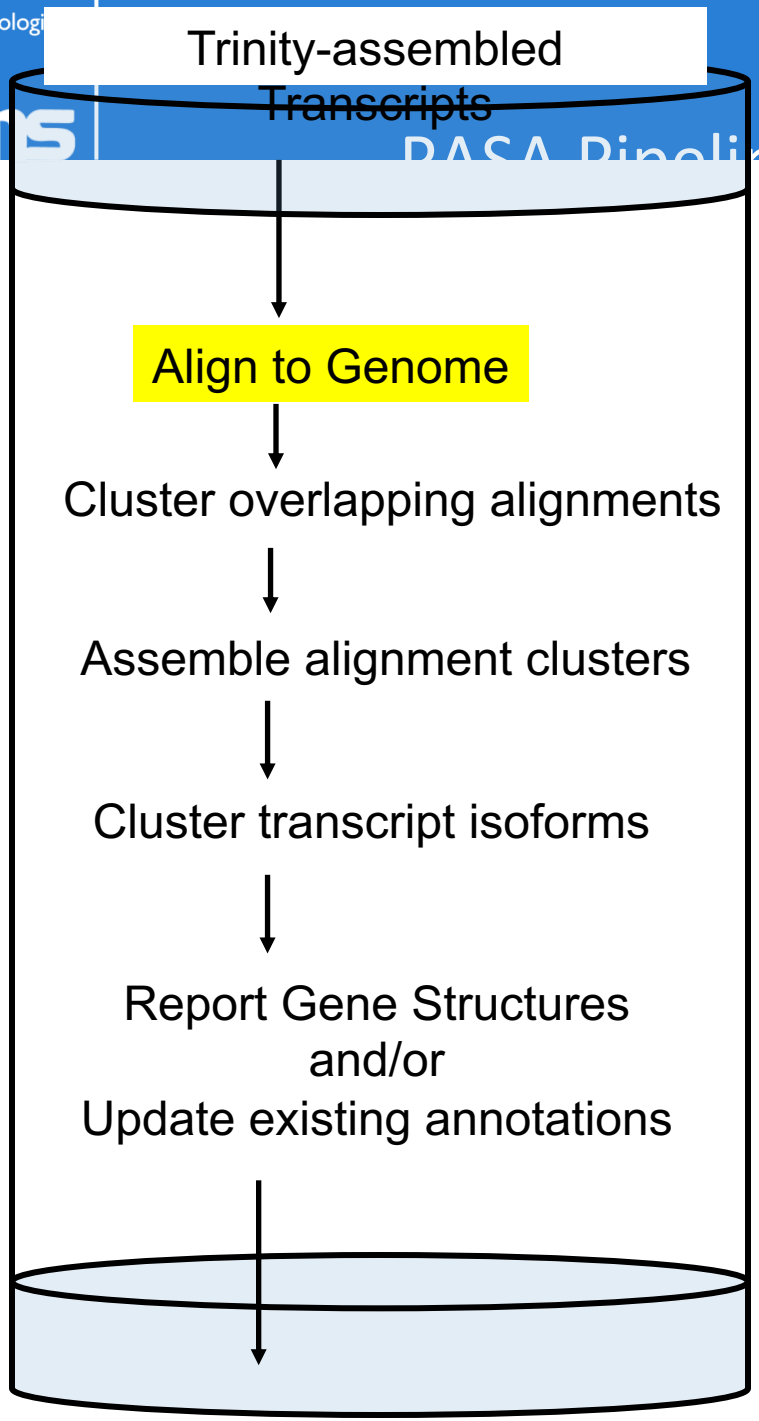
TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Developed (in 2003) to integrate ESTs and full-length cDNAs into gene structure annotations.

Compatible with RNA-Seq via Trinity.


Trinity-assembled
Transcripts
Align to Genome
Cluster overlapping alignments
Assemble alignment clusters
Cluster transcript isoforms
Report Gene Structures
and/or
Update existing annotations

PASA Pipeline



GMAP, BLAT, sim4
spliced transcript alignments

Valid alignment criteria:

- min *95%* Identity
  min *75%* transcript length aligned
  (configurable)
- Canonical splice sites
  - **GT-AG**
  - **GC-AG**
  - **AT-AC**

PASA Pipeline

Trinity-assembled Transcripts

Align to Genome

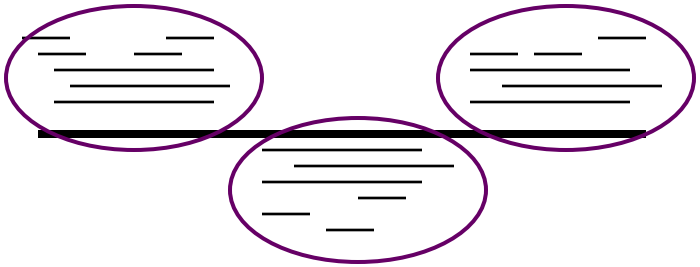
Cluster overlapping alignments

Assemble alignment clusters

Cluster transcript isoforms

Report Gene Structures
and/or
Update existing annotations

spliced alignments



Trinity-assembled Transcripts

Align to Genome

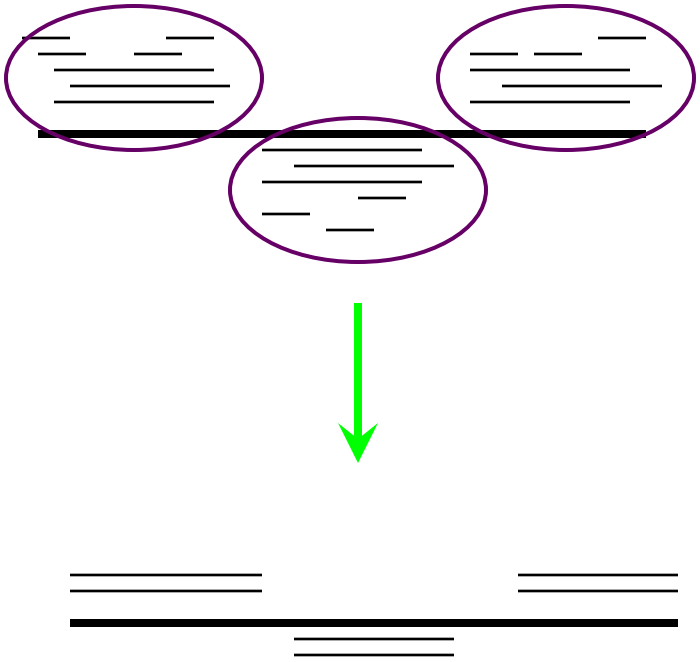
Cluster overlapping alignments

Assemble alignment clusters

Cluster transcript isoforms

Report Gene Structures
and/or
Update existing annotations

spliced alignments



PASA Pipeline

Trinity-assembled Transcripts

Align to Genome

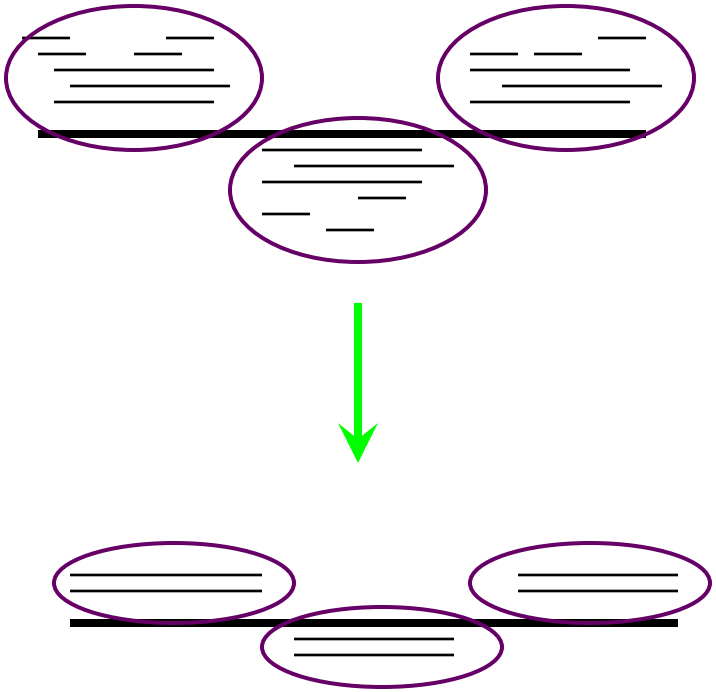
Cluster overlapping alignments

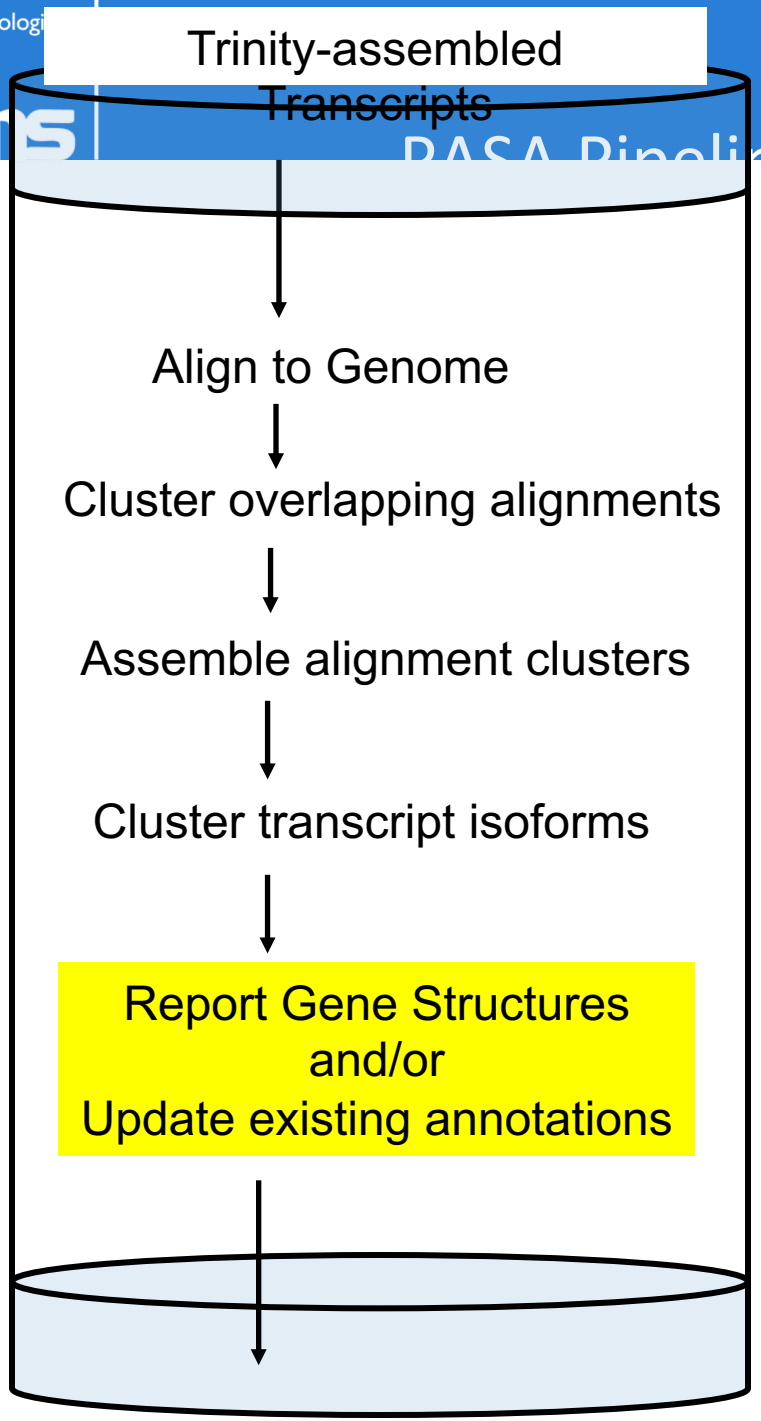
Assemble alignment clusters

Cluster transcript isoforms

Report Gene Structures
and/or
Update existing annotations

spliced alignments





Annotation output

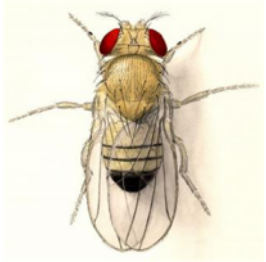
-gene structures
-alt splice isoforms
-predicted coding regions

(fasta, bed, gff3, gtf formats)

Annotation Updates

-exon modifications
-alt splice isoform additions
-gene merges
-gene splits
-new genes

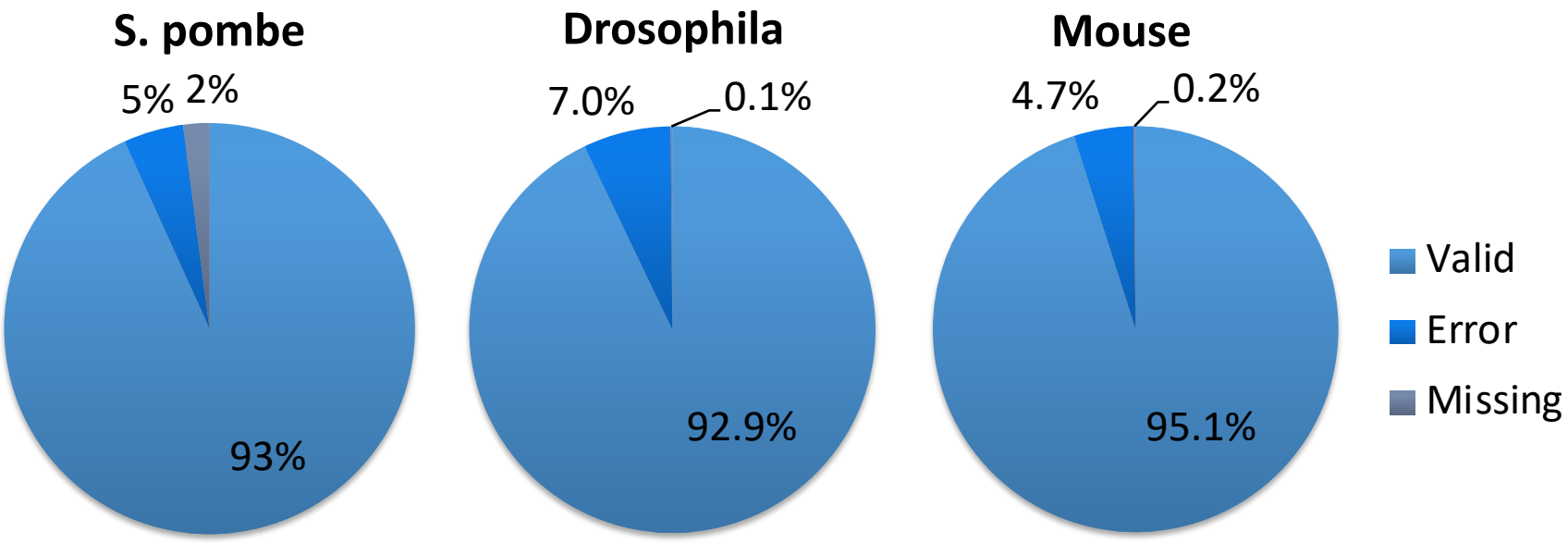
Evaluating Genome-based Transcript Reconstruction Using Reference Genomes + Transcriptomes

	Schizosaccharomyces pombe	Drosophila	Mouse
			
Genome size	12.5 Mb	170 Mb	2.7 Gb
Approx. # genes	5k	14k	20k

50M Paired-end Illumina ~75 base reads, each.

(100M total reads, each).

Nearly all (>98%) Trinity transcripts map to reference genomes



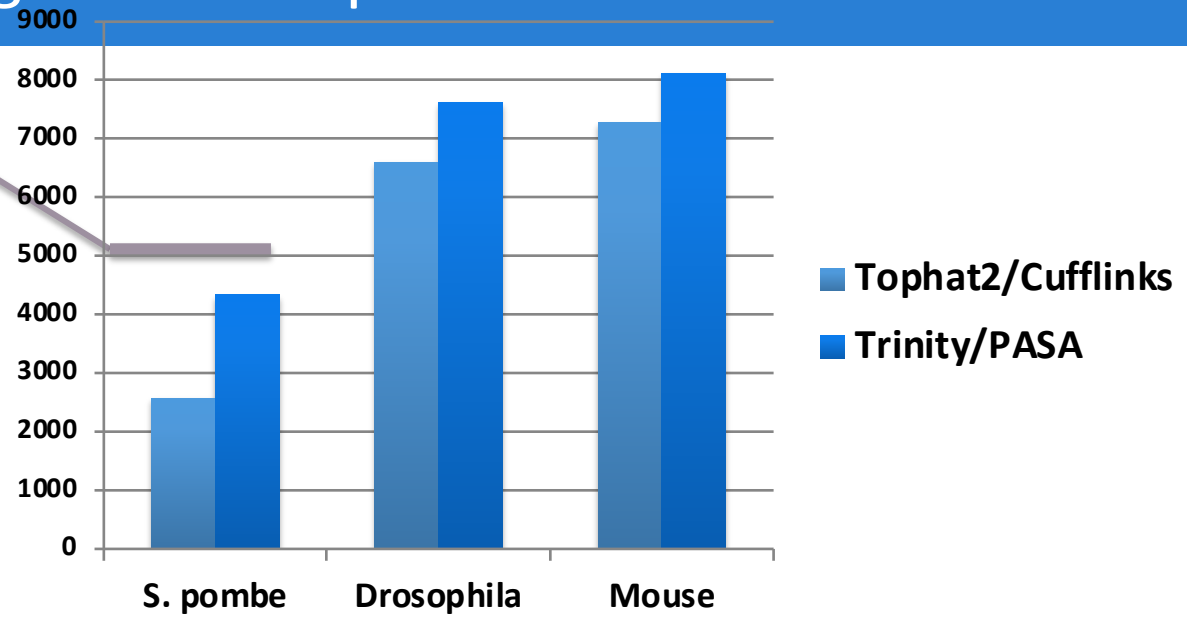
# Trinity Transcripts	14,548	36,320	81,516

~5% to 7% of assembled transcripts are problematic

Full-length Transcript Reconstruction from RNA-Seq

Total pombe genes

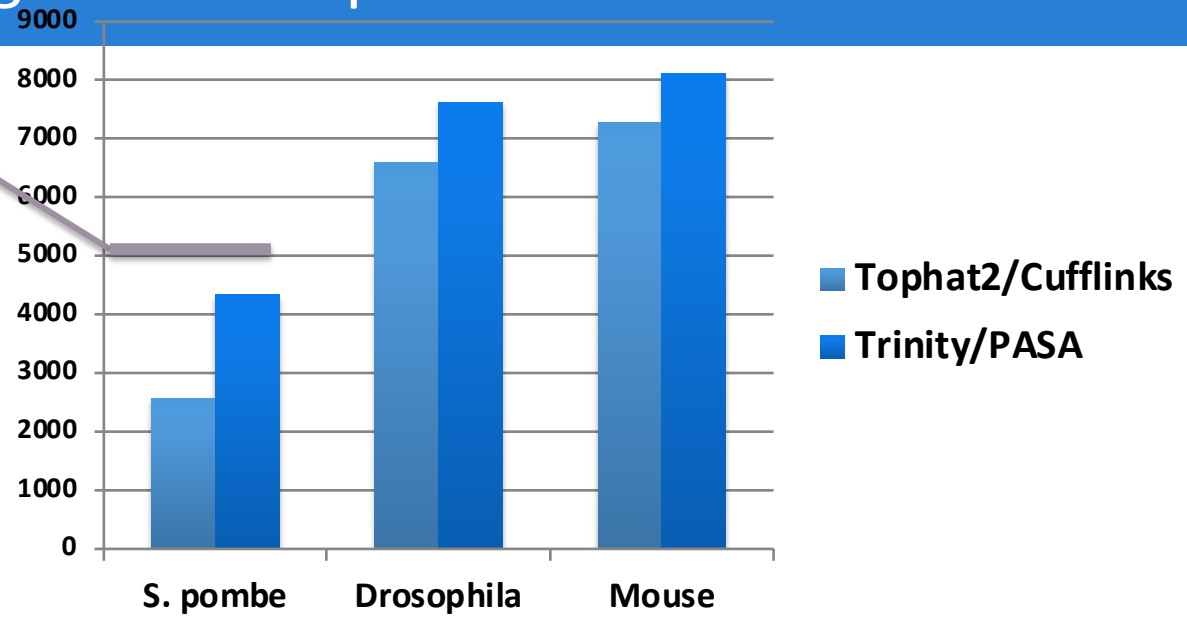
Number of genes with full length transcripts



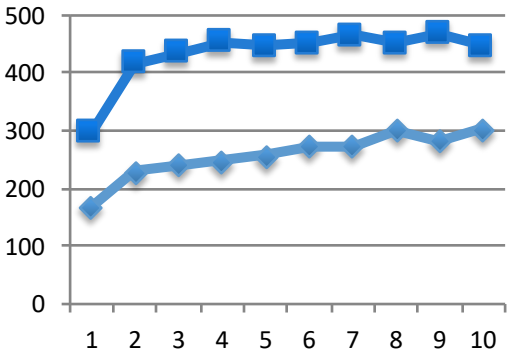
Full-length Transcript Reconstruction from RNA-Seq

Total pombe genes

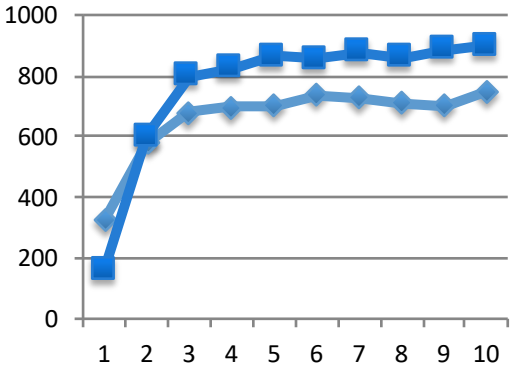
Number of genes with full length transcripts



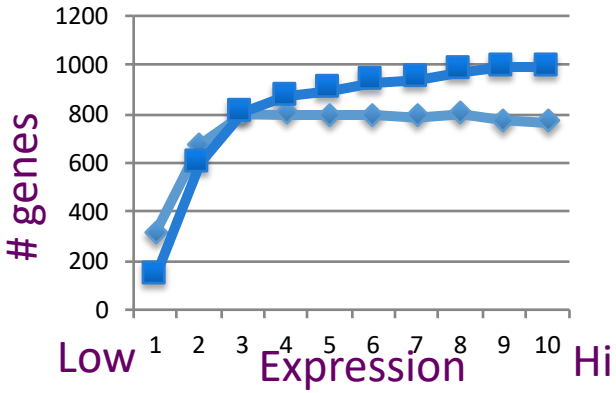
Schizosaccharomyces pombe



Drosophila



Mouse



Full-length Reconstruction by Expression Quintile