IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

IBiSA
INFRASTRUCTURES BIOLOGIE SANTÉ ET AGRONOMIE

EMBRC
EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE

OCEANOMICS

IDEALG
seaweed for the future

Biogenouest

ABiMS

12/06/2014

RNA Seq analysis

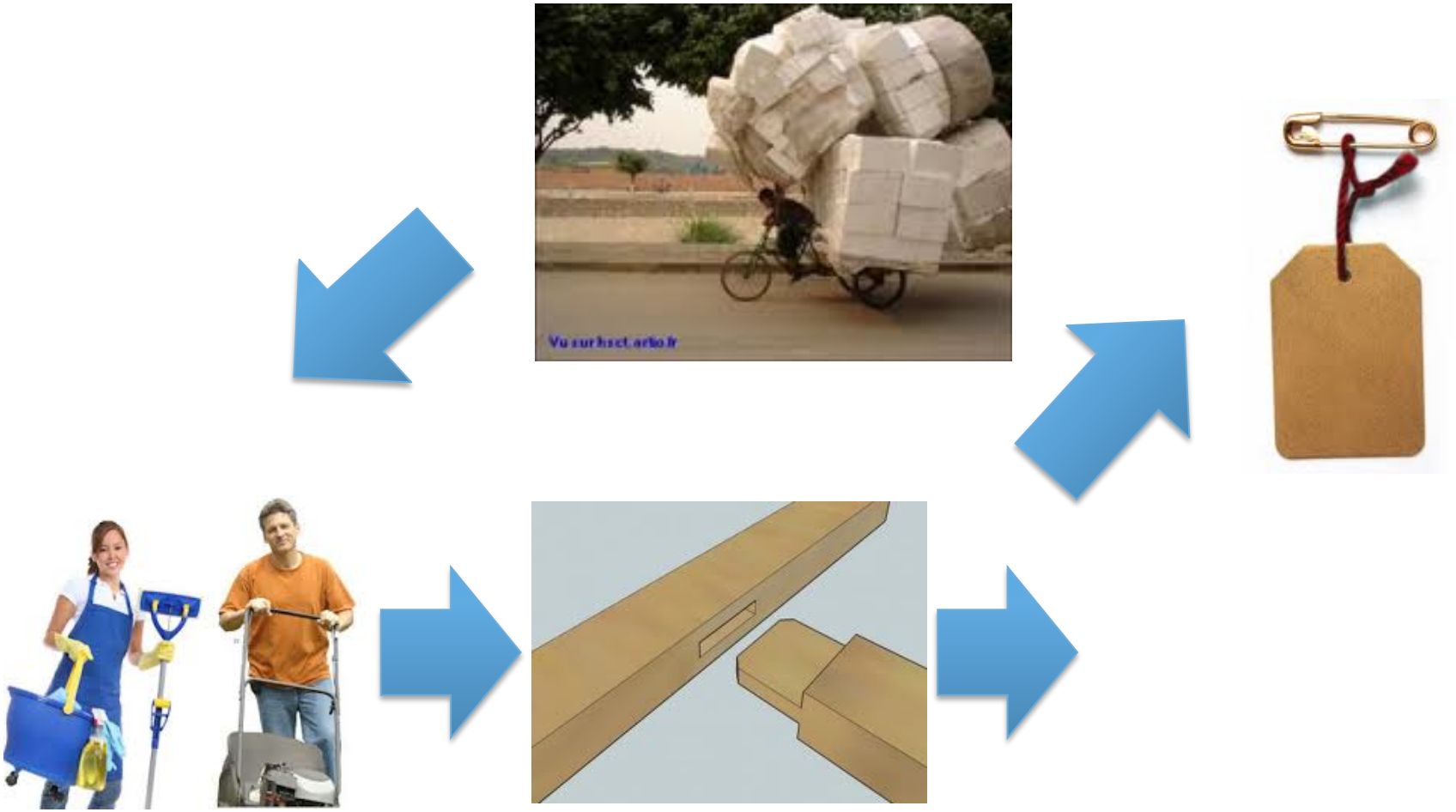
Transcriptome annotation

Plateforme ABiMS

UPMC
SORBONNE UNIVERSITÉS

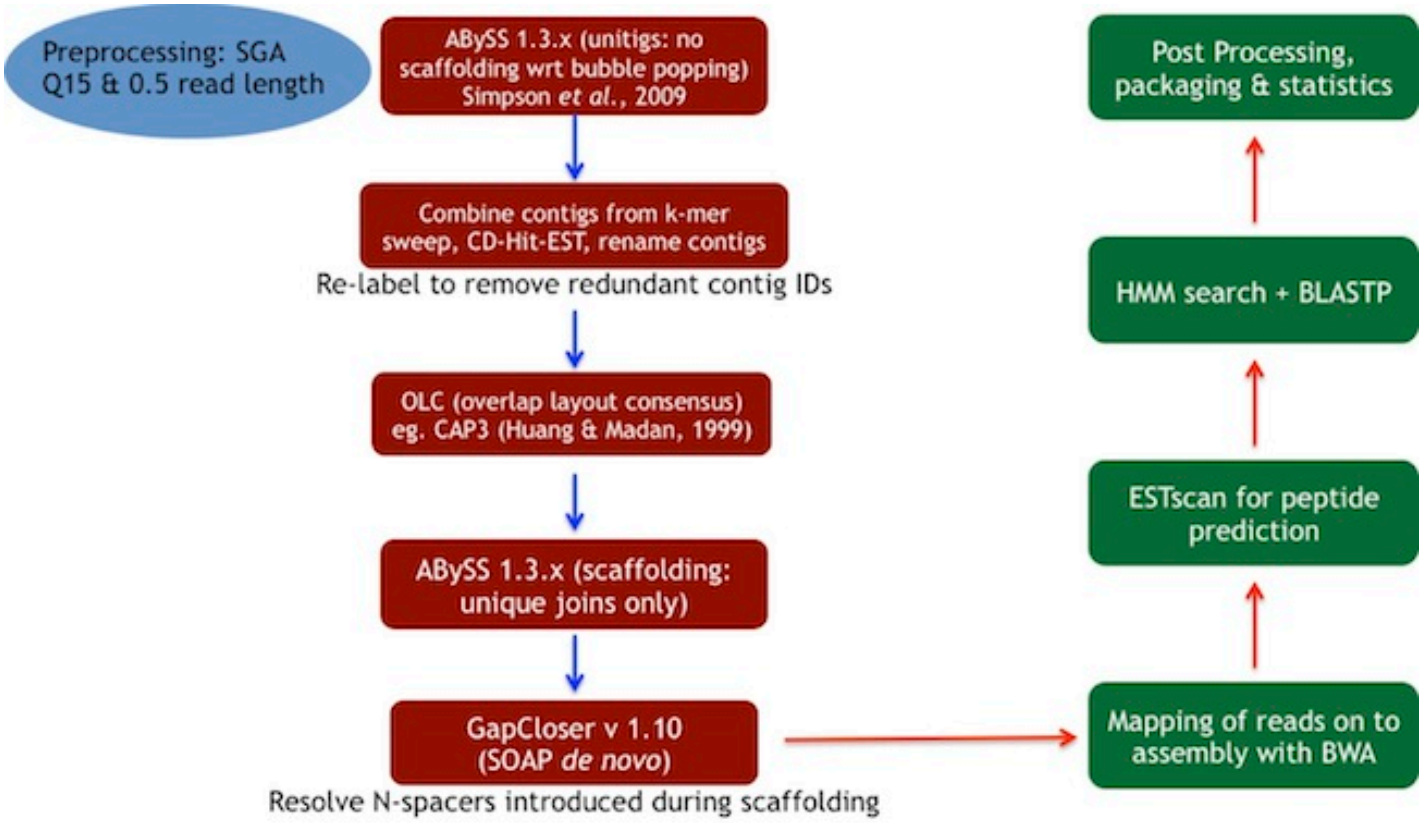
cnrs

RNA Seq analysis



Transcriptome annotation

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

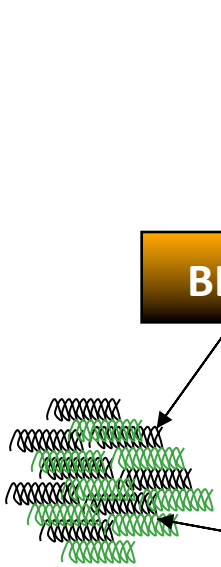


Annotation files based on hits to Swiss-Prot, Pfam-A, and TIGRFAMs include InterPro associations in the Ontology term attribute

Moore fondation pipeline

- **Transcriptome Annotation**
- Coding sequences were predicted using ESTScan [1, 2] with a close organisms scoring matrix.
- Sequence reads were aligned back to the nucleotide motifs of the predicted coding sequences using BWA [3].
- Peptide predictions over 30 amino acids in length were annotated.
- BLASTp [4] was used to generate hits against the UniProtKB/Swiss-Prot database.
- Protein sequences were also functionally characterized using HMMER3 [5] against Pfam-A [6], TIGRFAM [7], and SUPERFAMILY [8] databases.

Main functions within Blast2GO



BLAST

MAPPING

ANNOT.RULE

InterProScan

Annotation
(GO,IPR,EC)

GO Second Layer

GO-Slim

Manual Curation

Validation

Statistics

Graph Visualization

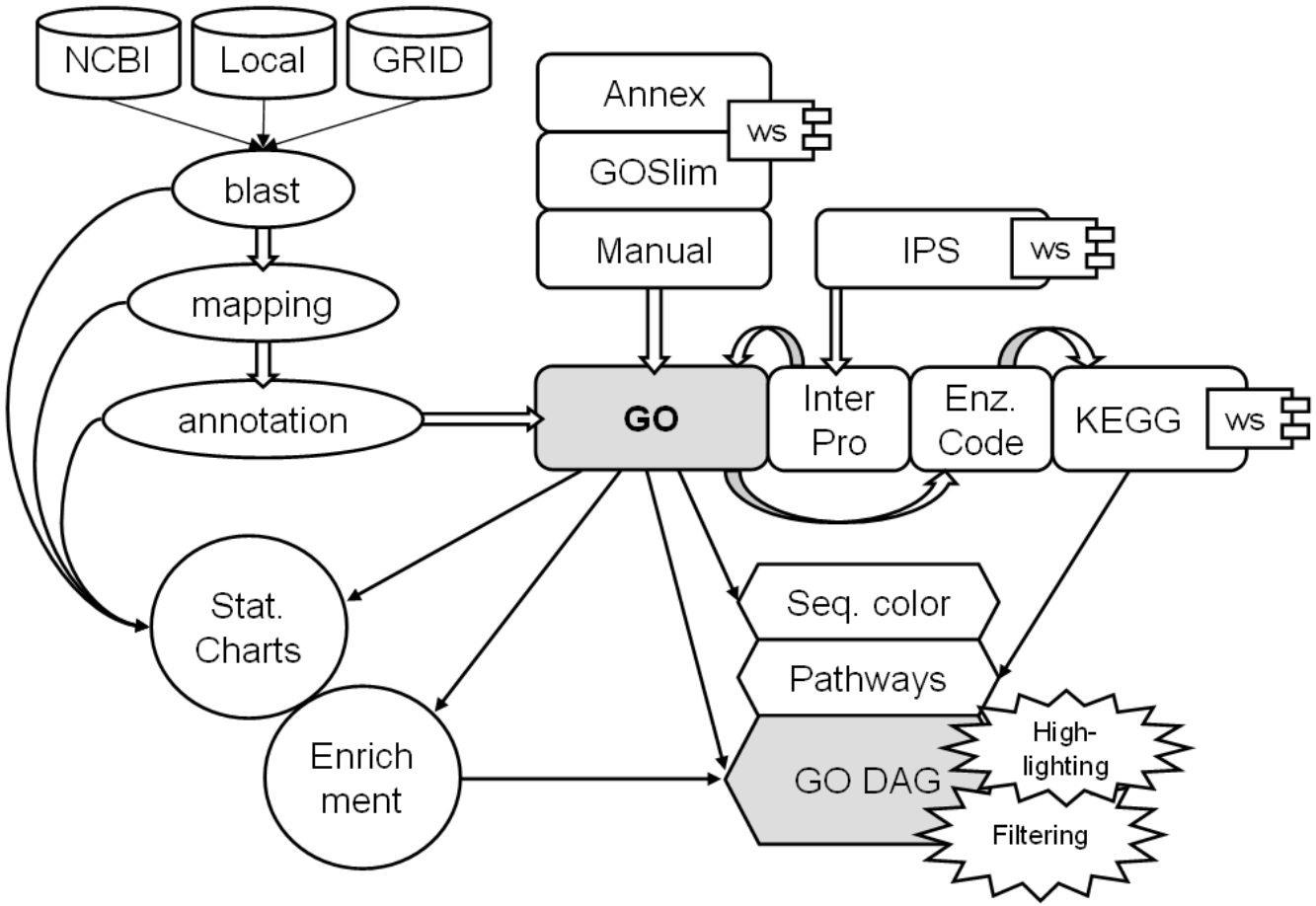
KEGG maps

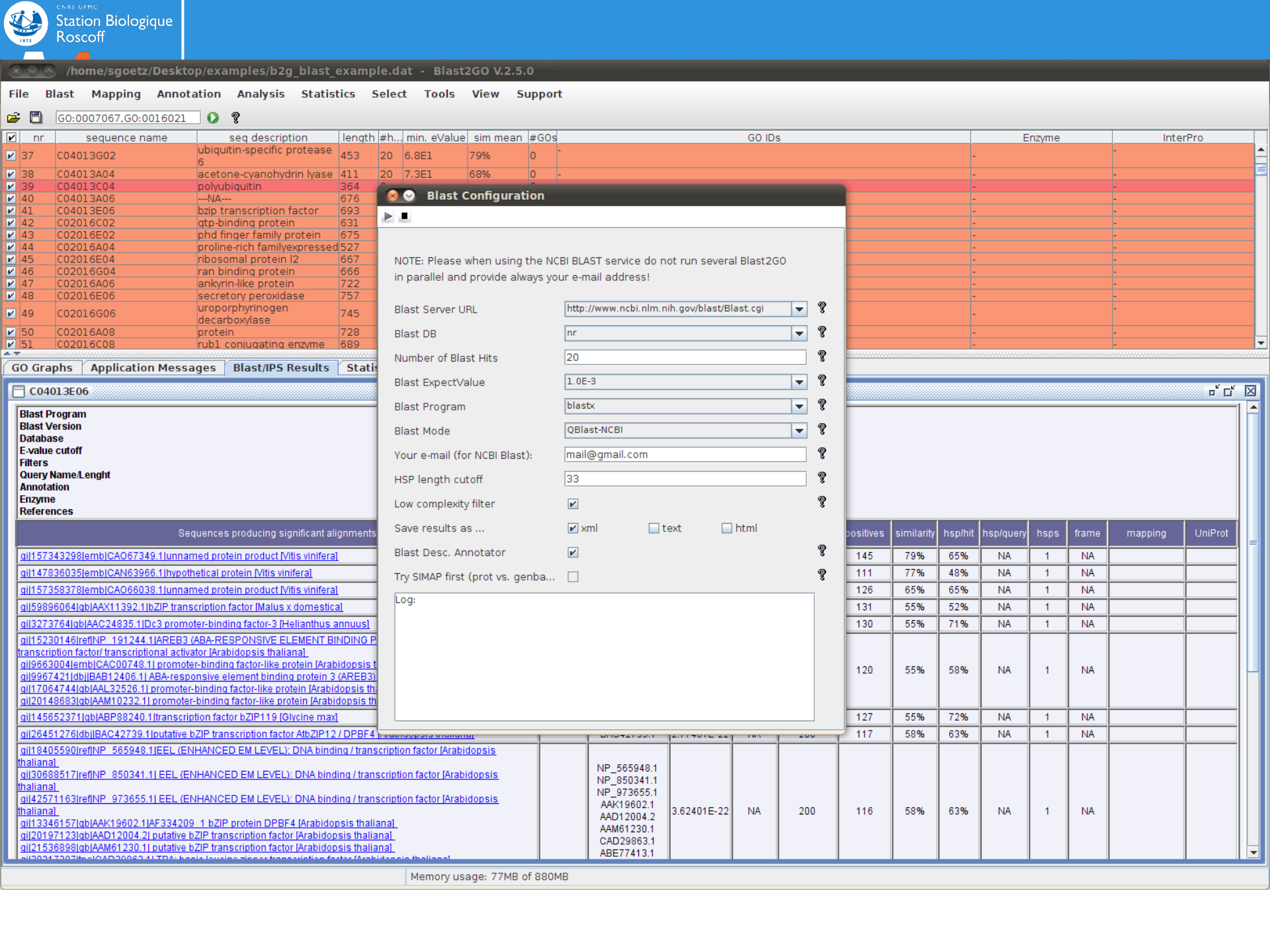
Enrichment

Additional Features:

Batch Mode	Pipeline	costumDB
localB2GDB	Compare	GeneIDs

Blast2GO Schema




CNRS UPMC
Station Biologique
Roscoff
/home/sgoetz/Desktop/examples/b2g_blast_example.dat - Blast2GO V.2.5.0
File Blast Mapping Annotation Analysis Statistics Select Tools View Support
GO:0007067,GO:0016021
nr sequence name seq description length #h... min. eValue sim mean #GOs GO IDs Enzyme InterPro
37 C04013G02 ubiquitin-specific protease 6 453 20 6.8E1 79% 0
38 C04013A04 acetone-cyanohydrin lyase 411 20 7.3E1 68% 0
39 C04013C04 polyubiquitin 364
40 C04013A06 ---NA--- 676
41 C04013E06 bzip transcription factor 693
42 C02016C02 gtp-binding protein 631
43 C02016E02 phd finger family protein 675
44 C02016A04 proline-rich familyexpressed 527
45 C02016E04 ribosomal protein l2 667
46 C02016G04 ran binding protein 666
47 C02016A06 ankyrin-like protein 722
48 C02016E06 secretory peroxidase 757
49 C02016G06 uroporphyrinogen decarboxylase 745
50 C02016A08 protein 728
51 C02016C08 rub1 conjugating enzyme 689
Blast Configuration
NOTE: Please when using the NCBI BLAST service do not run several Blast2GO in parallel and provide always your e-mail address!
Blast Server URL http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Blast DB nr
Number of Blast Hits 20
Blast ExpectValue 1.0E-3
Blast Program blastx
Blast Mode QBlast-NCBI
Your e-mail (for NCBI Blast): mail@gmail.com
HSP length cutoff 33
Low complexity filter
Save results as ... xml text html
Blast Desc. Annotator
Try SIMAP first (prot vs. genba...
Log:
GO Graphs Application Messages Blast/IPS Results
C04013E06
Blast Program
Blast Version
Database
E-value cutoff
Filters
Query Name/Lenght
Annotation
Enzyme
References
Sequences producing significant alignments positives similarity hsp/hit hsp/query hsps frame mapping UniProt
gi|157343298|emb|CAO67349.1|unnamed protein product [Vitis vinifera] 145 79% 65% NA 1 NA
gi|147836035|emb|CAN63966.1|hypothetical protein [Vitis vinifera] 111 77% 48% NA 1 NA
gi|157358378|emb|CAO66038.1|unnamed protein product [Vitis vinifera] 126 65% 65% NA 1 NA
gi|59896064|gb|AAX11392.1|bZIP transcription factor [Malus x domestica] 131 55% 52% NA 1 NA
gi|3273764|gb|AAC24835.1|DC3 promoter-binding factor-3 [Helianthus annuus] 130 55% 71% NA 1 NA
gi|15230146|ref|NP_191244.1|AREB3 (ABA-RESPONSIVE ELEMENT BINDING P transcription factor/ transcriptional activator [Arabidopsis thaliana]
gi|9663004|emb|CAC00748.1| promoter-binding factor-like protein [Arabidopsis t
gi|9967421|dbj|BAB12406.1| ABA-responsive element binding protein 3 (AREB3)
gi|17064744|gb|AAL32526.1| promoter-binding factor-like protein [Arabidopsis th
gi|20148683|gb|AAM10232.1| promoter-binding factor-like protein [Arabidopsis th
120 55% 58% NA 1 NA
gi|145652371|gb|ABP88240.1|transcription factor bZIP119 [Glycine max] 127 55% 72% NA 1 NA
gi|26451276|dbj|BAC42739.1|putative bZIP transcription factor AtbZIP12 / DPBF4 117 58% 63% NA 1 NA
gi|18405590|ref|NP_565948.1|EEL (ENHANCED EM LEVEL); DNA binding / transcription factor [Arabidopsis thaliana]
gi|30688517|ref|NP_850341.1| EEL (ENHANCED EM LEVEL); DNA binding / transcription factor [Arabidopsis thaliana]
gi|42571163|ref|NP_973655.1| EEL (ENHANCED EM LEVEL); DNA binding / transcription factor [Arabidopsis thaliana]
gi|13346157|gb|AAK19602.1|AF334209_1 bZIP protein DPBF4 [Arabidopsis thaliana]
gi|20197123|gb|AAD12004.2| putative bZIP transcription factor [Arabidopsis thaliana]
gi|21536898|gb|AAM61230.1| putative bZIP transcription factor [Arabidopsis thaliana]
NP_565948.1 NP_850341.1 NP_973655.1 AAK19602.1 AAD12004.2 AAM61230.1 CAD29863.1 ABE77413.1
3.62401E-22 NA 200 116 58% 63% NA 1 NA
Memory usage: 77MB of 880MB

CNRS UPMC
Station Biologique
Roscoff

/home/sgoetz/Desktop/examples/mapping.dat - Blast2GO -PRO- V.2.5.0

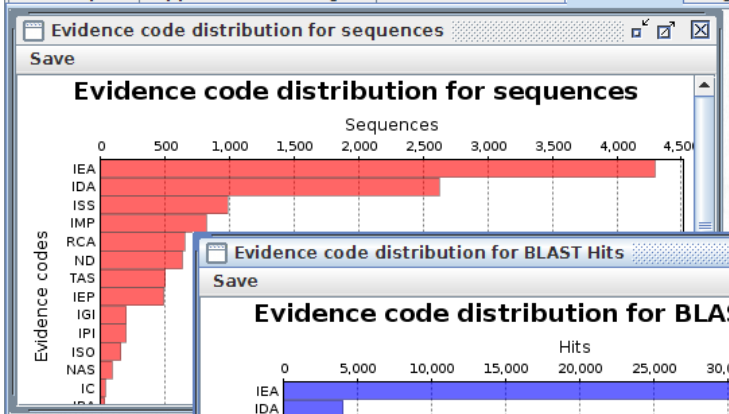
File Blast Mapping Annotation Analysis Statistics Select Tools View Support

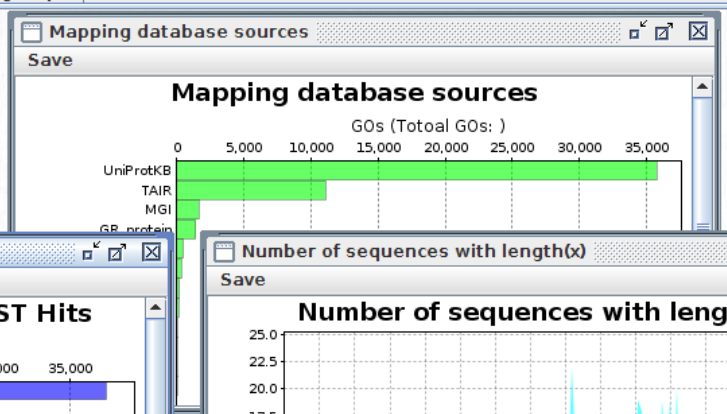
GO:0007067,GO:0016021 | transport;binding;apoptosis | SPO 2518,DDX18 HUMAN

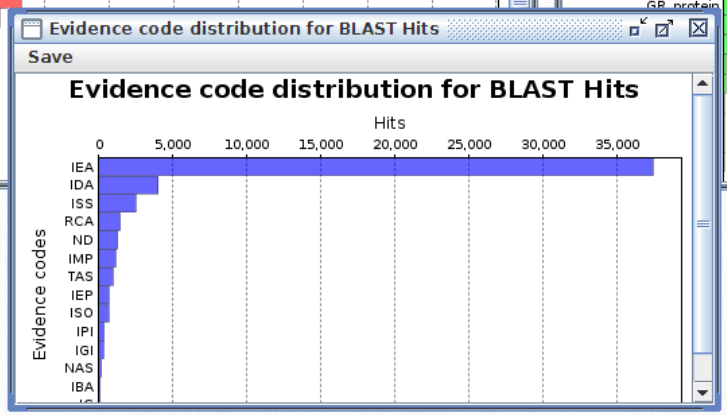
nr	sequence name	seq description	length	#h...	min. eValue	sim mean	#G...	GO IDs	Enzyme	InterPro
19	C04018C08	cinnamoylreductase	462	20	2.0E1	90%	10	F:binding; P:cellular metabolic process; F:catalytic activity; F:coenzyme binding; P:oxidation reduction; F:cinnamoyl-CoA reductase activity; F:oxidoreductase activity; F:3-beta-hydroxy-delta5-steroid dehydrogenase activity; F:oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor; P:steroid biosynthetic process	-	-
20	C04013G06	thaumatin-like protein	669	20	7.6E1	80%	5	P:response to biotic stimulus; P:defense response; P:killing of cells of another organism; P:defense respon	-	-
21	C04013A08	trypsin inhibitor	581	20	7.2E1	55%	5	P:nega F:endo C:apop		
22	C04013C08	j8 protein	600	10	3.0E1	57%	7	F:unfol shock P:resp		
23	C04013E08	peroxidase	677	20	8.4E1	80%	8	F:meta bindin oxidati peroxi		
		nam protein						P:regu binding; P:response to abscisic acid stimulus; P:response to water deprivation; P:multicellular organismal development; P:positive regulation of		

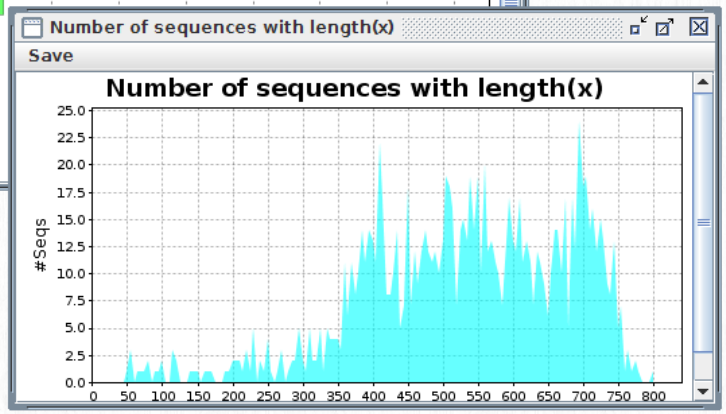
- Show Sequence
- Show Blast Result
- Show InterProScan Result
- Show GO Descriptions
- Load Kegg Pathway Map
- Annotate Seq
- Change Annotation and Description
- Draw Graph of Mapping-Results with highlighted Annotations
- Draw Graph of Annotations

GO Graphs | Application Messages | Blast/IPS Results | Statistics | Kegg Maps









Account: Stefan GÃ¶tz, Status: 32225, Server: DE2-b2g_sep11 | Memory usage: 177MB of 880MB

CNRS UPMC
Station Biologique
Roscoff

/home/sgoetz/Desktop/examples/annotation.dat - Blast2GO -PRO- V.2.5.0

File Blast Mapping Annotation Analysis Statistics Select Tools View Support

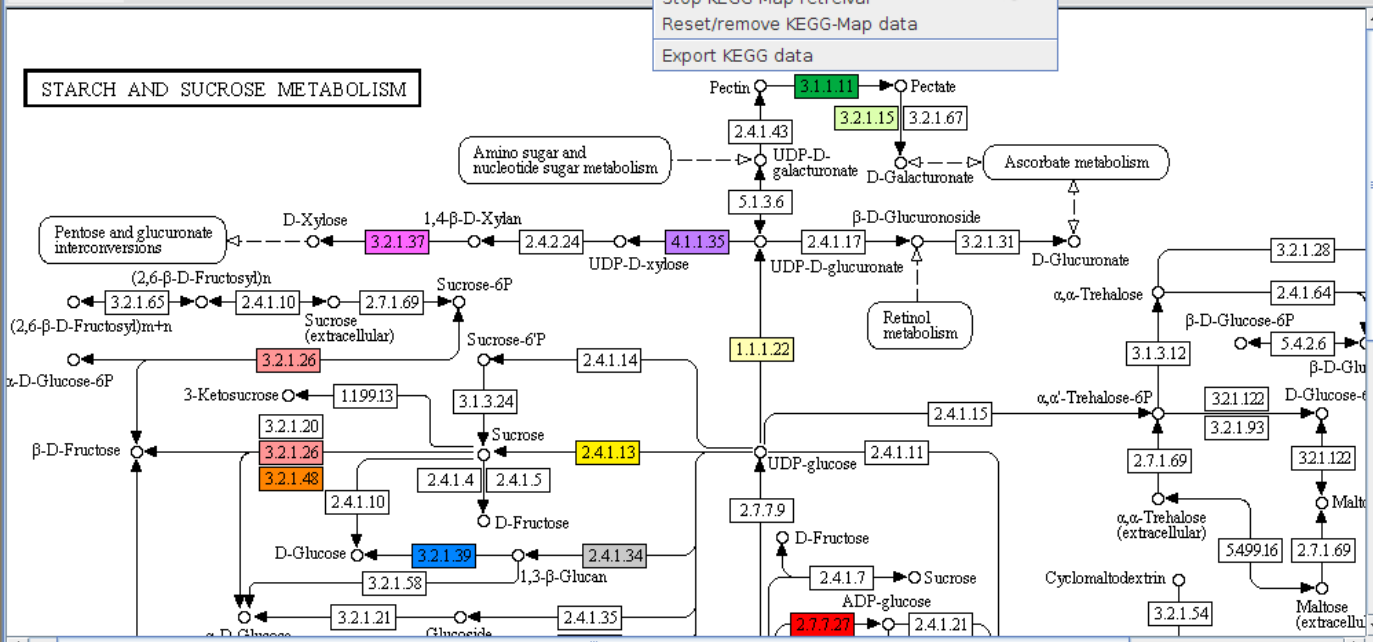
GO:0007067,GO:0 | SPO 2518,DDX18 HUMAN

Annotation menu:
- Run Annotation Step Alt-A
- Set Evidence Code Weights Alt-K
- Reset Annotation
- Validate Annotations
- Remove 1.Level Annotations
- Run ANNEX (Annotation Augmentation) Alt-N
- InterProScan
- Enzyme Code and KEGG
- GO-Slim

Enzyme Code and KEGG menu:
- Run GO-EnzymeCode Mapping Alt-E
- Reset EnzymeCode Mapping
- Load Pathway-Maps from KEGG (online) Alt-H
- Stop KEGG-Map retreival
- Reset/remove KEGG-Map data
- Export KEGG data

	nr	sequence nam	eValue	sim mean	#G...	GO IDs	Enzyme	InterPro
☑	1033	C05014G03	1	83%	3	C:chloroplast inner membrane; C:chloroplast thylakoid membrane; C:mitochondrion	-	no IPS match
☑	1034	C05014G05	1	97%	10	C:cytosol; F:calcium ion binding; P:regulation of photomorphogenesis; P:detection of calcium ion; C:vacuolar membrane; F:protein binding; P:pollen germination; P:protein catabolic process; P:calcium-mediated signaling; C:plasma	-	IPR002048; IPR011992; IPR018247; IPR018248; IPR018249; PTHR23050 (PANTHER), PTHR23050:SF20 (PANTHER), SSF47473

GO Graphs | Application Messages | Blast/IPS Results | Statist



Pathways
Amino sugar and nucleotide sugar metabolism
Phenylalanine metabolism
Starch and sucrose metabolism
Drug metabolism - other enzymes
Tropane, piperidine and pyridine alkaloid biosynthesis
Carbon fixation in photosynthetic organisms
Oxidative phosphorylation
Ascorbate and aldarate metabolism
Pyruvate metabolism
Metabolism of xenobiotics by cytochrome P450
Drug metabolism - cytochrome P450
Fatty acid metabolism
Pentose phosphate pathway
Arginine and proline metabolism
Pentose and glucuronate interconversions
Cysteine and methionine metabolism
Porphyrin and chlorophyll metabolism
Valine, leucine and isoleucine degradation
Aminobenzoate degradation
Galactose metabolism
Glycine, serine and threonine metabolism
Carbon fixation pathways in prokaryotes
alpha-Linolenic acid metabolism
Fructose and mannose metabolism
Tyrosine metabolism
Glycerolipid metabolism
Inositol phosphate metabolism
Lysine degradation

Color	Enzyme	Sequences
red	ec:2.7.7.27 - glucose-1-phosphate adenylyltransferase	C16001F12
yellow	ec:2.4.1.13 - sucrose synthase	C20010B06
orange	ec:3.2.1.48 - sucrose alpha-glucosidase	C04018H06, C18004F02
green	ec:3.1.1.11 - pectinesterase	C02016G10, C04023G02, C03006C02
blue	ec:3.2.1.39 - glucan endo-1,3-beta-D-glucosidase	C04018A08, C04019C06
pink	ec:3.2.1.37 - xylan 1,4-beta-xylosidase	C18004H02
violet	ec:4.1.1.35 - UDP-glucuronate decarboxylase	C02008D08

Account: Stefan GÃ¶tz, Status: 32225, Server: DE2-b2g_sep11 | Memory usage: 399MB of 880MB

/home/sgoetz/Desktop/examples/annotation.dat - Blast2GO -PRO- V.2.5.0

File Blast Mapping Annotation Analysis Statistics Select Tools View Support

GO:0007067,GO:0016021 | transport;binding;apoptosis | SPO 2518,DDX18 HUMAN

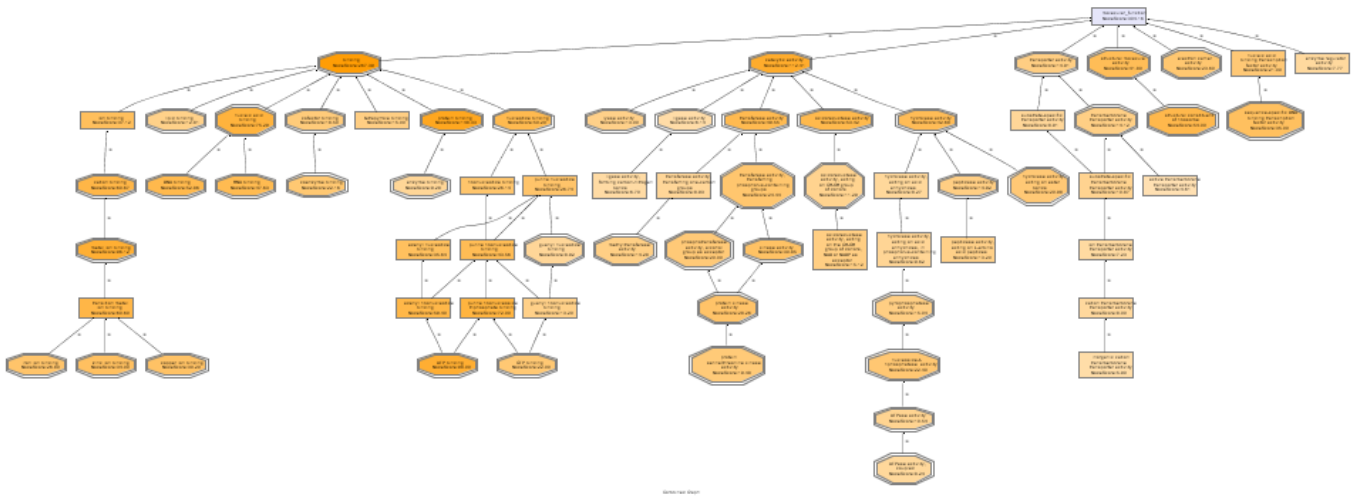
	nr	sequence name	seq description	length	#h...	min. eValue	sim mean	#G...	GO IDs	Enzyme	InterPro
✔	37	C04013G02	ubiquitin-specific protease 6	45					P:ubiquitin-dependent protein catabolic process; F:ubiquitin	EC:3.1.2.15	-
✔	38	C04013A04	acetone-cyanohydrin lyase	41						EC:3.1.1.1; EC:4.1.2....	-
✔	39	C04013C04	polyubiquitin	36						-	-
✔	40	C04013A06	---NA---	67						-	-
			bzip transcription								-

Graph Drawing Configuration

Tree Type	○ Process ◉ Function ○ Compon...
Seq Filter	20
Node Information	NodeScore
Mode of Graph-Colouring	byScore
Score alpha	0.6
Node Score Filter	0
Arrow labels	◉ on ○ off
Graph Title Text	Combined Graph

GO Graphs | Application Messages | Blast/IPS Re

Simple GOs : simpel GOs



Account: Stefan GÃ¶tz, Status: 32225, Server: DE2-b2g_sep11 | Memory usage: 331MB of 880MB

Blast2GO Annotation Rule

Lowest term satisfying the requirements

Similarity requirement

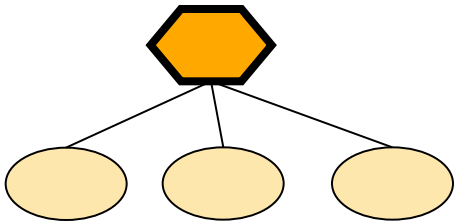
$$sim = \frac{\sum positives_{hsp}}{\sum alignmentlength_{hsp}}$$

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

How to use blast2go

- On your own computer ->

- Via qlogin on ABIMS Cluster ->

- On Command line B2G4Pipe on ABiMS cluster ->

- Soon with Galaxy ->

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

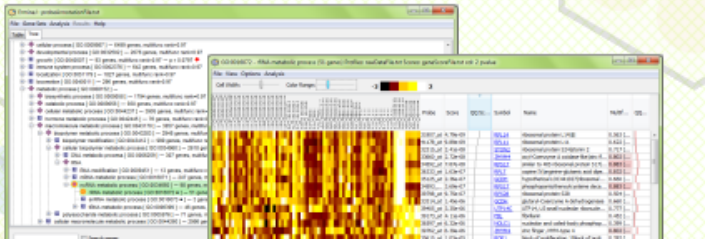
- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes 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.

Ermine J

ErmineJ performs analyses of gene sets in expression microarray data or other genome-wide data that results in rankings of genes

ermineJ

Gene set analysis software



Trinotate

NCBI BLAST

HMMER

Pfam

UniProt

eggNOG version 3.0

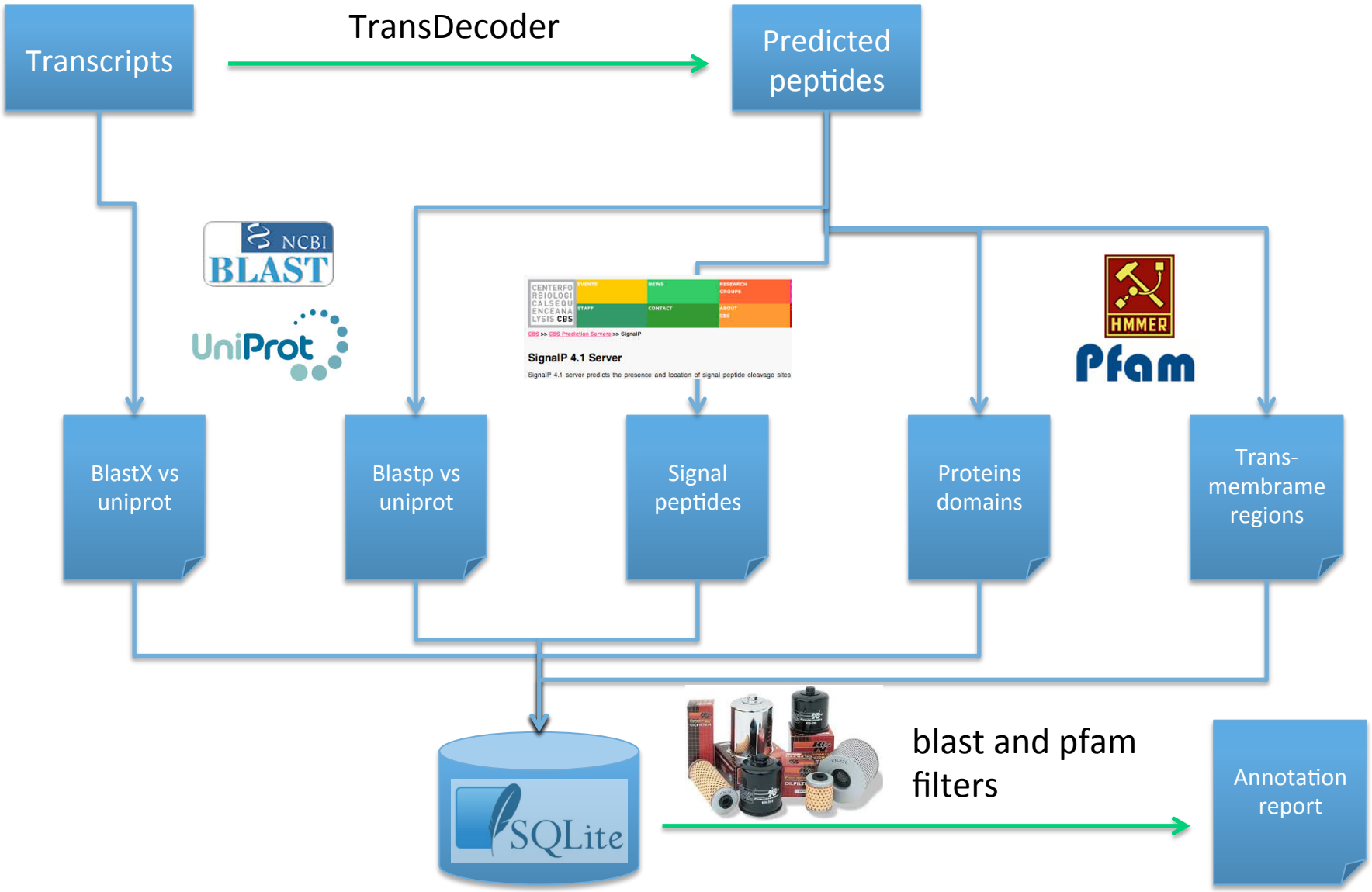
SQLite

the Gene Ontology

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

Automated Higher Order Biological Analysis

Trinotate pipeline



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 5 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).
- **optional** the putative peptide has a match to a Pfam domain above the noise cutoff score.

Trinnotate pipeline

2. Capturing BLASTP and BLASTX Homologies : uniprot-swissprot

3. Running HMMER to identify protein domains

4. Running signalP to predict signal peptides

5. Running tmHMM to predict transmembrane regions

Trinnotate pipeline

6 Loading Results into a Trinotate SQLite Database

7 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

Trinnotate pipeline

0 comp1507_c0

1 comp1507_c0_seq1:405-1415(+)

2 m.772

3 sp|Q7Z8R5|PALI_YARLI`Q7Z8R5`Q:1-236,H:3-234`30.13%ID`E:3e-21`RecName: Full=pH-response regulator protein palI/RIM9;`Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Dipodascaceae; Yarrowia.

4 PF06687.7^SUR7^SUR7/PalI family^7-165^E:2.4e-18

5 sigP:1`23`0.564`YES

6 ExpAA=91.29`PredHel=4`Topology=i9-31o89-111i118-140o150-172i

7 NOG12793^ Calcium ion binding protein

8 GO:0016021^cellular_component^integral to membrane`GO:0005886^cellular_component^plasma membrane

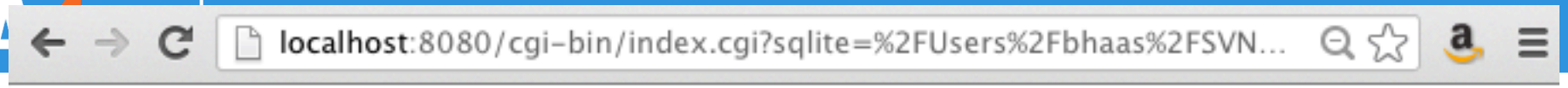
9
MRIRSATPSLILLVIAIVFFVLAICTPPLANNLTLGKYGDVRFGVFGYCLNSNCSKPLVGYNSDYLDEHAKDGFRTSVIVRQ
RASYGLVIVPVSACICLISTIMTIFAHIGAIARSPGFFNVIGTITFFNIFITAIAFVICVITFVPHIQWPSWLVLANVGIQLIVLL
LLLVARRQATRLQAKHLRRATSGSLGYNPYSLQNSSNIFSTSSRKGDLPKFSDYSAEKPMYDTISEDDGLKRGGSVSKLK
PTFSNDSRSLSSYAPTVREPVPVPKSNSGFRFPFMRNKPAEQAPENPFRDPENPFKDPASAPAPNPWSINDVQANND
KKPSRFSWGRS*

0	#component	5	SignalP
1	trans_derived	6	TmHMM
2	prot_id	7	eggnog
3	TopBlastHit	8	gene_ontology
4	Pfam	9	prot_seq

Not yet implemented on the cluster

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



Trinotate Web for Annotation and Expression Analysis

Stats

Various summary stats go here...

Got 8694 genes and 9299 transcripts

Search

Text search of transcript annotations

search

Still needed: search based on specific attribute: pfam, go, kegg, etc.

Pairwise Expression Comparisons (Volcano and MA plots)

	Sp_hs	Sp_log	Sp_plat
Sp_ds	Sp_ds vs. Sp_hs	Sp_ds vs. Sp_log	Sp_ds vs. Sp_plat
Sp_hs		Sp_hs vs. Sp_log	Sp_hs vs. Sp_plat
Sp_log			Sp_log vs. Sp_plat

Multi-sample Comparisons (Expression Profiling)

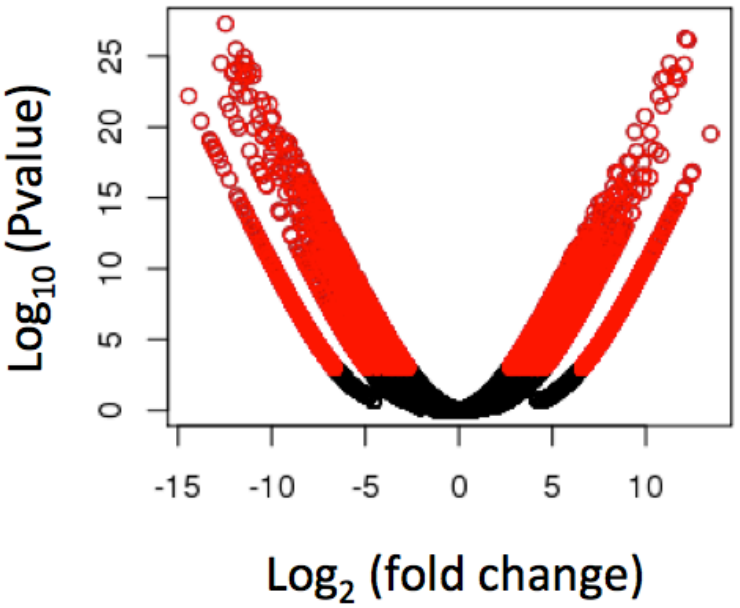
Go to the interactive heatmap for all DE transcripts.

Analyses of clusters of expression profiles:

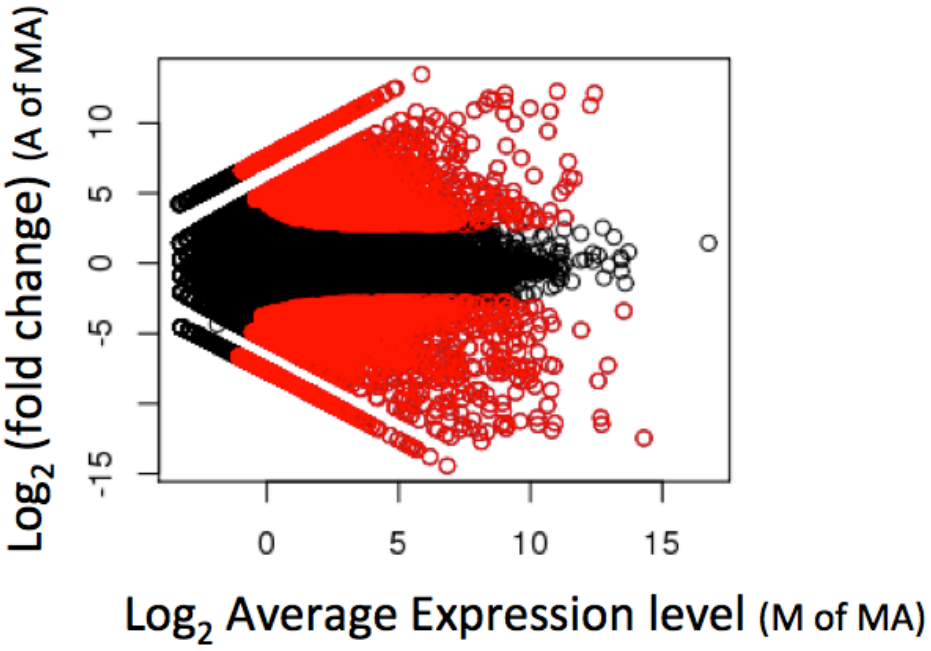
- edgeR_trans/diffExpr.P0.001_C2.matrix.R.all.RData.clusters_fixed_P_20 with 55 clusters.

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

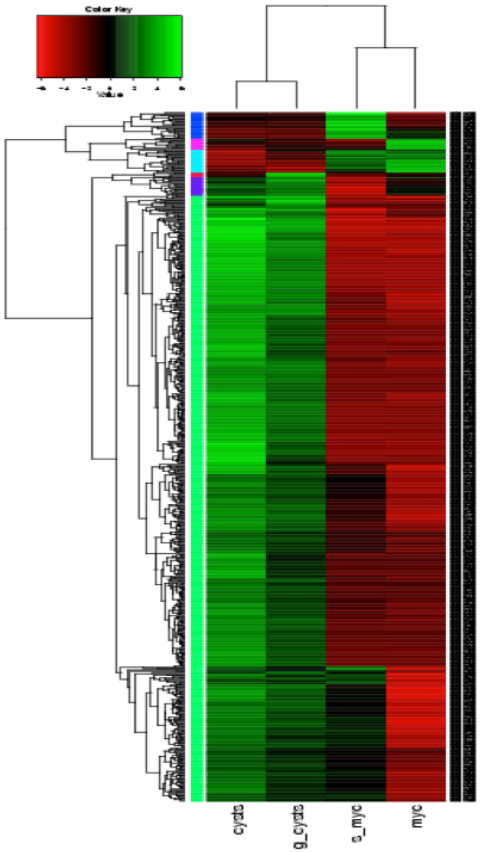


MA plot
(abundance vs. fold change)



Significantly differently expressed transcripts have FDR <= 0.001 (shown in red)

Comparing Multiple Samples



Heatmaps provide an effective tool for navigating differential expression across multiple samples.

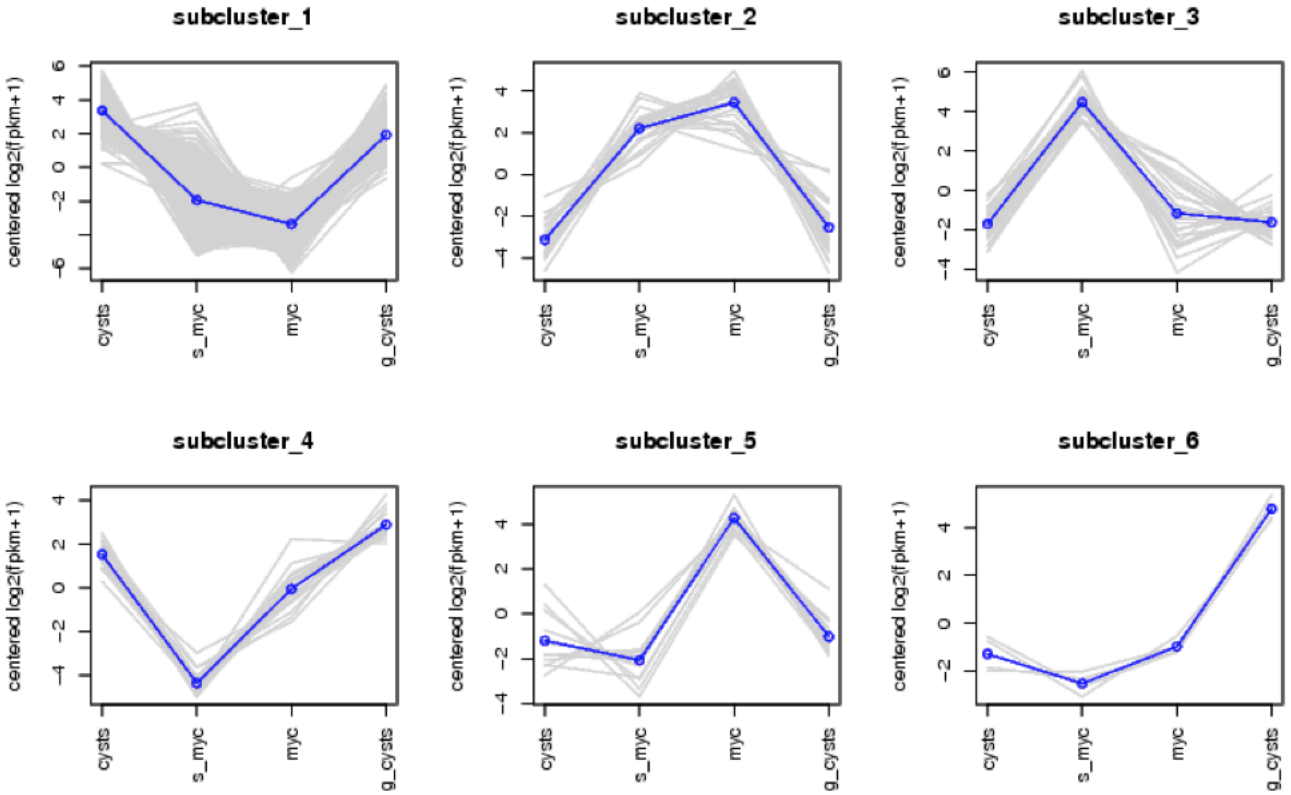
Clustering can be performed across both axes:

-cluster transcripts with similar expression patterns.

-cluster samples according to similar expression values among transcripts.

Examining Patterns of Expression Across Samples

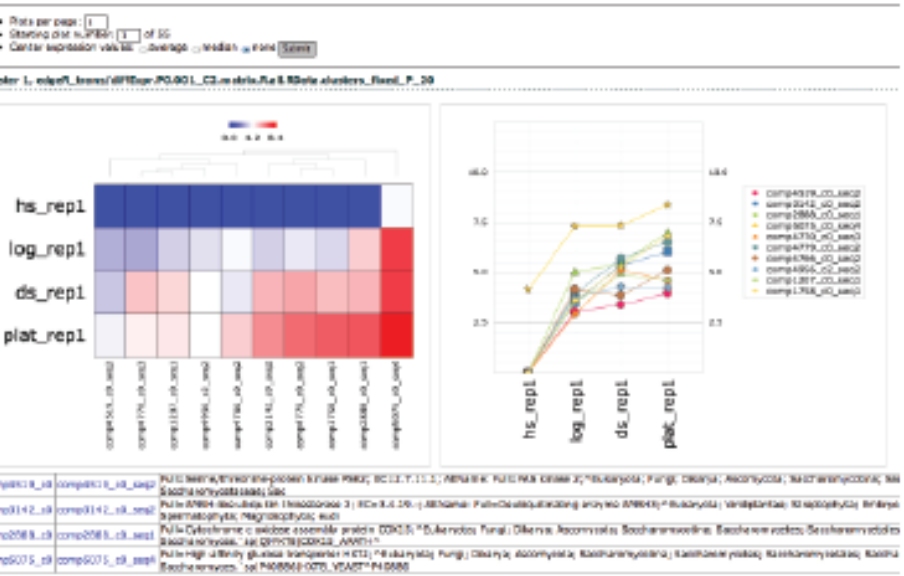
Can extract clusters of transcripts and examine them separately.



Station Biologique Roscoff

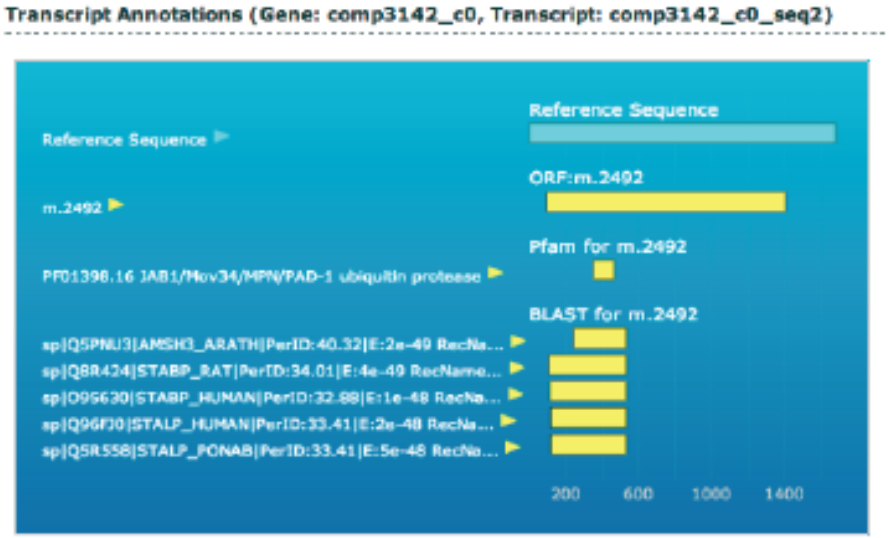
Abims

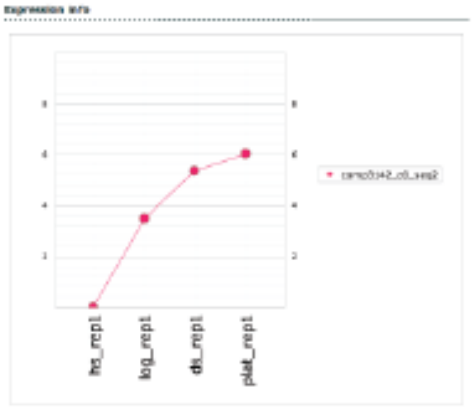
Clustered Expression Profiles



Very Early Release and
Just Scratching the Surface

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



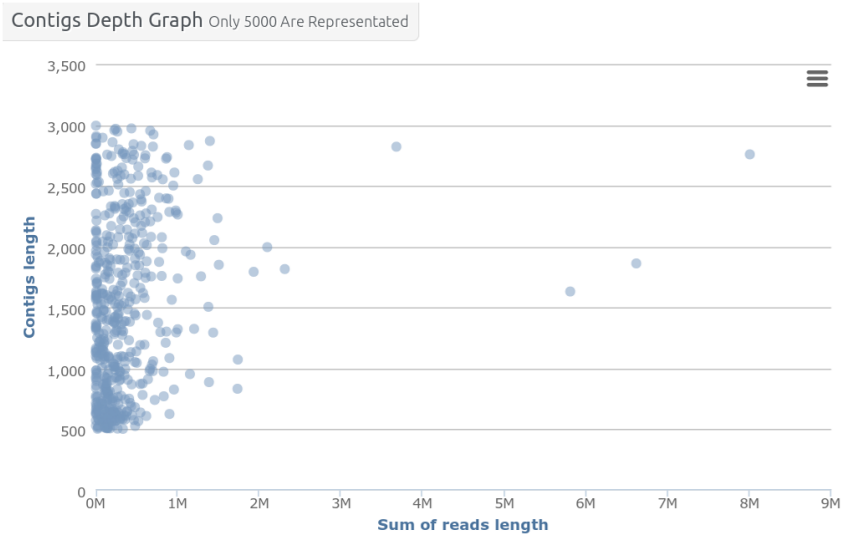


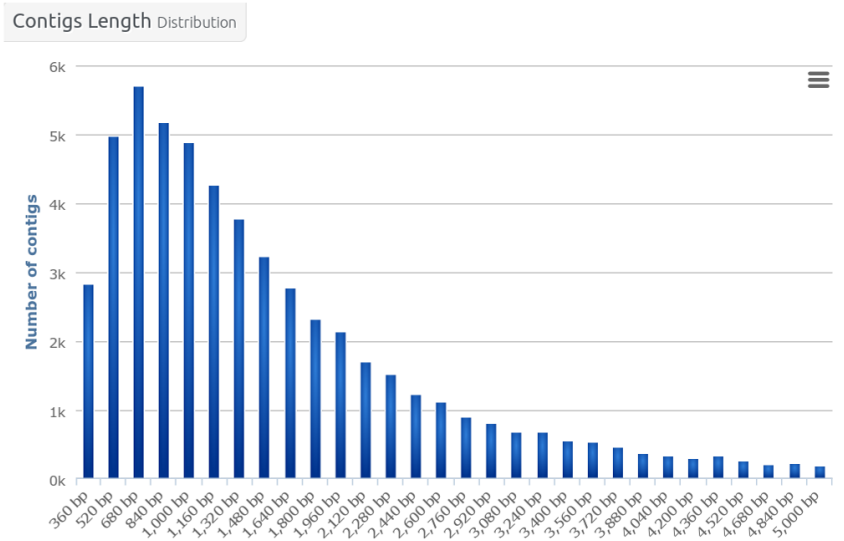
Individual Transcript
Expression Profiles

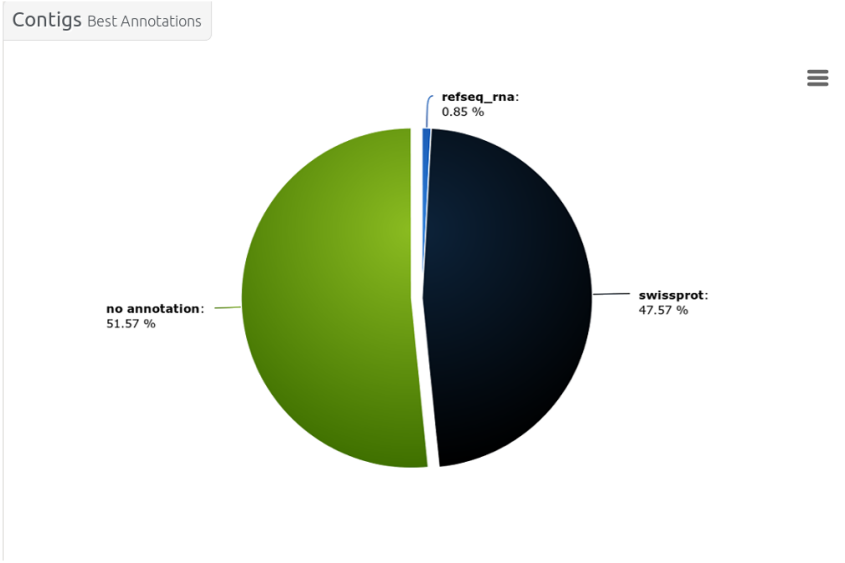
Transcript and
Protein Sequences

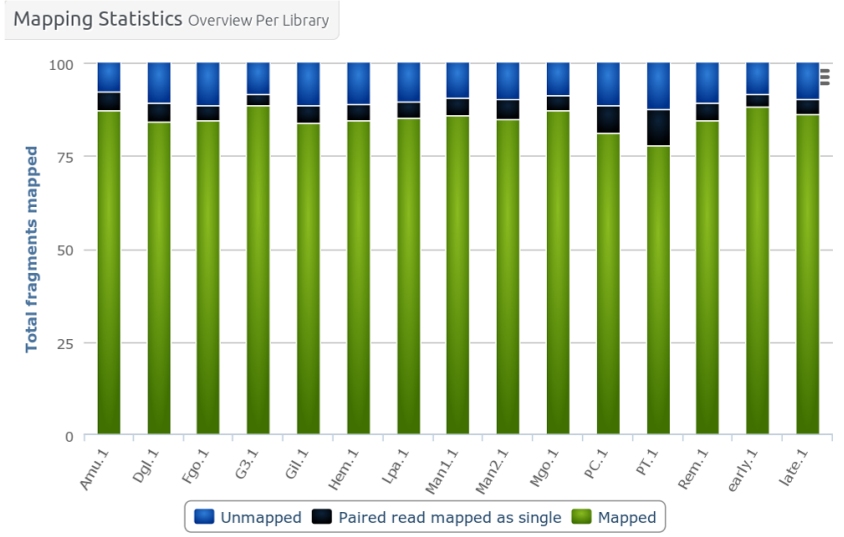
RNAbrowse

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





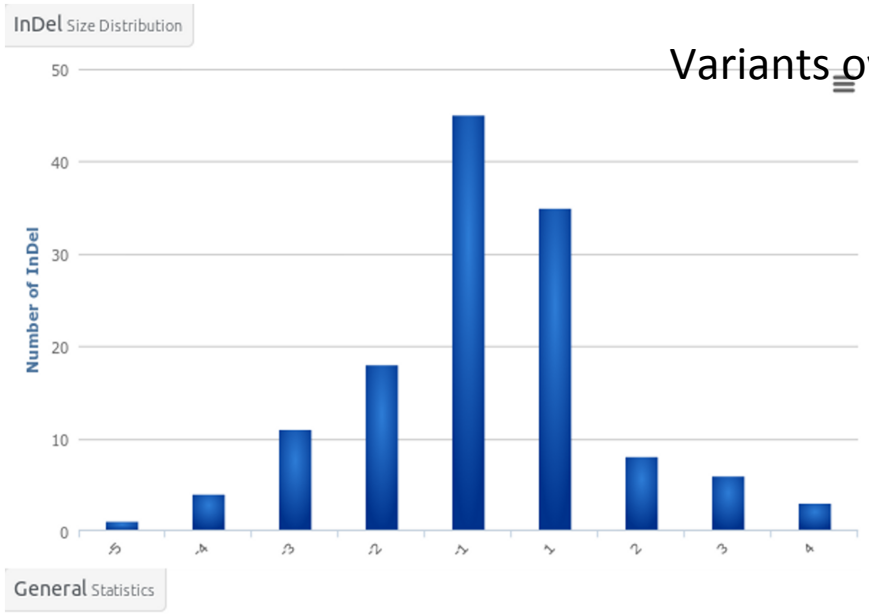


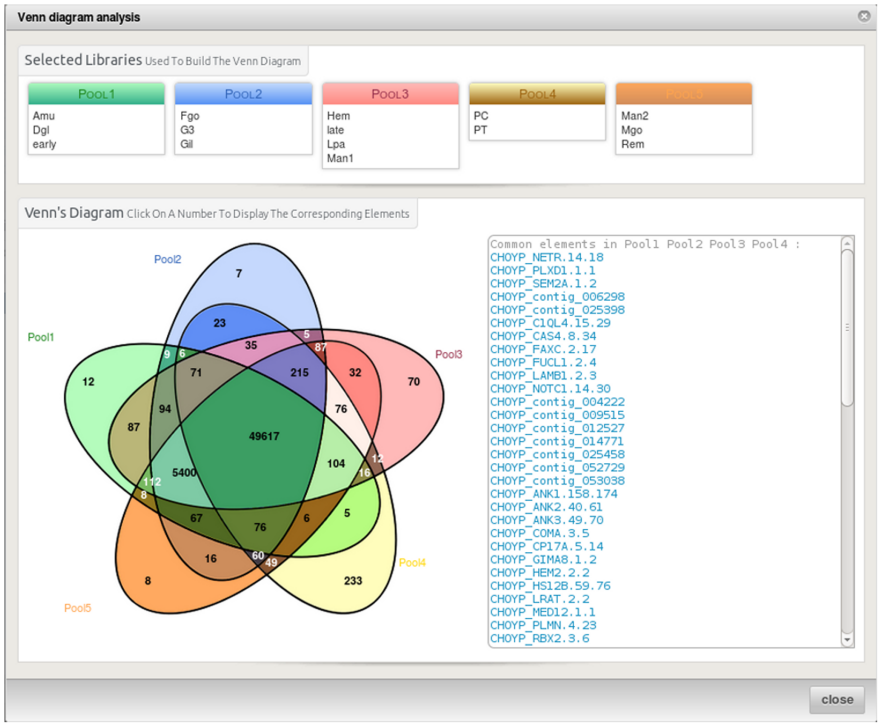


Contigs overview figures

RNAbrowse

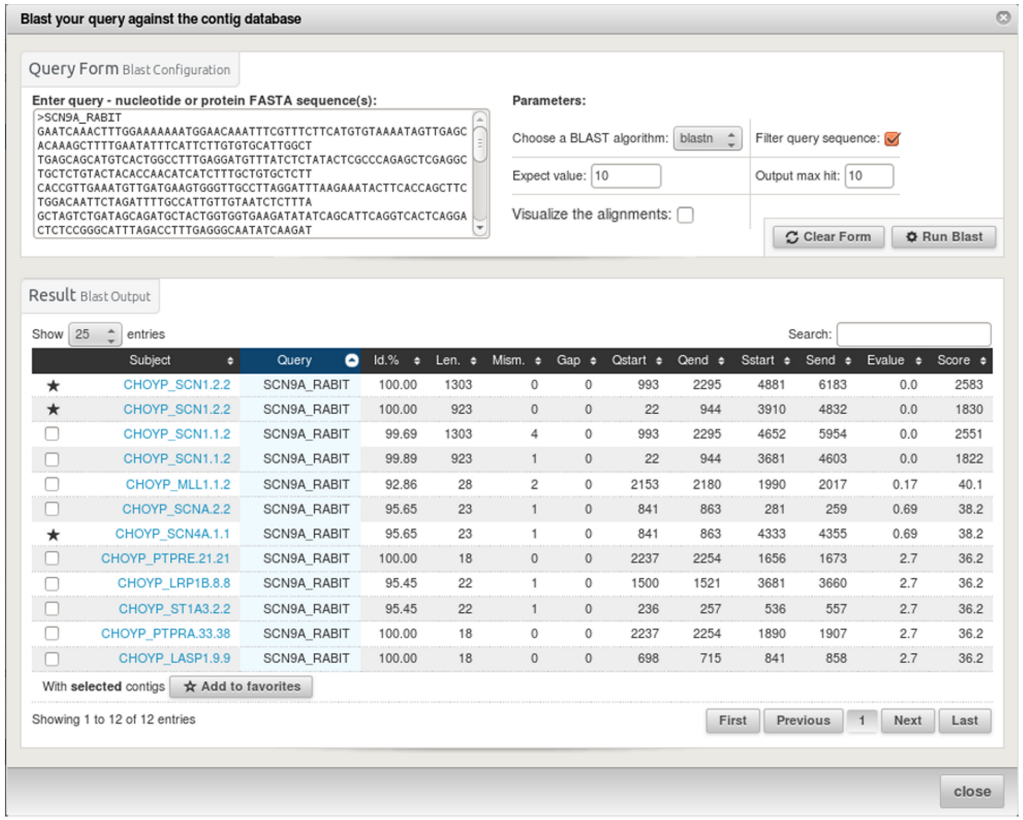
Variants overview figures



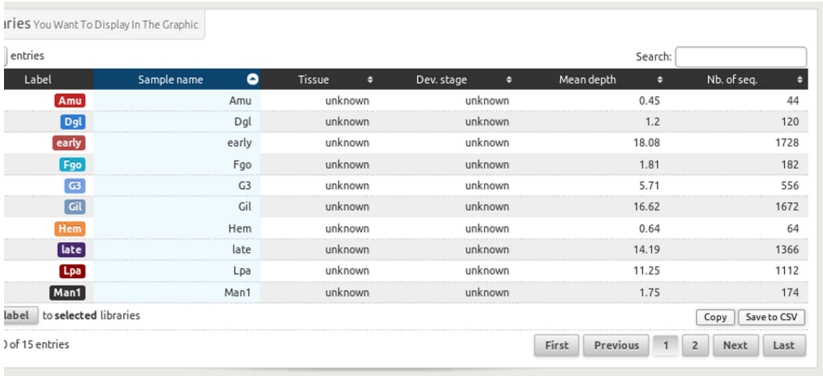


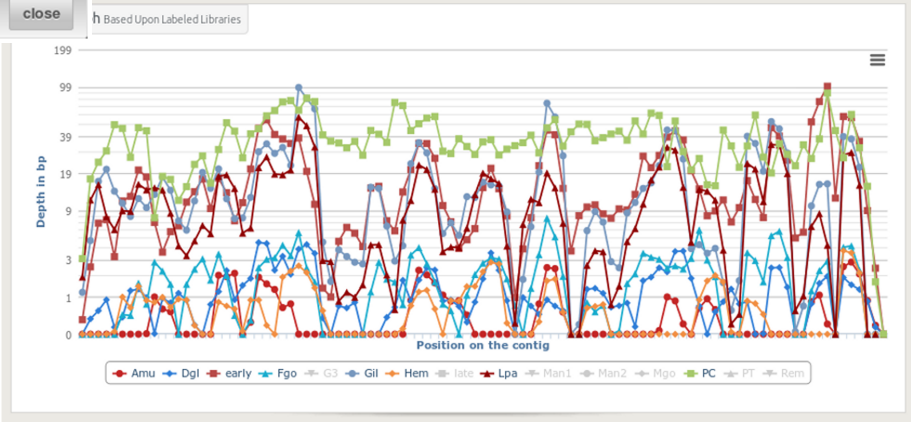
The Venn diagram shows the number of contigs shared between libraries

RNAbrowse



Blast interface





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

TRAPID: Rapid Analysis of Transcriptome Data

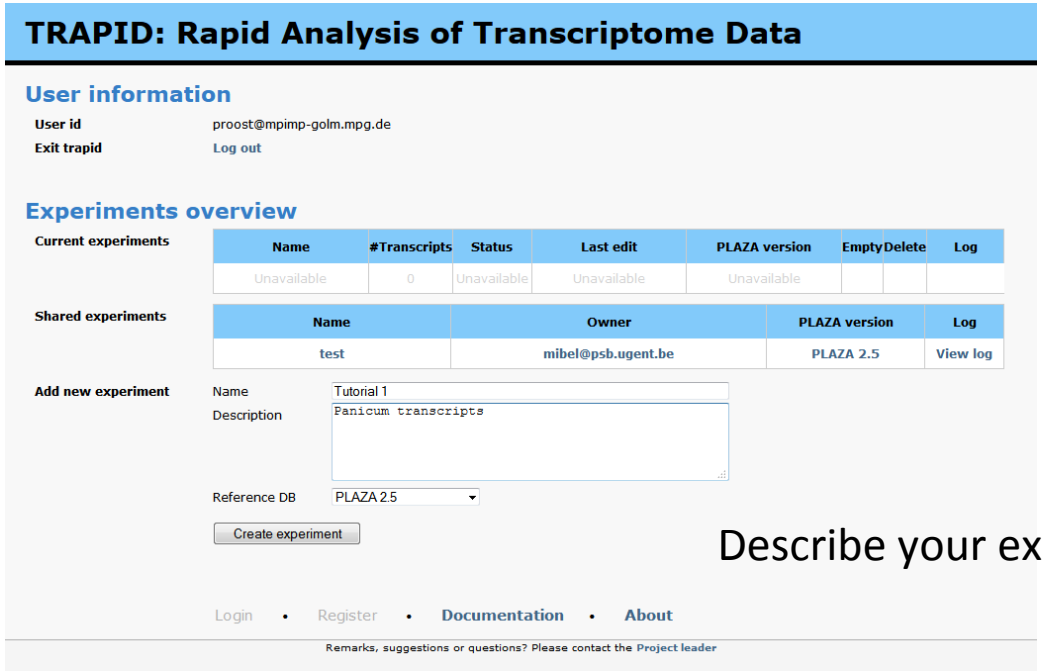
http://bioinformatics.psb.ugent.be/webtools/trapid/

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

TRAPID: Rapid Analysis of Transcriptome Data

Process transcripts

Experiments
Manage jobs
Documentation

Experiment overview

Name	Tutorial 1
Processing status	finished
Last edit	2013-05-06 14:33:19
Data source	PLAZA 2.5
Transcript count	25392

Overview

The transcript pipeline of the PLAZA workbench can be used to analyze transcripts (provided by the user) of species not present in the PLAZA database. This is useful for e.g. transcriptome analyzes during specific conditions or for species for which no genome is present, only a transcriptome. Transcripts are initially associated with PLAZA gene families using a translational approach. Further analyzes are then done on a per-family basis.

Options

Use NCBI Taxonomy to find the closest relative species or best clade.

Similarity Search Database Type	Single Species
Similarity Search Database	Arabidopsis lyrata
Similarity Search E-value	10e-5
Gene Family type	Gene Families
Functional annotation	Transfer based on gene family

Start transcriptome pipeline

Start the analysis

Remarks, suggestions or questions? Please contact the Project leader

TRAPID: Rapid Analysis of Transcriptome Data

TRAPID: Rapid Analysis of Transcriptome Data

Statistics

Experiment overview

Name	Tutorial 1
Processing status	finished
Last edit	2013-05-06 14:33:19
Data source	PLAZA 2.5
Transcript count	25392

Experiments
Documentation

Transcript information

#Transcripts	25392
Average sequence length	777.8
Average ORF length	532.9
#ORFs with a start codon	7186 (28.3%)
#ORFs with a stop codon	15755 (62%)

Frameshift information

#Transcripts with putative frameshift	2805 (11%)
#Transcripts with corrected frameshift	0 (0%)

Meta annotation information

#Meta annotation full-length	1894 (7.5%)
#Meta annotation quasi full-length	8028 (31.6%)
#Meta annotation partial	6361 (25.1%)
#Meta annotation no information	9109 (35.9%)

Similarity search information

Best similarity search hit for each transcript.

Sorghum bicolor	8372 (49.9%)
Zea mays	4453 (26.6%)
Oryza sativa ssp. indica	1470 (8.8%)
Oryza sativa ssp. japonica	1288 (7.7%)
Brachypodium distachyon	1184 (7.1%)
Total	16767

Gene family information

#Gene families	4989
#Transcripts in GF	16767 (66%)
Largest GF	114_HOM000002 (156 transcripts)
#single copy	2247

Functional annotation information

Gene Ontology

#GO terms	3092
#Transcripts with GO	12044 (47.4%)

InterPro

#InterPro domains	3639
#Transcripts with Protein Domain	13818 (54.4%)

Export

PDF export

General statistics

TRAPID: Rapid Analysis of Transcriptome Data

Search

Experiment overview

Name	Tutorial 1
Processing status	finished
Last edit	2013-05-06 14:33:19
Data source	PLAZA 2.5
Transcript count	25392

Search results

Click table-header(s) to enable sorting

GO term	GO description	#transcripts
GO:0005618	cell wall	31
GO:0006037	cell wall chitin metabolic process	25
GO:0007047	cellular cell wall organization	5
GO:0009273	peptidoglycan-based cell wall biogenesis	4
GO:0009505	plant-type cell wall	11
GO:0009664	plant-type cell wall organization	7
GO:0010382	cellular cell wall macromolecule metabolic process	5
GO:0010383	cell wall polysaccharide metabolic process	26
GO:0016998	cell wall macromolecule catabolic process	6
GO:0042545	cell wall modification	13
GO:0042546	cell wall biogenesis	6
GO:0044036	cell wall macromolecule metabolic process	36
GO:0044038	cell wall macromolecule biosynthetic process	5
GO:0052386	cell wall thickening	6
GO:0052482	cell wall thickening during defense response	6
GO:0052543	callose deposition in cell wall	6
GO:0052544	callose deposition in cell wall during defense response	6
GO:0070592	cell wall polysaccharide biosynthetic process	1
GO:0070882	cellular cell wall organization or biogenesis	11
GO:0071554	cell wall organization or biogenesis	37
GO:0071555	cell wall organization	25
GO:0071669	plant-type cell wall organization or biogenesis	7

Click table-header(s) to enable sorting

Perform new search

GO description | cell wall | Search

Remarks, suggestions or questions? Please contact the Project leader

Functional enrichment analysis
Phylogenetic analysis

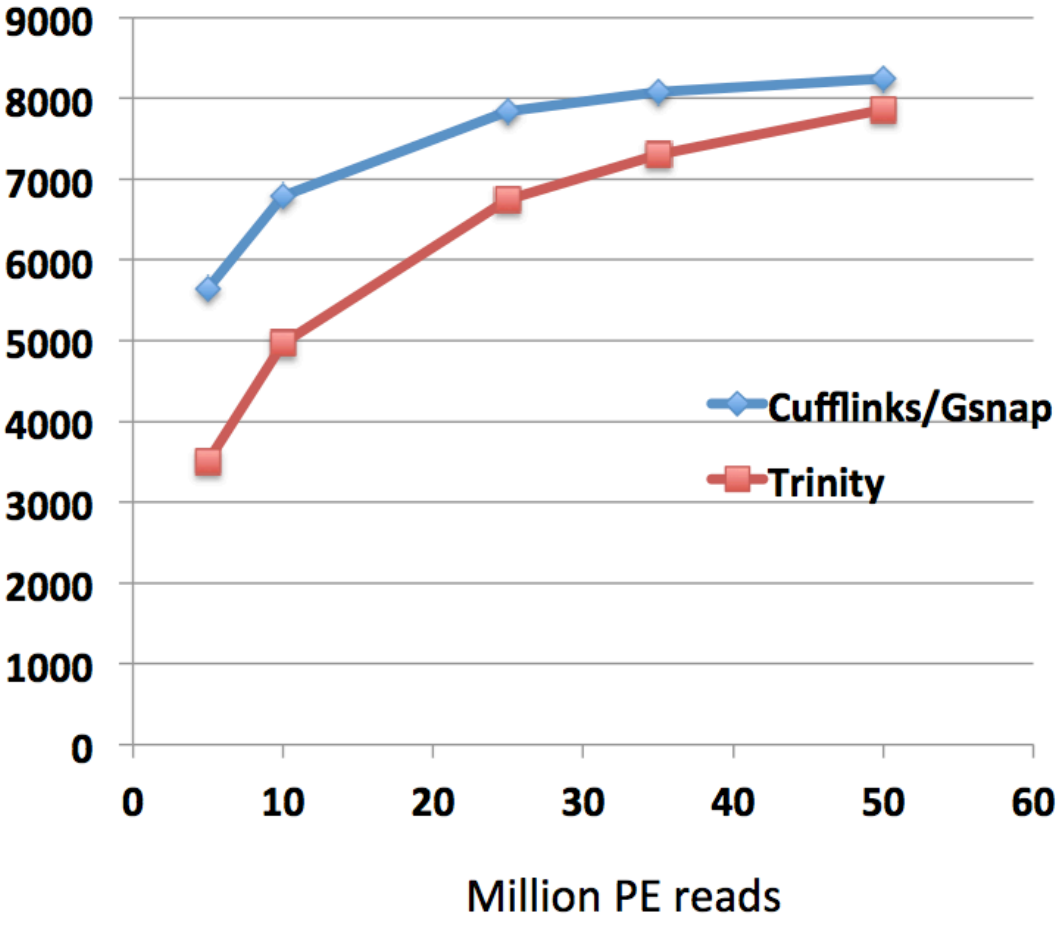
To infinity... and beyond

With ref. vs de novo

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

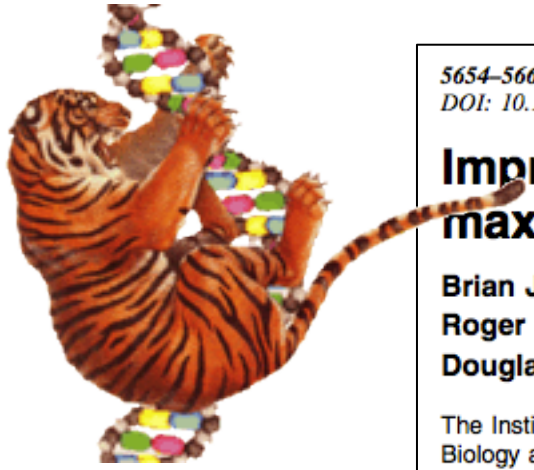
Genes w/ fully reconstructed transcripts

Mouse data



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

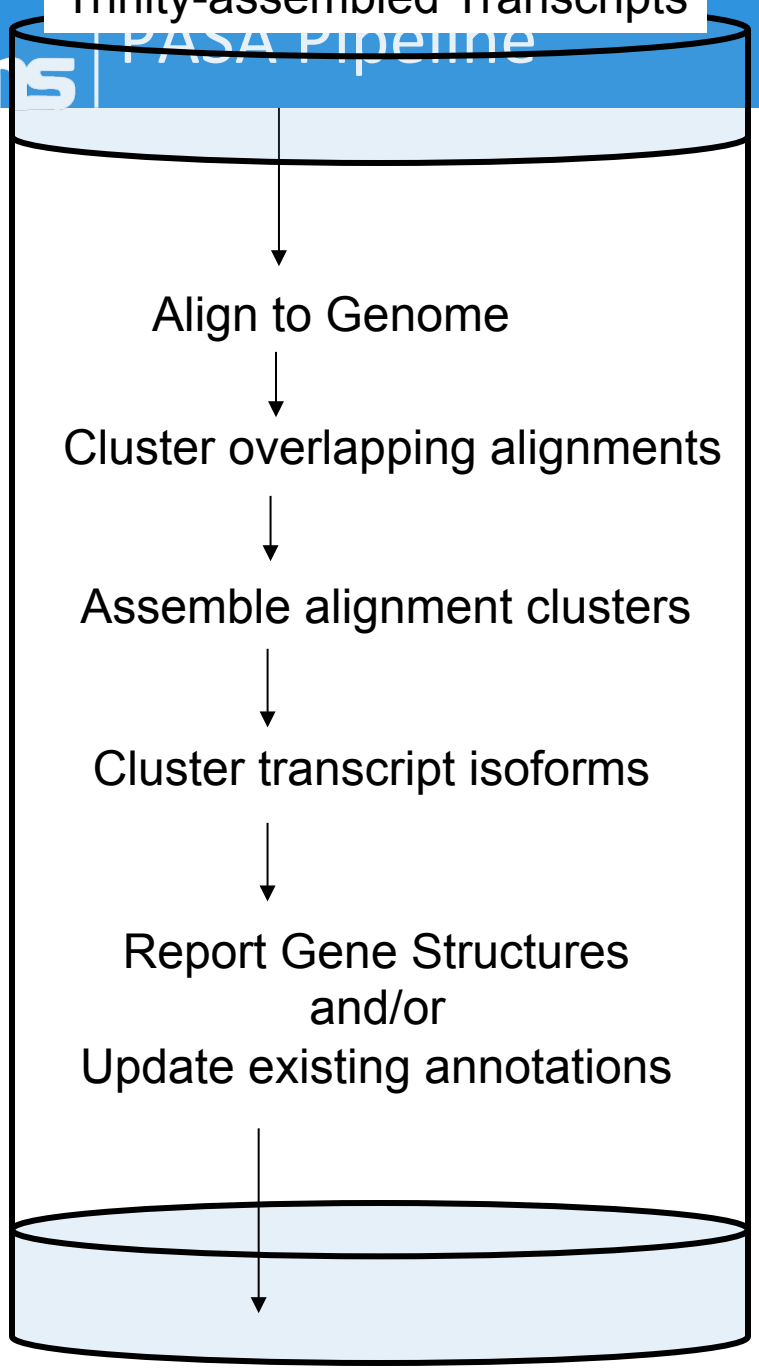
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, [1]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA and [2]The Center for Advancement of Genomics, 1901 Research Boulevard, Rockville, MD 20850, USA

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

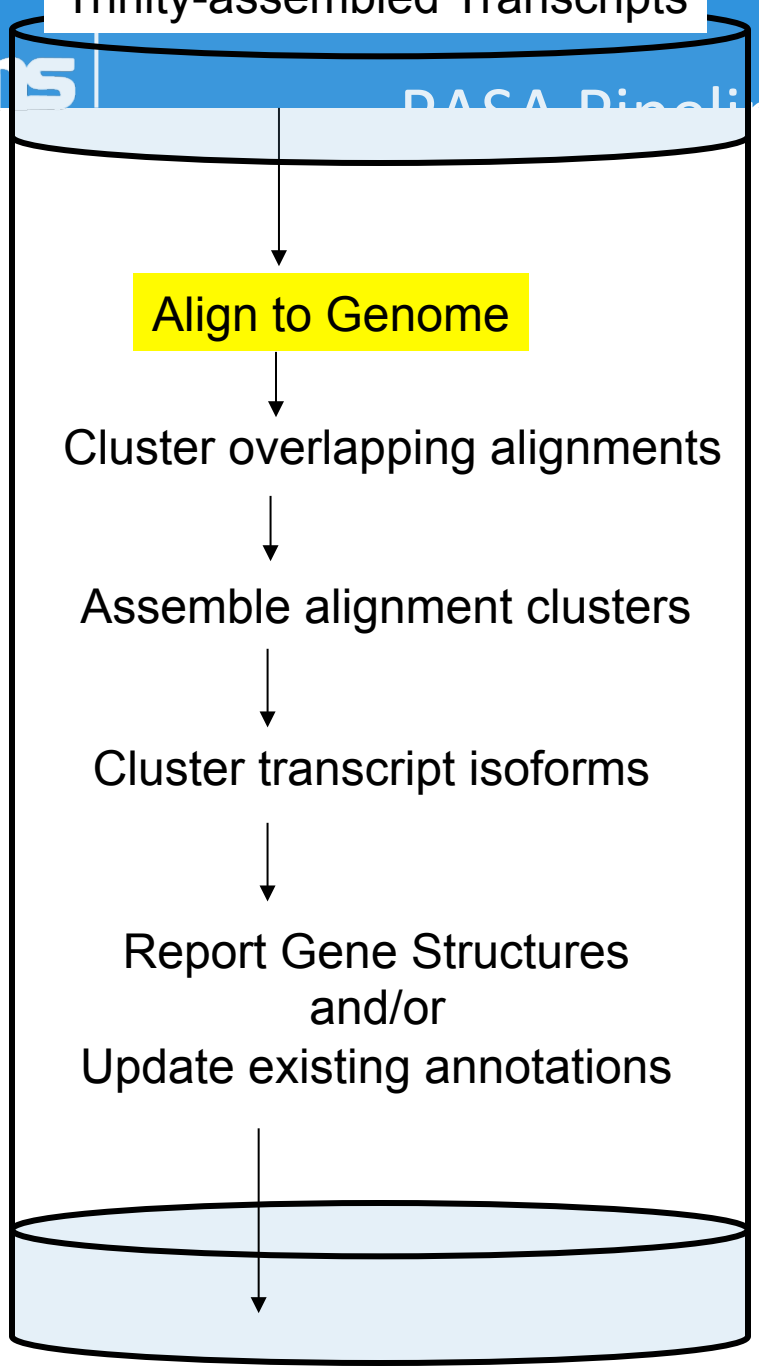
Developed (in 2003) to integrate ESTs and full-length cDNAs into gene structure annotations.

Compatible with RNA-Seq via Trinity.

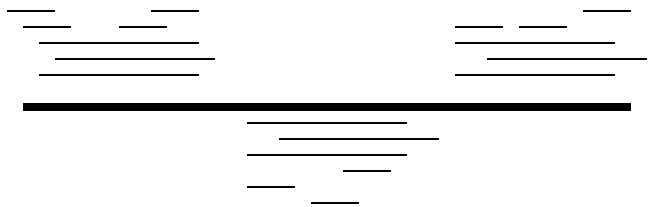
PASA Pipeline



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**

Trinity-assembled Transcripts

Align to Genome

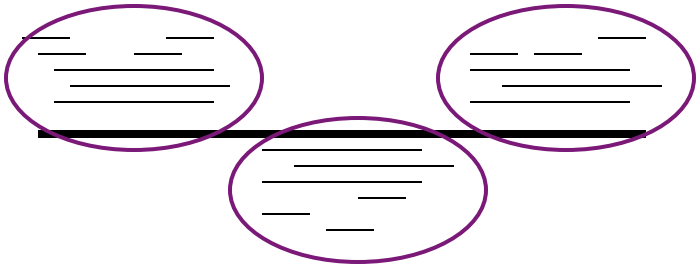
Cluster overlapping alignments

Assemble alignment clusters

Cluster transcript isoforms

Report Gene Structures
and/or
Update existing annotations

spliced alignments



Trinity-assembled Transcripts

Align to Genome

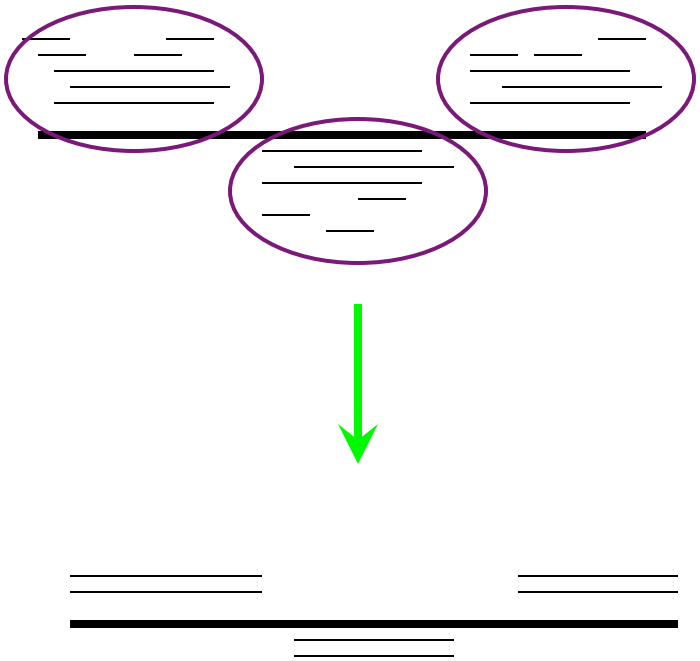
Cluster overlapping alignments

Assemble alignment clusters

Cluster transcript isoforms

Report Gene Structures
and/or
Update existing annotations

spliced alignments



PASA Pipeline

Trinity-assembled Transcripts

Align to Genome

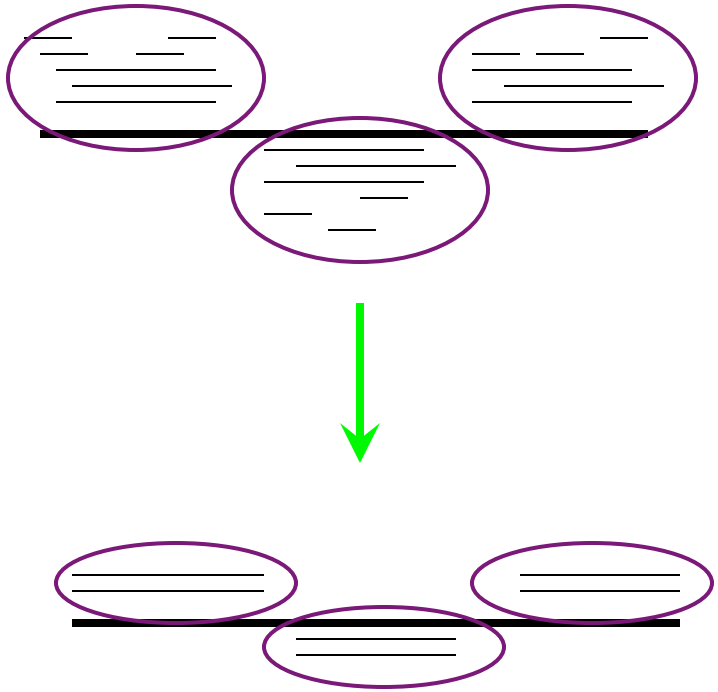
Cluster overlapping alignments

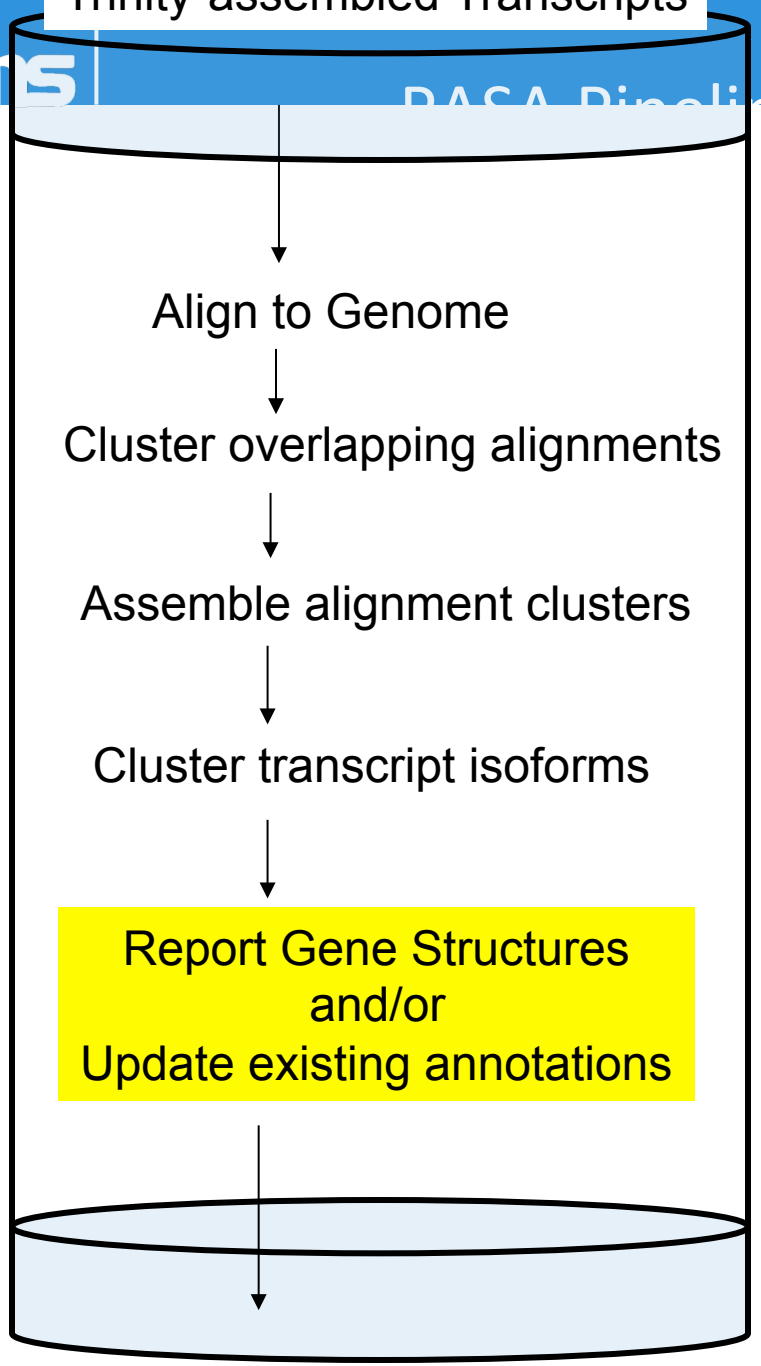
Assemble alignment clusters

Cluster transcript isoforms

Report Gene Structures
and/or
Update existing annotations

spliced alignments





Annotation output

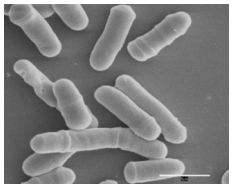
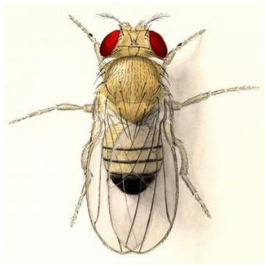
-gene structures
-alt splice isoforms
-predicted coding regions

(fasta, bed, gff3, gtf formats)

Annotation Updates

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

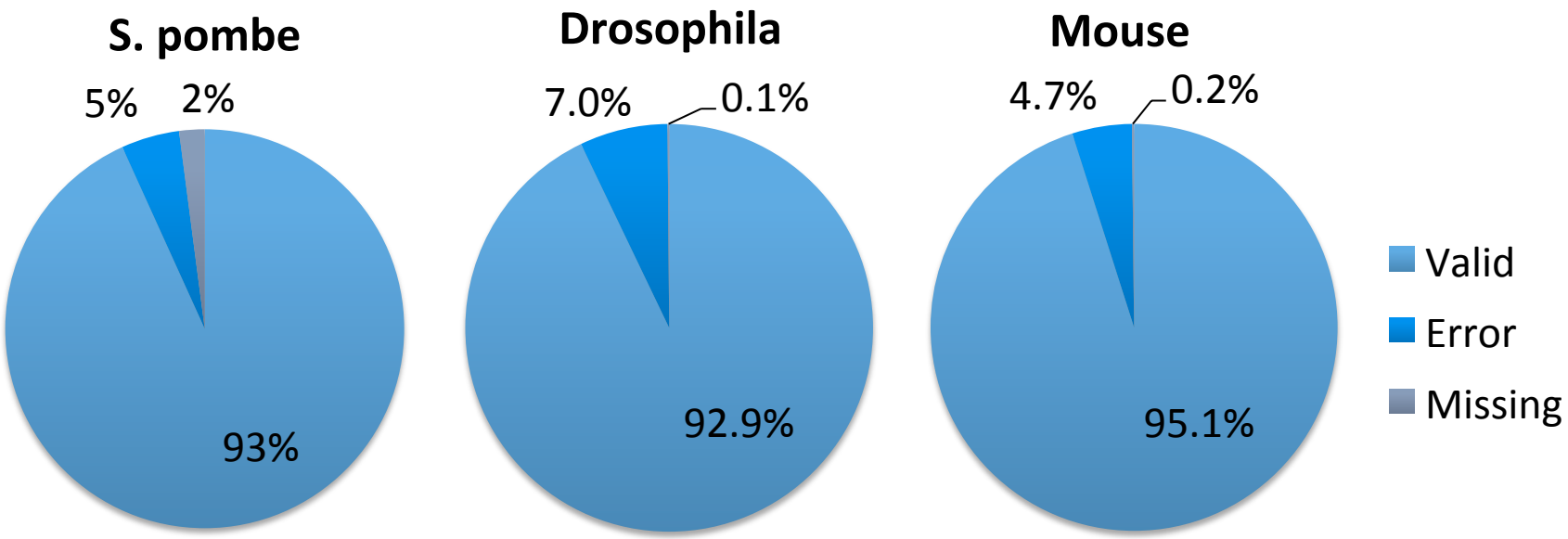
Evaluating Genome-based Transcript Reconstruction Using Reference Genomes + Transcriptomes

	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



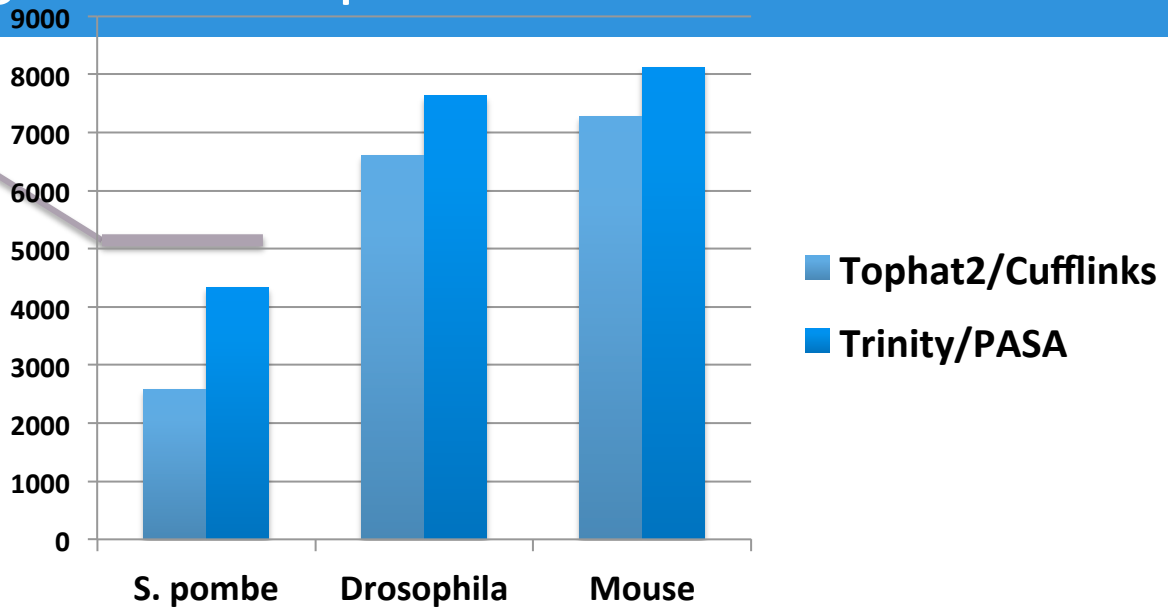
# Trinity Transcripts	14,548	36,320	81,516

~5% to 7% of assembled transcripts are problematic

Full-length Transcript Reconstruction from RNA-Seq

Total pombe genes

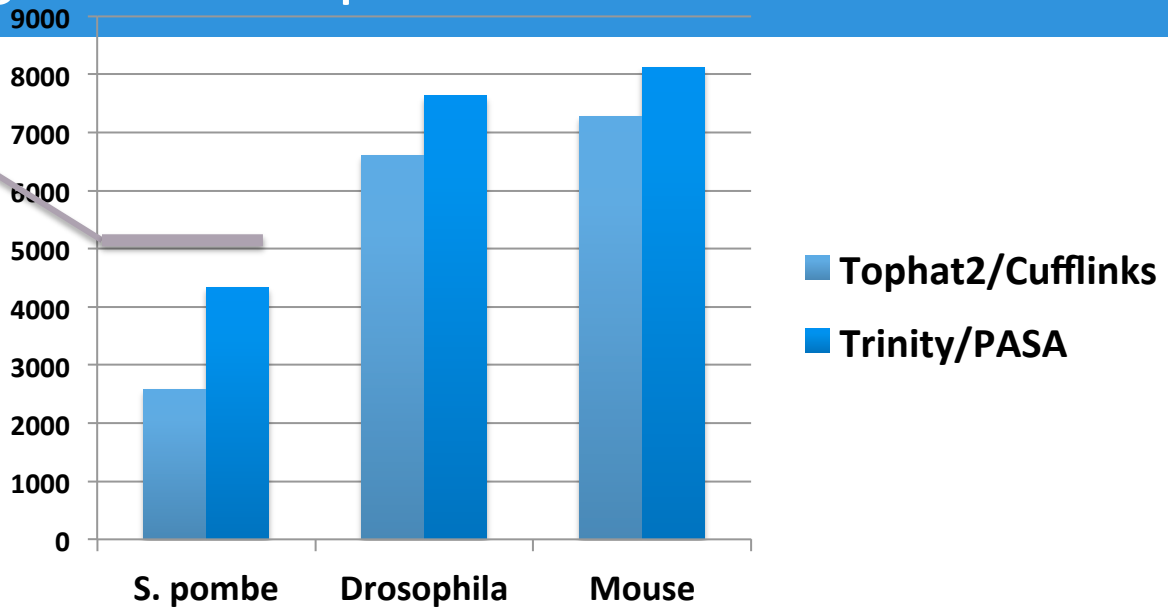
Number of genes with full length transcripts



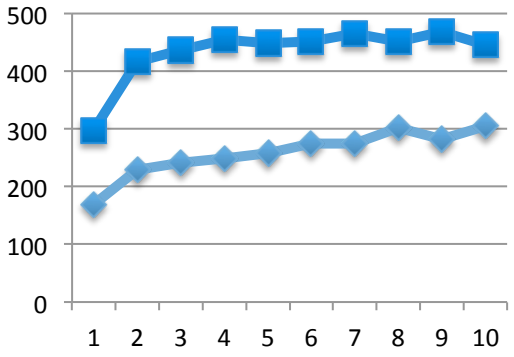
Full-length Transcript Reconstruction from RNA-Seq

Total pombe genes

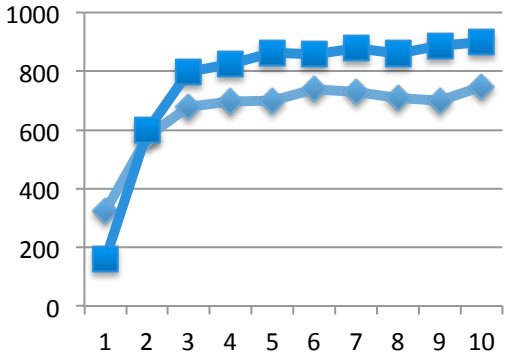
Number of genes with full length transcripts



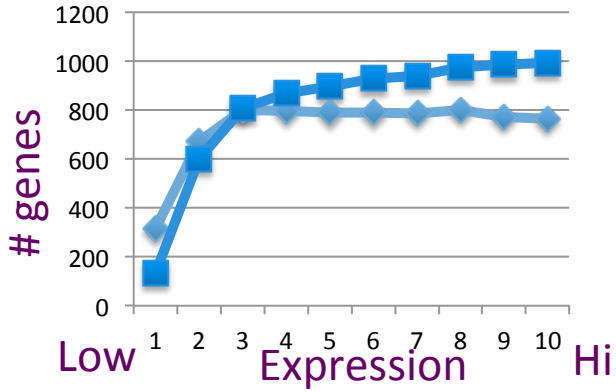
Schizosaccharomyces pombe



Drosophila



Mouse



Full-length Reconstruction by Expression Quintile