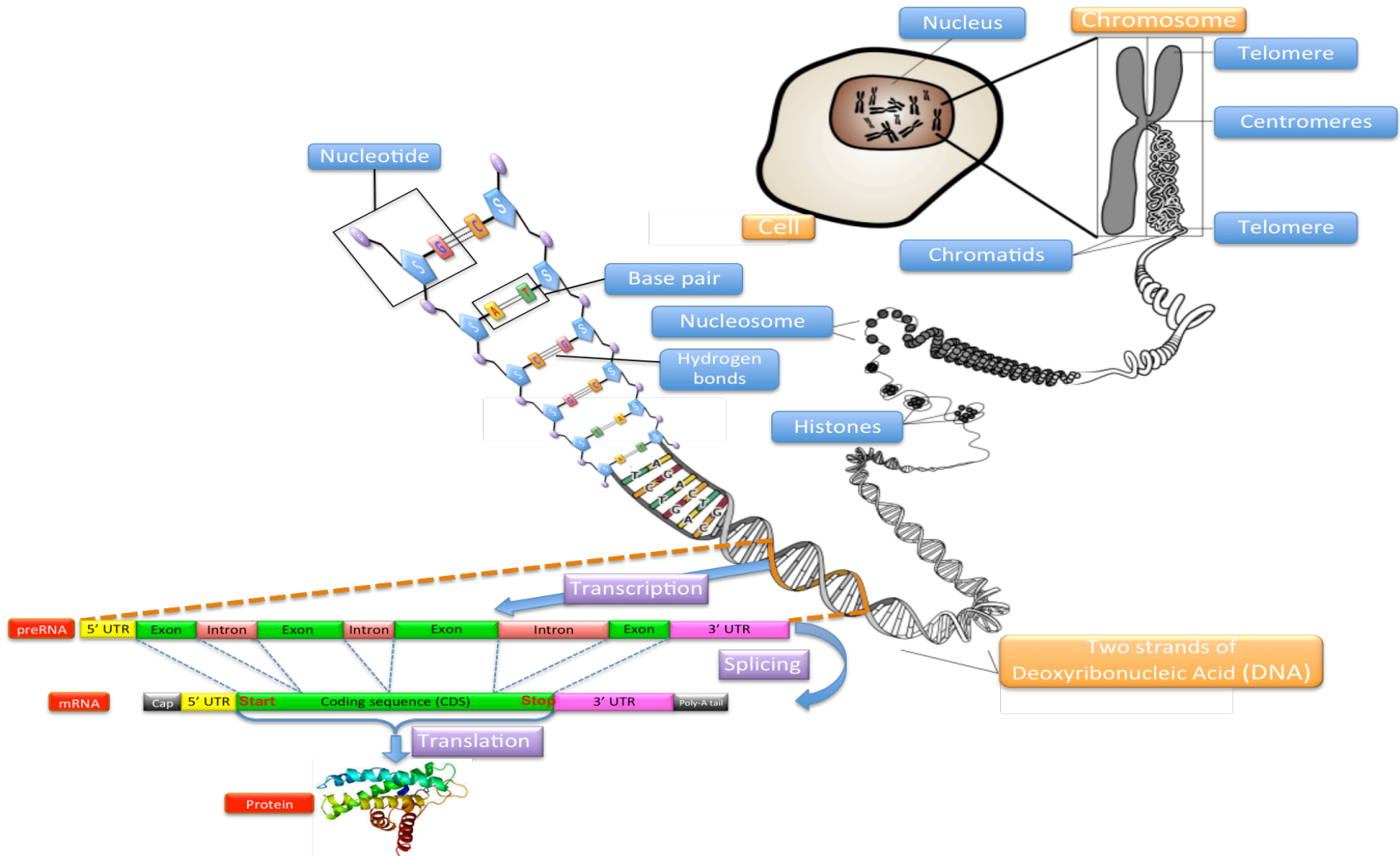
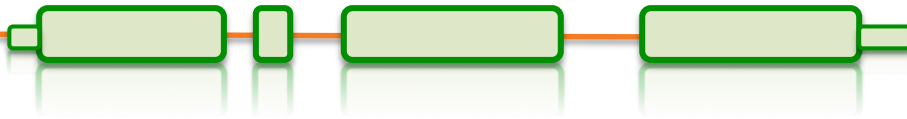
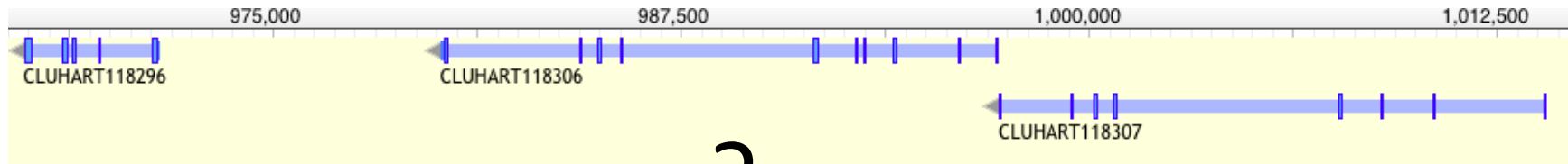


Functional annotation





Right, now we have our genes, but what do they do ?



?

?

?

Insulin receptor?

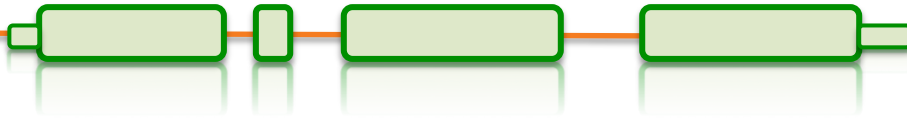
Vesicle-trafficking protein?

Alcohol dehydrogenase?

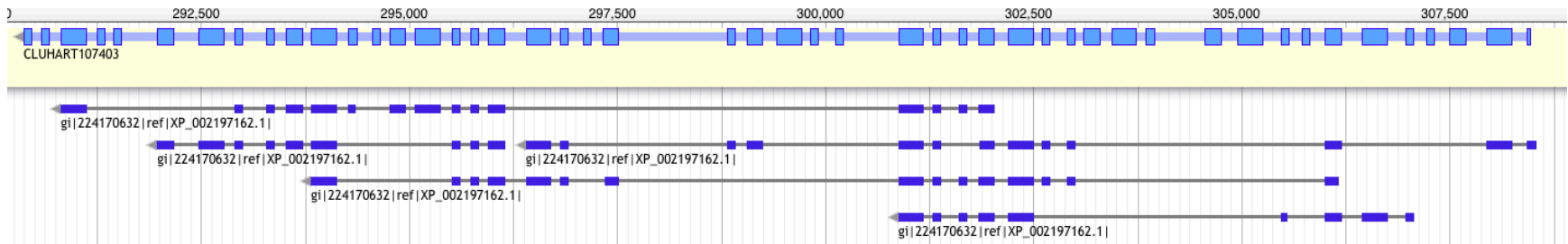
Aquaporin?

Transcription factor

MAP kinase kinase kinase?

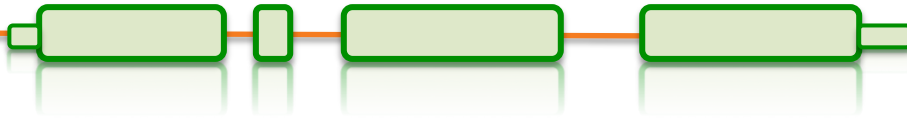


But we have used proteins in our annotation!



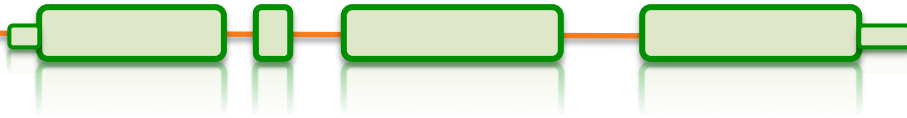
It is actually kind of complex...

... and most of pipelines do not do this for you.



First you need the sequences

- Extract sequences from the browser (Webapollo)
- GFF3 => fasta : Use gffread (in Cufflinks package)
- Fasta available (Biomart, FTP, output of annotation tools)
- If CDS=> translate in AA : Use gffread (in Cufflinks package)



Methods

- **Based on blast**
 - Best blast hit
 - Clustering
- **Based on synteny**
 - ⇒ Whole genome alignment (lastZ)
 - (NBIS) Satsuma + kraken + custom script
- **Based on phylogeny**

Tool	Approach	Comment
Trinotate	Best blast hit + protein domain identification (HMMER/PFAM) + protein signal peptide and transmembrane domain prediction (signalP/tmHMM), and leveraging various annotation databases (eggNOG/GO/Kegg databases).	Not automated
Annocript	Best blast hit	Collects the best-hit and related annotations (proteins, domains, GO terms, Enzymes, pathways, short)
Annot8r	Best blast hits	A tool for Gene Ontology, KEGG biochemical pathways and Enzyme Commission EC number annotation of nucleotide and peptide sequences.
Sma3s	Best blast hit + Best reciprocal blast hit + clusterisation	3 annotation levels
afterParty	BLAST, InterProScan	web application
Interproscan	Run separate search applications HMMs, fingerprints, patterns => InterPro	Created to unite secondary databases
Blast2Go	Best* blast hits	Retrieve only GO Commercial !

Database	Information	Comment
KEGG	Pathway	Kyoto Encyclopedia of Genes and Genomes
MetaCyc	Pathway	Curated database of experimentally elucidated metabolic pathways from all domains of life (NIH)
Reactome	Pathway	Curated and peer reviewed pathway database
UniPathway	Pathway	Manually curated resource of enzyme-catalyzed and spontaneous chemical reactions.
GO	Gene Ontology	Three structured, controlled vocabularies (ontologies) : biological processes, cellular components and molecular functions
Pfam	Protein families	Multiple sequence alignments and hidden Markov models
Interpro	Protein families, domains and functional sites	Run separate search applications, and create a signature to search against Interpro.
<p>Have a look on the Interpro web page: All the database they search into are listed. It gives a nice overview of different types of databases available.</p>		

Starting point for downstream analysis

GO term prediction

Biological Process

- [GO:0006631](#) fatty acid metabolic process
- [GO:0006635](#) fatty acid beta-oxidation
- [GO:0008152](#) metabolic process
- [GO:0055114](#) oxidation-reduction process

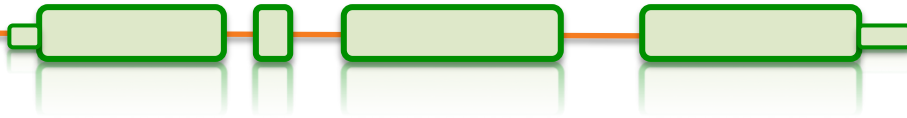
More than 60 000 terms

Molecular Function

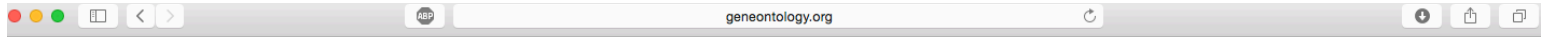
- [GO:0003824](#) catalytic activity
- [GO:0003857](#) 3-hydroxyacyl-CoA dehydrogenase activity
- [GO:0004300](#) enoyl-CoA hydratase activity
- [GO:0016491](#) oxidoreductase activity
- [GO:0016616](#) oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
- [GO:0050662](#) coenzyme binding

Cellular Component

- [GO:0005739](#) mitochondrion
- [GO:0016507](#) mitochondrial fatty acid beta-oxidation multienzyme complex



Starting point of downstream analysis



Enrichment analysis

Your gene IDs here...

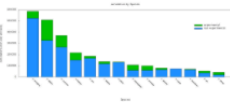
biological process

Homo sapiens

Submit

[Advanced options / Help](#)
Powered by [PANTHER](#)

Statistics



Other GOC tools

Explore other GOC tools in the AmiGO software suite.

[Tweets about #geneontology OR @news4go](#)

Gene Ontology Consortium

Search GO data

Search for terms and gene products...

Search

Ontology

[Filter classes](#)

[Download ontology](#)

Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:

- molecular function**
molecular activities of gene products
- cellular component**
where gene products are active
- biological process**
pathways and larger processes made up of the activities of multiple gene products.

[more](#)

Annotations

[Download annotations](#) (standard files)

[Filter and download](#) (customizable files <10k lines)

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt signaling pathway. Each statement is based on a specified piece of evidence. [more](#)

The mission of the GO Consortium is to develop an up-to-date, comprehensive, **computational model of biological systems**, from the molecular level to larger pathways, cellular and organism-level systems. [more](#)

Search documentation

Search

User stories

Explore documentation related to your personal [user story](#).

What is the Gene Ontology?

- [An introduction to the Gene Ontology](#)
- [What are annotations?](#)
- [Ten quick tips for using the Gene Ontology](#) Important
- [Enrichment analysis](#)
- [Downloads](#)

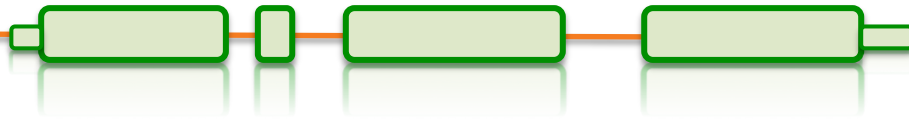
Recent news

[Paper on extending GO in the context of extracellular RNA and vesicle communication](#)
Post date: 04/21/2016 - 06:42

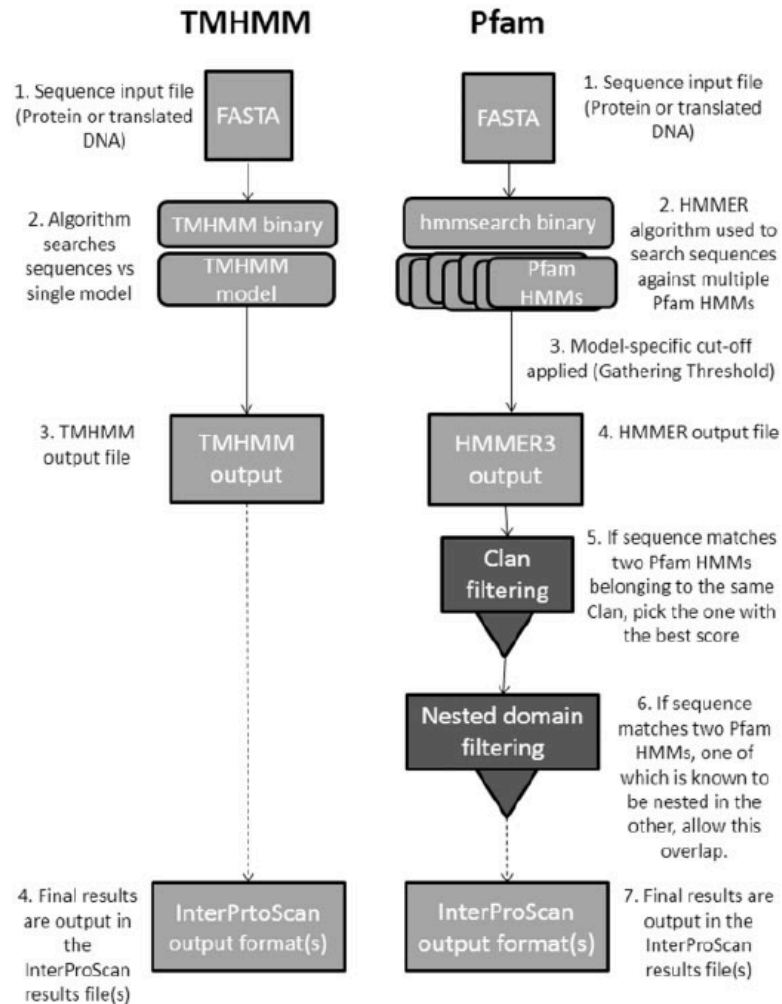
There is an equivalence table between GO and Interpro

GO annotation is given with an Evidence Code:

- IDA: inferred by direct assay
- ISS: inferred by sequence similarity
- IEA: electronic annotation
-

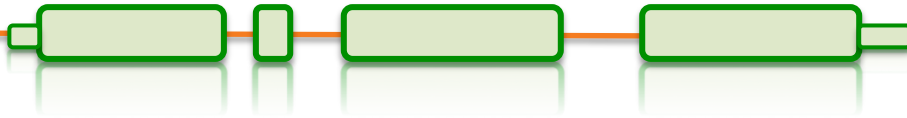


Search applications have two main modalities



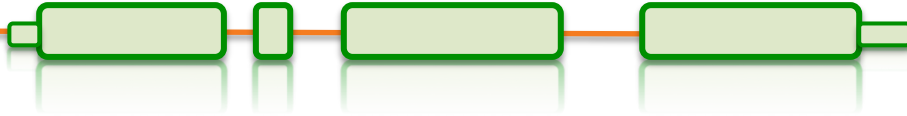
Jones,P.etal.InterProScan5:genome-scale protein function classification. Bioinformatics 30, 1236–1240 (2014).

Fig. 1. Comparison of the processing steps used by two different member database applications, TMHMM and Pfam



Annotate the sequences functionally using Blast

- Blast the protein-sequences from your maker-run to a protein blast-database (e.g., uniprot) using blastp from the Blast+ package
- Use Annie to extract best hits from blast-hit list and the corresponding description from uniprot-headers
- Add this information to the annotation.gff using custom-script



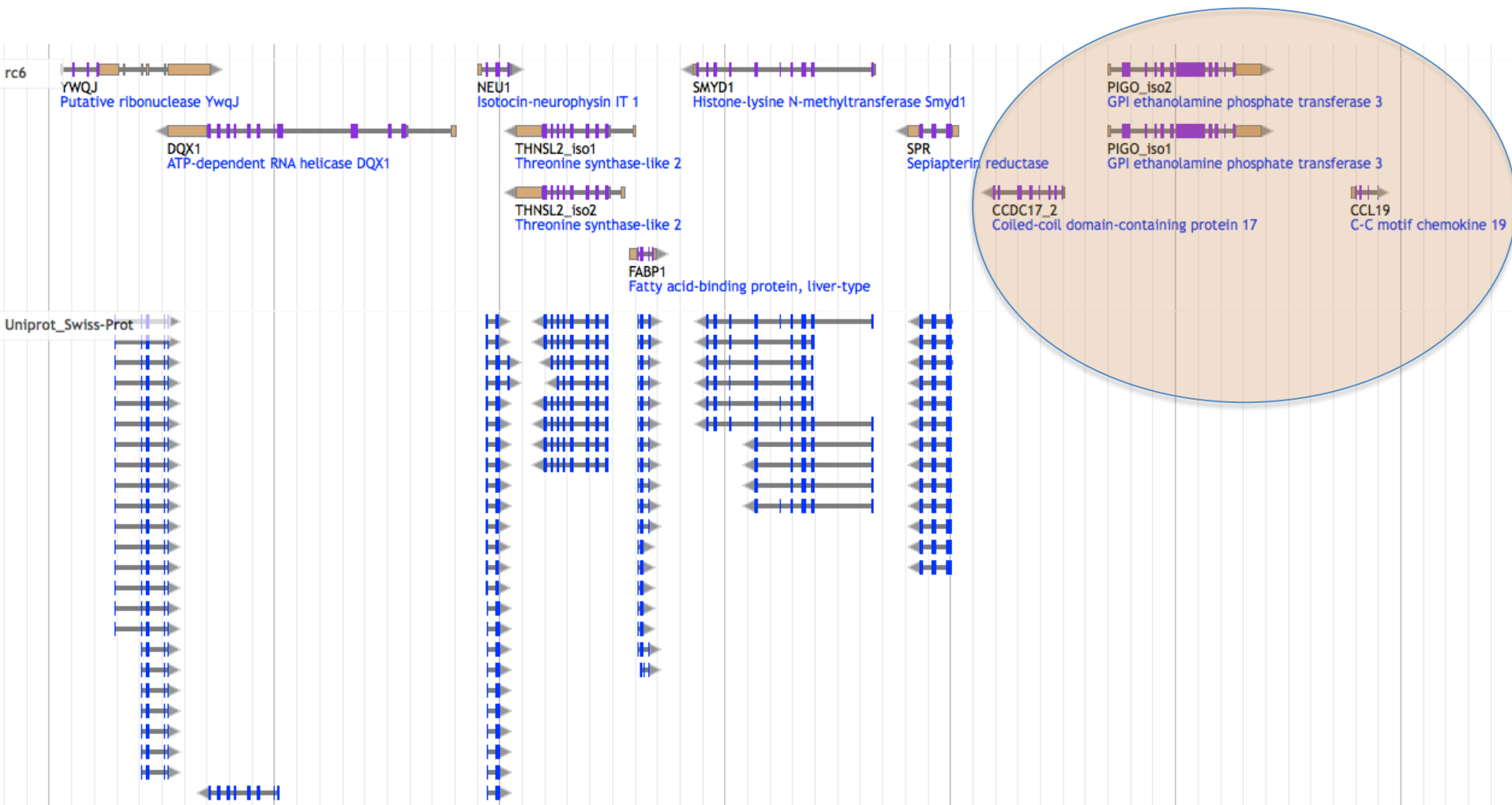
Strengths

- Fairly fast and easy
- Allow gene naming

Limits

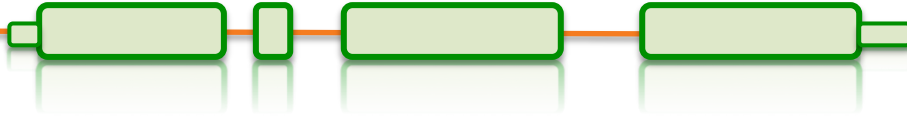
- Orthology not certain - best blast-hit does not equal orthologous!
- Bias due to well conserved domains
- Best Hit (use as template) is not necessary the best annotated sequence to use => Could apply a prioritization rule (Human first, then mouse, etc).

Aligned proteins ≠ Blast hits



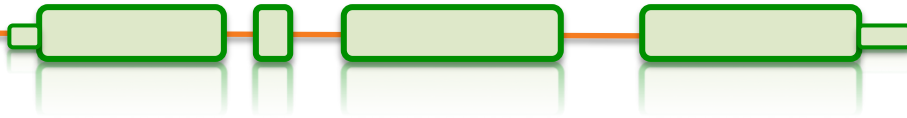
Blast-based approach

The NBIS annotation service

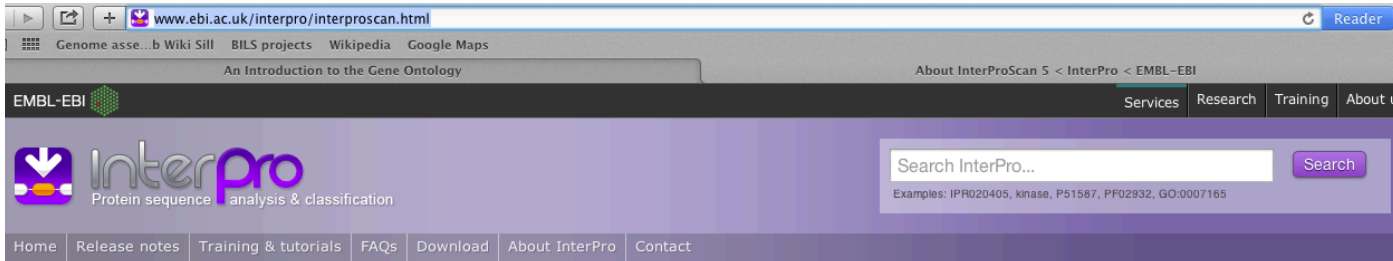


Blast-based annotation are tightly dependent to the quality of the annotation

- Gene Fusion
- Gene split
- Gene Partial (Well conserved domain)



- Annotate the sequences functionally using Interproscan



About InterProScan

What is InterProScan?

InterProScan is the software package that allows sequences (protein and nucleic) to be scanned against InterPro's signatures. Signatures are predictive models, provided by several different databases (referred to as member databases), that make up the InterPro consortium.

The software is available:


- As a web-based tool, using the sequence search box on the [InterPro homepage](#), for the analysis of single protein sequences (also available in the [EBI tool section](#))
- Programmatically via Web services that allow up to 25 sequences to be analysed per request (both [SOAP](#) and [REST](#)-based services are available)
- As a downloadable package for local installation from the EBI's FTP server, for instructions see the [detailed documentation](#) pages.

InterProScan is run regularly against UniProtKB and the results are made available via the InterPro website.

More information

For more information, and for instructions on how to obtain, install and run InterProScan, please see the [detailed documentation](#) pages.

Publications

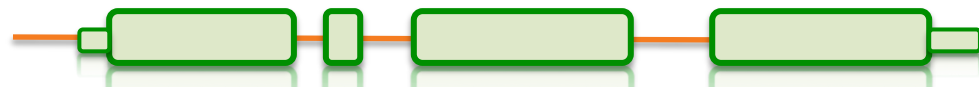


InterProScan 5: genome-scale protein function classification
 Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter
Bioinformatics, Jan 2014
 (doi:10.1093/bioinformatics/btu031)
[HTML](#) - [PDF \(324Kb\)](#)

Jones, P. et al. InterProScan5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240 (2014).

Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N., Apweiler R., et al. (2005). InterProScan: protein domains identifier. *Nucleic Acids Res.* 33, W116–W120. 10.1093/nar/gki442

Interproscan



Contents and coverage of InterPro 57.0

InterPro protein matches are now calculated for all UniProtKB and UniParc proteins. The following statistics are for all UniProtKB proteins. InterPro release 57.0 contains [29175](#) entries (last entry: [IPR033481](#)), representing:

- F** Family (19597)
- D** Domain (8393)
- R** Repeat (284)
- S** Sites
 - ↳ Active site (129)
 - ↳ Binding site (75)
 - ↳ Conserved site (680)
 - ↳ PTM (17)

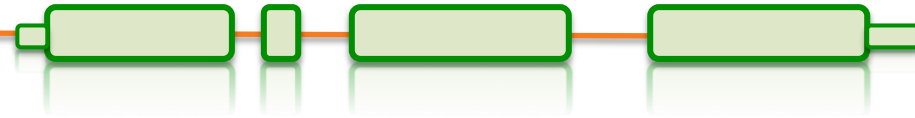
InterPro cites 47947 publications in PubMed.

Member database information

Signature database	Version	Signatures*	Integrated signatures**
CATH-Gene3D	3.5.0	2626	1725
HAMAP	201511.02	2045	2037
PANTHER	10.0	95118	5179
PIRSF	3.01	3285	3223
PRINTS	42.0	2106	2002
PROSITE patterns	20.119	1309	1291
PROSITE profiles	20.119	1136	1108
Pfam	29.0	16295	15700
ProDom	2006.1	1894	1130
SMART	7.1	1312	1265
SUPERFAMILY	1.75	2019	1416
TIGRFAMs	15.0	4488	4454

* Some signatures may not have matches to UniProtKB proteins.

** Not all signatures of a member database may be integrated at the time of an InterPro release



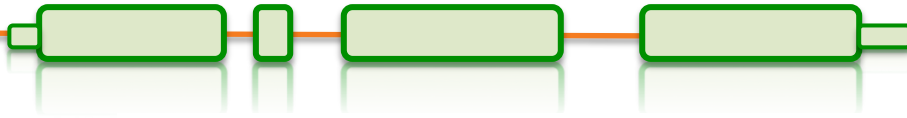
Sequence database	Version	Count	Count of proteins matching	
			any signature	integrated signatures
UniProtKB	2016_04	64237017	53062429 (82.6%)	51384050 (80.0%)
UniProtKB/TrEMBL	2016_04	63686057	52526761 (82.5%)	50852748 (79.8%)
UniProtKB/Swiss-Prot	2016_04	550960	535668 (97.2%)	531302 (96.4%)

InterPro2GO

Total number of GO terms mapped to InterPro entries - 31668

Interproscan results

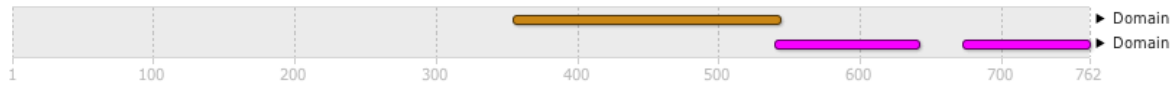
The NBIS annotation service



Protein family membership

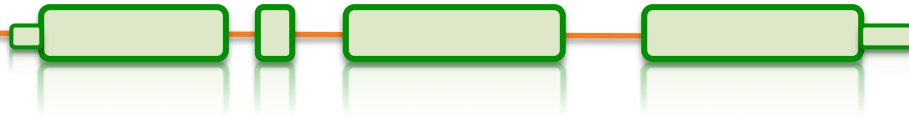
- [-] F Crotonase superfamily (IPR001753)
- [-] F Fatty acid oxidation complex, alpha subunit, mitochondrial (IPR012803)

Domains and repeats



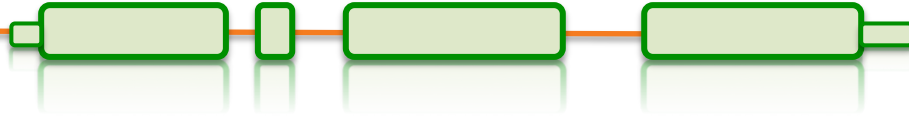
Detailed signature matches

F IPR001753	Crotonase superfamily	PF00378 (ECH)
F IPR012803	Fatty acid oxidation complex, alpha subunit, mitochondrial	TIGR02441 (fa_ox_al...)
D IPR016040	NAD(P)-binding domain	G3DSA: 3.40.50...
D IPR006176	3-hydroxyacyl-CoA dehydrogenase, NAD binding	PF02737 (3HCDH_N)
D IPR008927	6-phosphogluconate dehydrogenase, C-terminal-like	SSF48179
D IPR013328	Dehydrogenase, multihelical	G3DSA: 1.10.10...
D IPR006108	3-hydroxyacyl-CoA dehydrogenase, C-terminal	PF00725 (3HCDH)
? no IPR	Unintegrated signatures	G3DSA: 3.90.22... PTHR23309 SSF51735 SSF52096

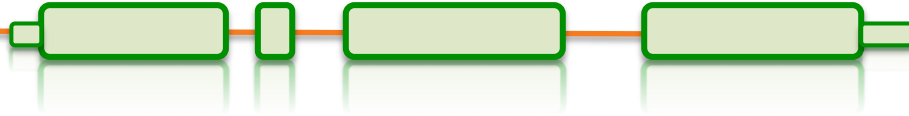


Output: TSV, XML, SVG, etc

gene-2.44-mRNA-1	a9deba5837e2614a850c7849c85c8e9c	447	Pfam	PF02458	Transferase family	98	425
1.4E-15	T 31-10-2015	IPR003480	Transferase	GO:0016747			
gene-0.13-mRNA-1	61882f1a46b15c8497ed9584a0eb1a35	459	Pfam	PF01490	Transmembrane amino acid transporter protein	49	439
2.0E-39	T 31-10-2015	IPR013057	Amino acid transporter, transmembrane				
gene-1.4-mRNA-1	b867bbb377084bba6ea84dcda9f27f4e	511	SUPERFAMILY	SSF103473	Major facilitator superfamily domain, general substrate transporter	42	481
4.19E-50	T 31-10-2015	IPR016196					
gene-1.4-mRNA-1	b867bbb377084bba6ea84dcda9f27f4e	511	Pfam	PF07690	Major Facilitator Superfamily	67	447
3.5E-30	T 31-10-2015	IPR011701	Major facilitator superfamily	GO:0016021 GO:0055085			



- Run Interproscan on the protein fasta file created by maker
- Use Maker-supplied scripts to merge the interproscan-results to the Maker annotations.gff file



- Looks for conserved domains, so might be more reliable than blast?
- How to go from conserved domains to assigning a function for your protein?

Another way : use the (mostly) commercial alternative



- Combines a blast-based search with a search for functional domains
- Blast at NCBI -> picks out GO terms based on blast hits and uniprot -> statistical significance test -> done!
- Blast2Go relies entirely on sequence similarity ... but InterProScan searches can also be launched within blast2go
- Command line tool or Plugin for Geneious or CLC bio Workbench

=> Contain nice downstream analysis/visualization components



/Users/hobbe/Documents/Artemis_files_current/blast2go_20101001_0816.dat - Blast2GO V.2.4.4

File Blast Mapping Annotation Analysis Statistics Select Tools View Info

GO:0007067,GO:0016021 transport;binding;apoptosis SPO_2518,DDX18_HUMAN

nr	sequence name	seq description	length	#...	min. eValue	sim mean	#G...	GO IDs	Enzyme	InterPro	
<input type="checkbox"/>	3884	gene_3884 GeneMar...	c6 transcription	977	20	1.0E-171	59.85%	7	F:transcription factor activity; F:zinc ion binding; P:regulation of transcription, DNA-dependent; C:transcription factor complex; F:transporter activity; C:membrane; P:transmembrane transport		IPR005829; IPR007219
<input type="checkbox"/>	3885	gene_3885 GeneMar...	hypothetical protein NFIA_039100 [Neosartorya fischeri NRRL 181]	312	20	1.0E-39	63.15%	1	C:viral capsid		no IPS match
<input type="checkbox"/>	3886	gene_3886 GeneMar...	sin3 complex subunit	870	20	0.0	73.2%	0			
<input type="checkbox"/>	3887	gene_3887 GeneMar...	mitochondrial intermembrane space translocase subunit	87	20	1.0E-40	88.55%	5	F:metal ion binding; P:protein import into mitochondrial inner membrane; C:mitochondrial inner membrane; C:mitochondrial intermembrane space protein transporter complex; P:transmembrane transport		IPR004217; PTHR11038 (PANTHER); PTHR11038:SF8 (PANTHER)
<input type="checkbox"/>	3888	gene_3888 GeneMar...	lysyl-tRNA synthetase	592	20	0.0	73.55%	7	C:cytoplasm; P:auxin biosynthetic process; F:nucleic acid binding; F:lysine-tRNA ligase activity; P:lysyl-tRNA aminoacylation; F:ATP binding; P:lysine biosynthetic process	EC:6.1.1.6	IPR004364; IPR004365; IPR006195; IPR012340; IPR016027; IPR018149; IPR018150; G3DSA:3.30.930.10 (GENE3D); SSF5568 (SUPERFAMILY)
<input type="checkbox"/>	3889	gene_3889 GeneMar...	transcription factor conserved	1569	20	0.0	70.9%	0			
<input type="checkbox"/>	3890	gene_3890 GeneMar...	hypothetical protein [Aspergillus clavatus NRRL 1]	240	20	1.0E-51	56.25%	0			
<input type="checkbox"/>			udp-glc gal endoplasmic reticulum nucleotide						C:integral to membrane; C:endoplasmic reticulum membrane; P:transmembrane transport; P:carbohydrate transport		IPR013657; PTHR10778 (PANTHER)

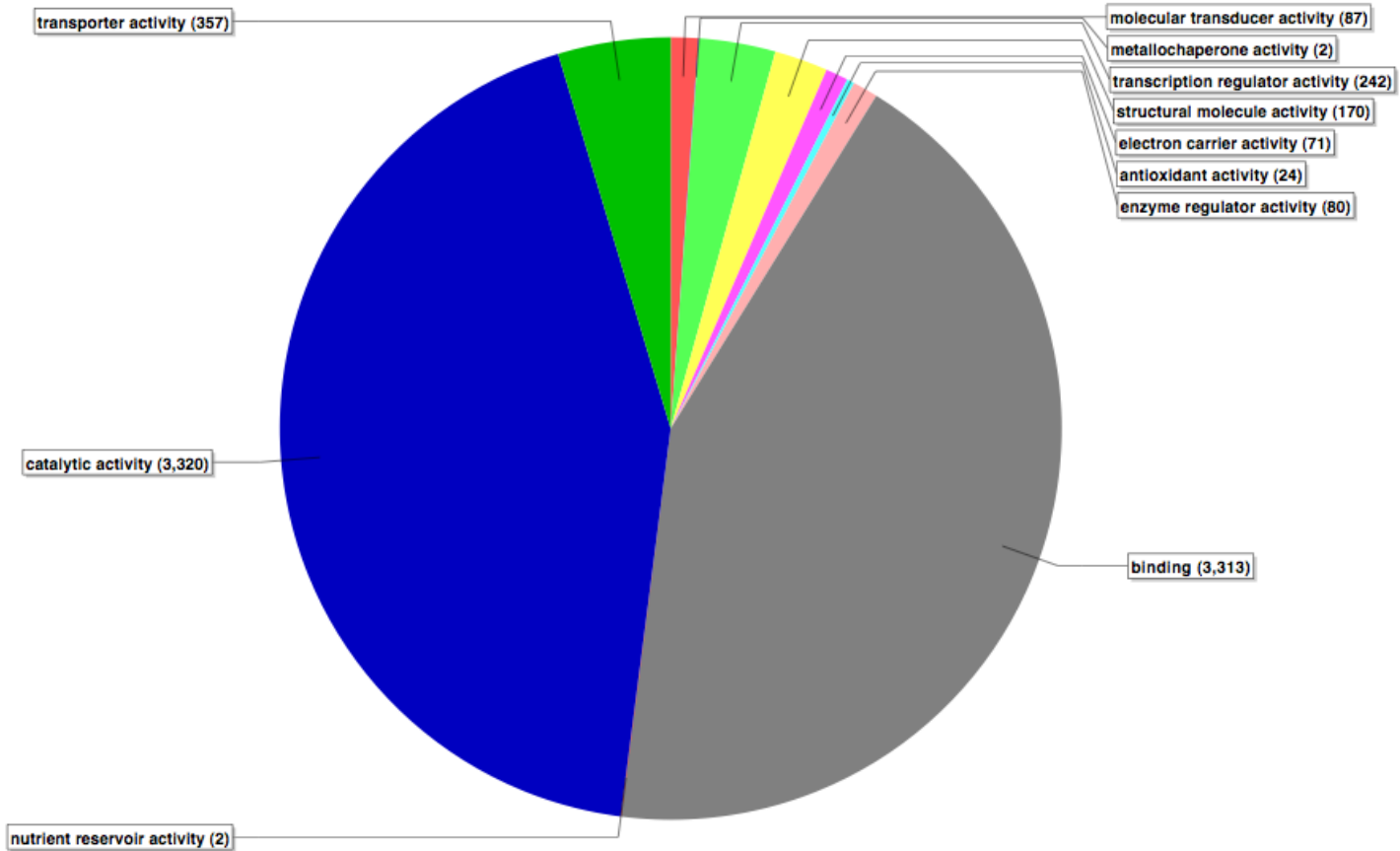
GO Graphs Application Messages Blast/IPS Results Statistics Kegg Maps

```

17:59 InterProScan for gene_8871|GeneMark.hmm|286_aa done.
17:59 -----
17:59 InterProScan Result:
17:59 InterProId: IPR001715
17:59 InterProName: Calponin-like actin-binding
17:59 InterProType: Domain
17:59 DB-Name: GENE3D - G3DSA:1.10.418.10
17:59 InterProId: IPR016146
17:59 InterProName: Calponin-homology
17:59 InterProType: Domain
17:59 DB-Name: SUPERFAMILY - SSF47576
17:59 InterProId: noIPR
17:59 InterProName: unintegrated
17:59 InterProType: unintegrated
17:59 DB-Name: PANTHER - PTHR19961
17:59 DB-Name: PANTHER - PTHR19961:SF9
    
```

Annotation already running

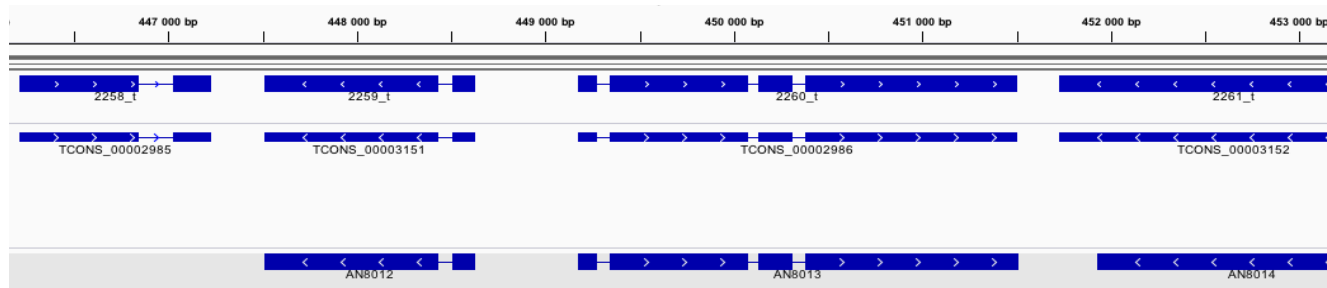
molecular_function Level 2





Liftovers are very useful for orthology determination

- Align the two genomes (Satsuma)
- Transfer annotations between aligned regions (Kraken)



The END

supplement

KEGG-mapping

file blast mapping Annotation Analysis Statistics Select Tools view info

GO:0007067,GO:0016021 transport;binding;apoptosis SPO_2518,DDX18_HUMAN

nr	sequence name	seq description	length	#...	min. eValue	sim mean	#G...	GO IDs	Enzyme	InterPro
		succinyl- synthetase subunit						F:ATP binding; F:succinate-CoA ligase (GDP-forming) activity; P:tricarboxylic acid cycle; C:succinate-CoA ligase		IPR003781; IPR005810

GO Graphs Application Messages Blast/IPS Results Statistics Kegg Maps

GLYCEROLIPID METABOLISM

Pathways

- Pentose phosphate pathway
- Fructose and mannose metabolism
- Butanoate metabolism
- Carbon fixation in photosynthetic organisms
- Lysine degradation
- Tyrosine metabolism
- Methane metabolism
- Glyoxylate and dicarboxylate metabolism
- Glycerolipid metabolism**
- Glutathione metabolism
- Selenoamino acid metabolism
- Phenylalanine metabolism
- Benzoate degradation via CoA ligation
- Valine, leucine and isoleucine biosynthesis
- Reductive carboxylate cycle (CO2 fixation)
- Galactose metabolism
- Phenylalanine, tyrosine and tryptophan biosynthesis
- N-Glycan biosynthesis
- Photosynthesis
- Drug metabolism - other enzymes
- Sulfur metabolism
- Fatty acid biosynthesis
- Inositol phosphate metabolism
- beta-Alanine metabolism
- Drug metabolism - cytochrome P450
- Pantothenate and CoA biosynthesis
- Biosynthesis of unsaturated fatty acids
- Cyanoamino acid metabolism
- Terpenoid backbone biosynthesis
- Histidine metabolism
- T cell receptor signaling pathway
- Tropane, piperidine and pyridine alkaloid biosynthesis
- One carbon pool by folate
- Pentose and glucuronate interconversions
- Phosphatidylinositol signaling system

Color	Enzyme	Sequences
red	ec:1.1.1.2 - alcohol dehydrogenase (NADP+)	gene_674 GeneMark.hmm 333_aa, gene_5801 GeneMark.hmm 312_aa
yellow	ec:2.3.1.158 - phospholipid:diacylglycerol acyltransferase	gene_2604 GeneMark.hmm 188_aa, gene_6532 GeneMark.hmm 505_aa
orange	ec:2.3.1.51 - 1-acylglycerol-3-phosphate O-acyltransferase	gene_176 GeneMark.hmm 429_aa, gene_6693 GeneMark.hmm 292_aa
green	ec:2.3.1.20 - diacylglycerol O-acyltransferase	gene_176 GeneMark.hmm 429_aa, gene_7213 GeneMark.hmm 521_aa, gene_8170 GeneMark.hmm 470_aa
blue	ec:2.3.1.15 - glycerol-3-phosphate O-acyltransferase	gene_886 GeneMark.hmm 748_aa, gene_2640 GeneMark.hmm 823_aa
pink	ec:1.1.1.72 - glycerol dehydrogenase (NADP+)	gene_3376 GeneMark.hmm 325_aa, gene_4577 GeneMark.hmm 326_aa
violet	ec:1.2.1.3 - aldehyde dehydrogenase (NAD+)	gene_2201 GeneMark.hmm 497_aa, gene_5247 GeneMark.hmm 502_aa, gene_5611 GeneMark.hmm 471_aa
light-red	ec:2.7.1.107 - diacylglycerol kinase	gene_5292 GeneMark.hmm 409_aa

Annotation already running