

Functional annotation of transcripts



Lucile Soler
SciLifeLab RNAseq workshop
November 2018

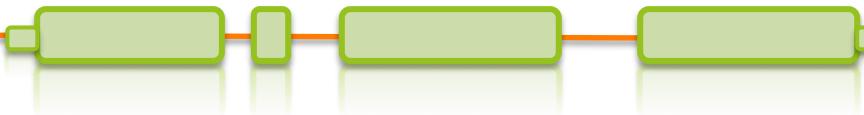
The accelerate logo consists of the word "accelerate" in a bold, black, sans-serif font, oriented diagonally upwards from bottom-left to top-right. To the left of the text is a small graphic element: a blue square with a white dot inside, followed by three orange dots of increasing size, and a short blue line segment extending from the top of the blue square.

1. Introduction to functional annotation

Functional annotation

- **What is functional annotation :**
 - Find out what the proteins/genes/transcripts do : function, domains ...
- **Why annotate RNAseq :**
 - To use annotated transcript for a first annotation (reduce noise, select annotated)
 - To use annotated transcript after annotation to for instance improve genome annotation
 - Know which genes are expressed depending on different tissues or life stages

Functional annotation – HOW?



- Experimentally
=> Mutants, knockout, etc.

Precise

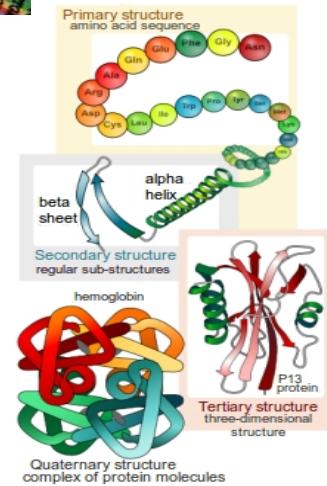


Mice homozygous for the diabetes 3J spontaneous mutation

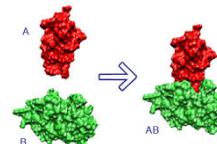
- Computationally
 - Sequence-based



- Structure based



- Protein-protein interaction data



limited accuracy

Methods - Sequence-based

- Based on similarity
=>Best blast hit

Q GLMDTAFEHIKATGGLTTESNYPYKGEDATCNS-KJ
GLM+ AFE+IK +GG+TTES YPY+ + TC++ +
S GLMENAFEYIKHSGGITTESAYPYRAANGTCDAVR

- Based on Motif/Patern
⇒Proscan, MEME, QuasiMotifFinder

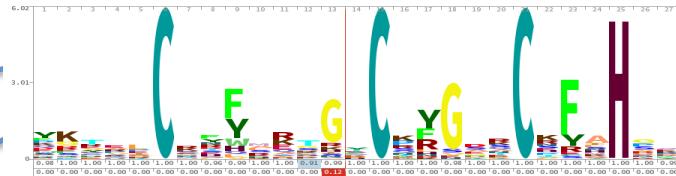
D-X-[KR]-P-{WYF}-X5

- Based on Profile (HMM or other statistical signature)

Whole sequence
e.g. Psi-BLAST*, PIRSF

domain
e.g. PFAM

Localization (e.g membrane, golgi, secreted)
e.g. SignalP, TMHMM



...
structural classification
e.g. SUPERFAMILY

- Based on evolutionary relationship (Orthology)
 - Gene are part of functional groups : KOG / COG
 - Based on synteny to check gene order *
 - ⇒ Whole genome alignment (lastZ)
 - (NBIS) Satsuma + kraken + custom script
 - Based on phylogeny to look at the evolution of set of genes*
 - ⇒ Quite complicated at large scale

* Can not be done on transcripts

- Similarity to known structures.

- Global structure-comparison
 - CATH and SCOP, the two most comprehensive structure-based family resources
- localized regions
 - might be relevant to function: clefts, pockets and surfaces
- active-site residues (catalytic clusters and ligand-binding sites)
 - active-site residues is often more conserved than the overall fold
⇒PDBSiteScan

no single method is always successful

Functional annotation – HOW?

It is actually kind of complex...

- Multi-dimensional problem :
 - e.g. A protein can have a molecular function, a cellular role, and be part of a functional complex or pathway
- Molecular function can be illustrated by multiple descriptive levels
 - (e.g. '**enzyme**' category versus a more specific '**protease**' assignment).

Functional annotation – HOW?



It is actually kind of complex...

- Similarities (structural or in sequence)  function.
 - Similar sequence but different function (new domain => new combination => different function)
 - Different sequence may have same function (convergence) : Profiles helpful
 - Two proteins may have a similar fold but different functions
- Looks for conserved domains more reliable than whole sequence ?
 - How to go from conserved domains to assigning a function for your protein?

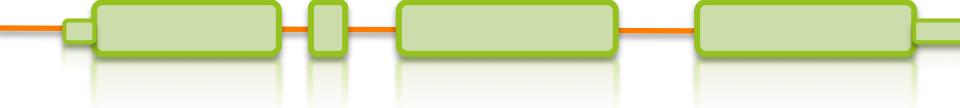
=> Importance to gathering as much information as possible

Sequence-based methods

- The most used (popular)
- Quick
- Easy to use
- Accurate (>70%)

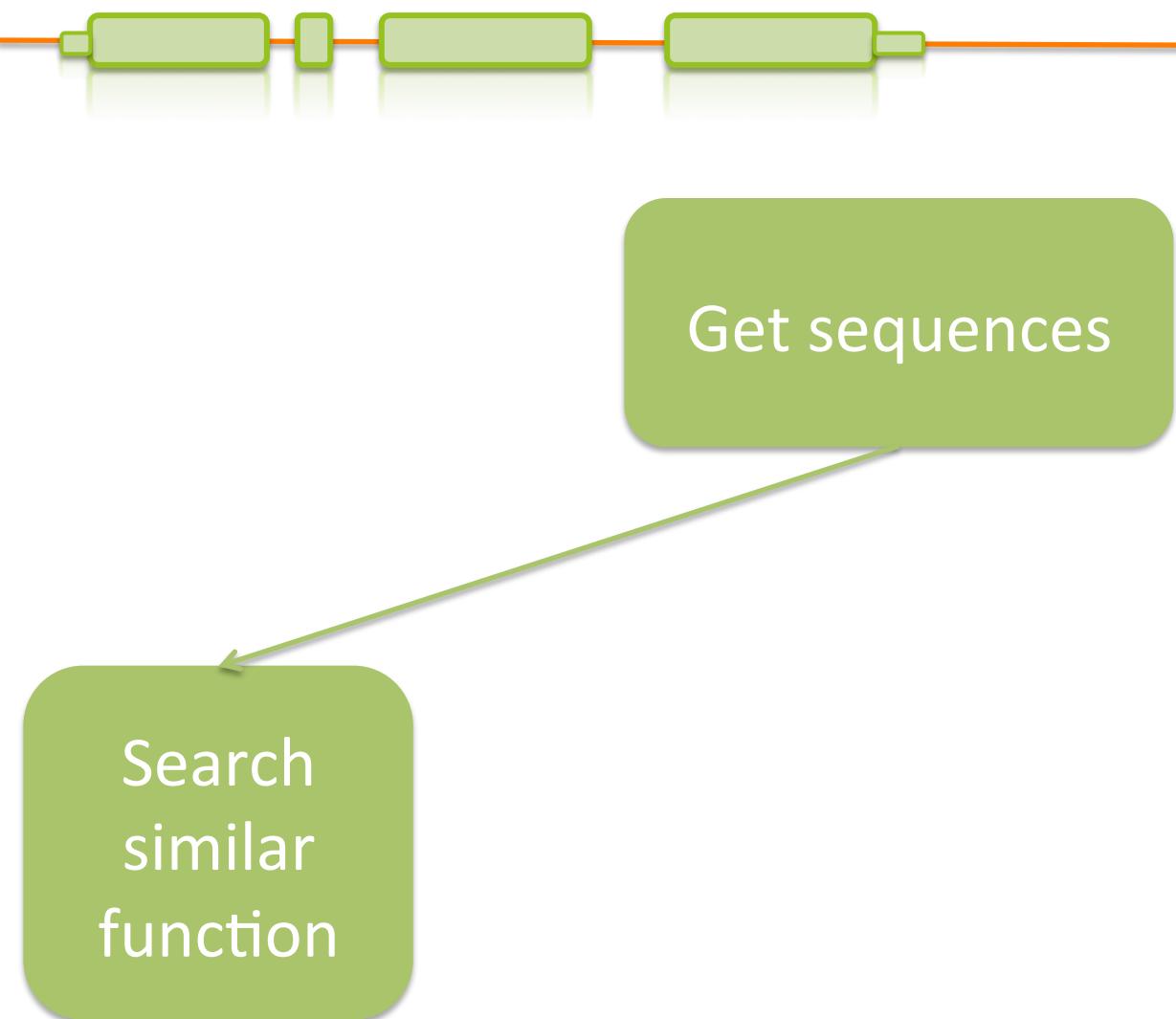
Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM: Towards fully automated structure-based function prediction in structural genomics: a case study. *J Mol Biol.* 2007, 367: 1511-1522. 10.1016/j.jmb.2007.01.063.

- Many resources: even structural domains information
- Less computationally demanding



2. Blast based approach

Functional annotation – HOW?



Get sequences

Search
similar
function

Blast-based
approach

Blast-based approach



Annotate the sequences functionally using Blast

Choose
database

Uniprot	Swissprot
exhaustive	reliable

Blast-based approach

Annotate the sequences functionally using Blast

Choose
database

Fasta
aa

blast



! Minimum Threshold

Blast-based approach

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).

Blast-based approach

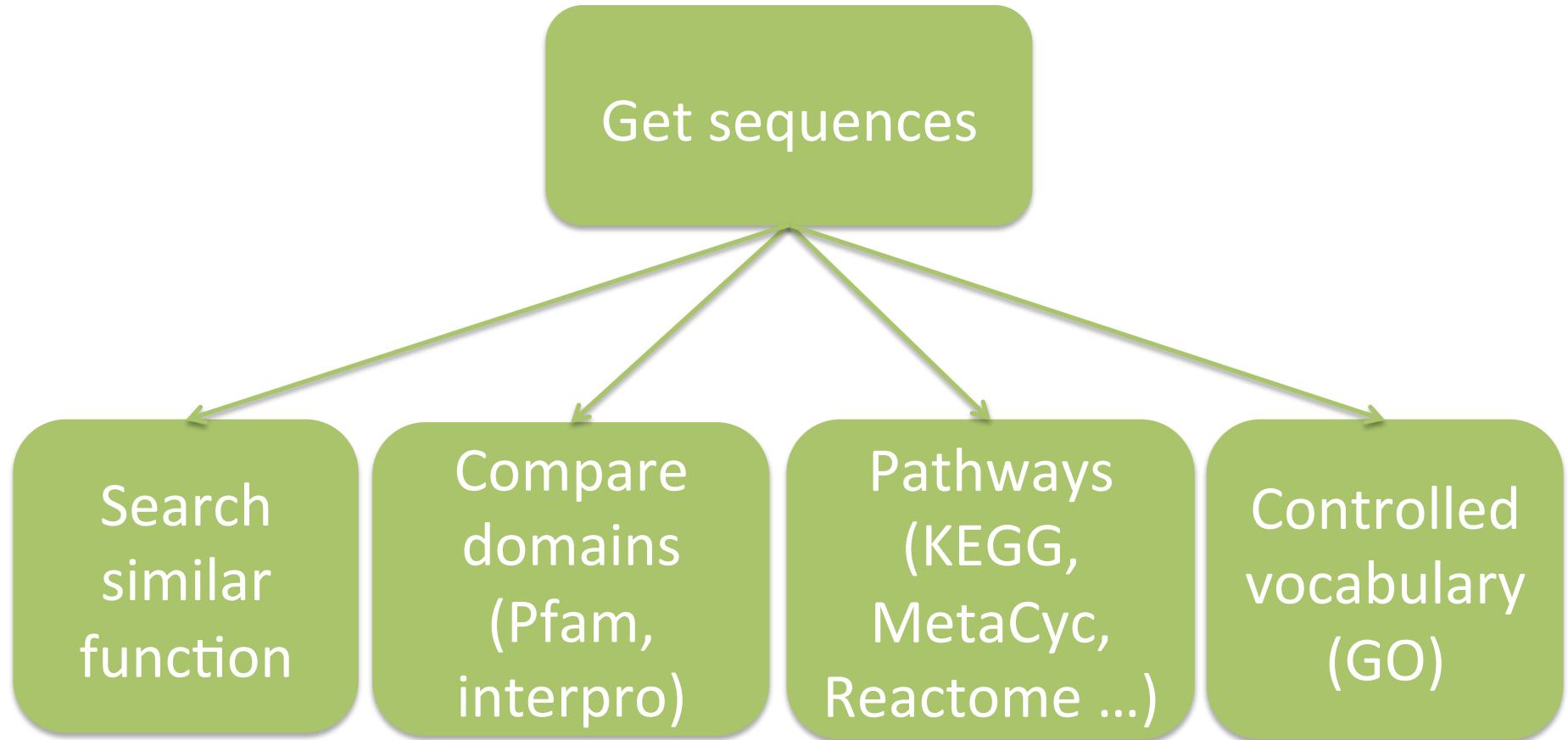
Blast-based annotation are tightly dependent to the quality of the transcript assembly

- Gene Fusion
- Gene split
- Gene Partial (Well conserved domain)
- Over prediction
- Wrong ORF



3. Domains/profiles/patterns approach

Functional annotation – HOW?



Databases

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.

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.

Gene 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:

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

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

More than 60 000 terms

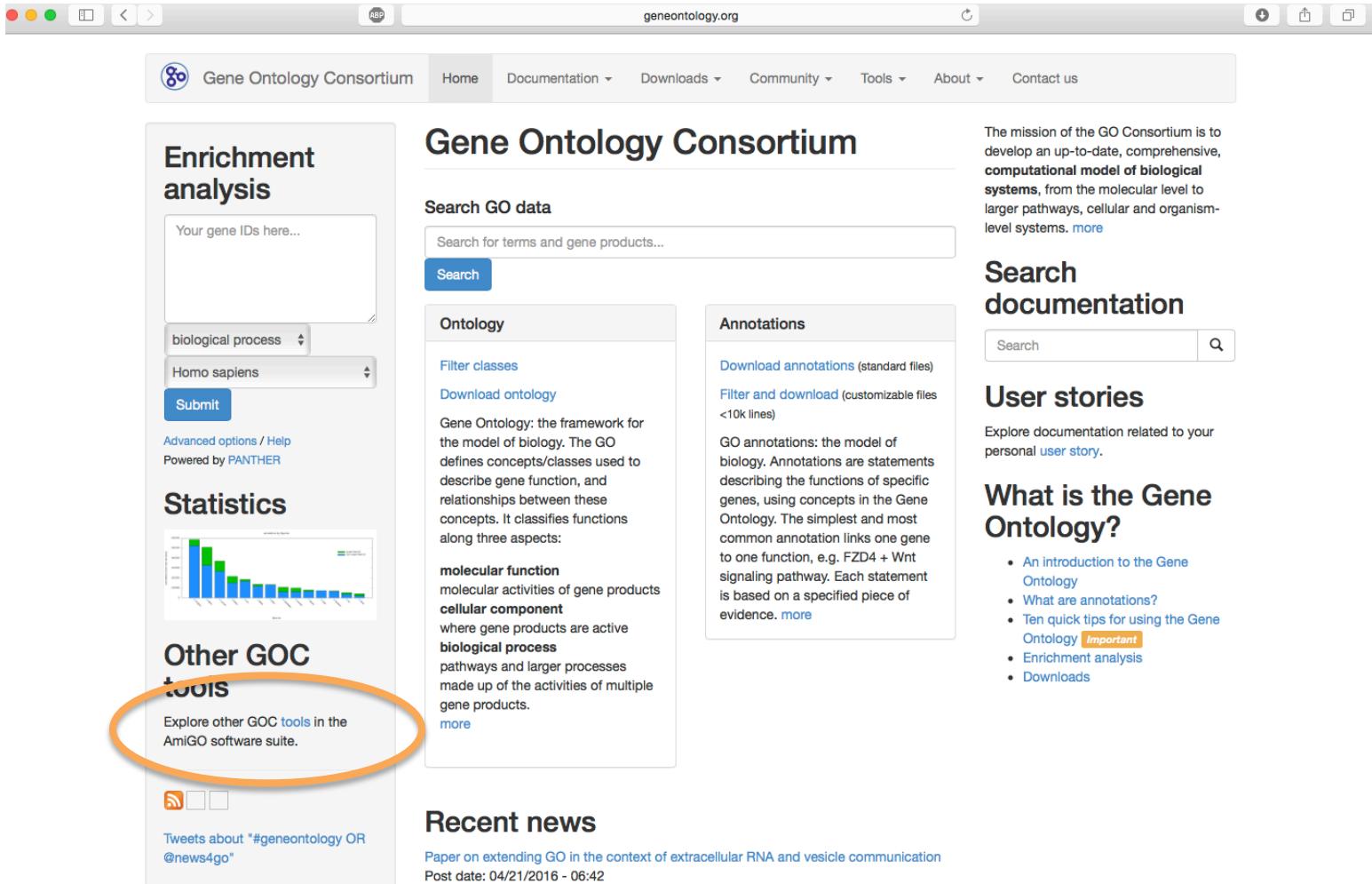
pathways and larger processes made up of the activities of multiple gene products.

molecular activities of gene products

where gene products are active

Gene Ontology

<http://www.geneontology.org/>



The screenshot shows the homepage of the Gene Ontology Consortium. At the top, there's a navigation bar with links for Home, Documentation, Downloads, Community, Tools, About, and Contact us. Below the navigation is a main content area.

Enrichment analysis section:

- A text input field labeled "Your gene IDs here..."
- A dropdown menu set to "biological process".
- A dropdown menu set to "Homo sapiens".
- A blue "Submit" button.
- Links for "Advanced options / Help" and "Powered by PANTHER".

Statistics section:

- A small bar chart.

Other GOC tools section (circled in orange):

- Text: "Explore other GOC tools in the AmiGO software suite."
- Social media icons for RSS, GitHub, and Bitbucket.
- Tweets about "#geneontology OR @news4go" (with a count of 1).

Gene Ontology Consortium section:

Gene Ontology Consortium

Search GO data

Search for terms and gene products...

Ontology

- Filter classes
- Download ontology
- Text: "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
 - cellular component
 - biological process
- Text: "pathways and larger processes made up of the activities of multiple gene products." [more](#)

Annotations

Download annotations (standard files) [<10k lines](#)

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](#)

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

3. Tools

Tools

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
Annocrypt	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 GO and other domains Commercial !

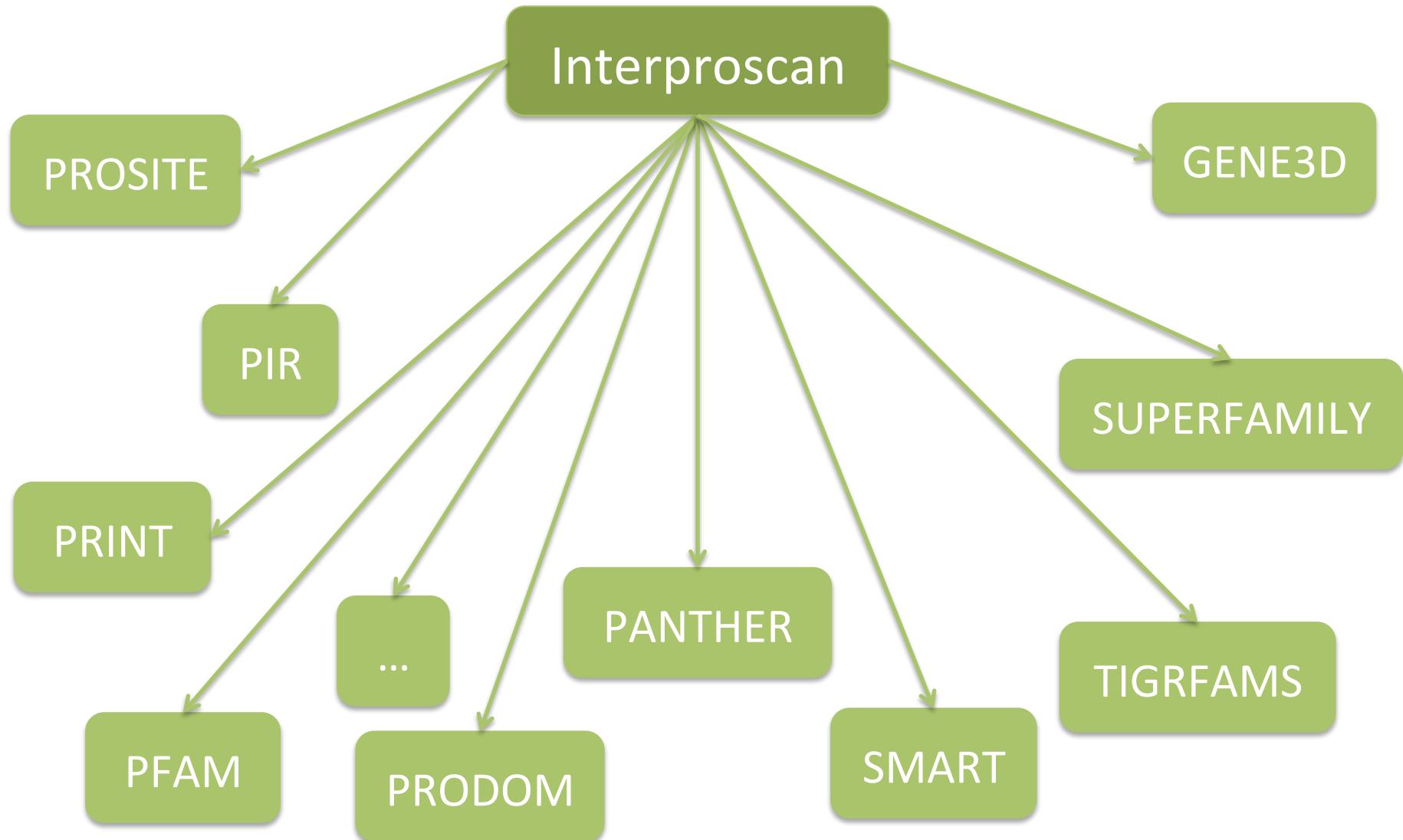


Interproscan

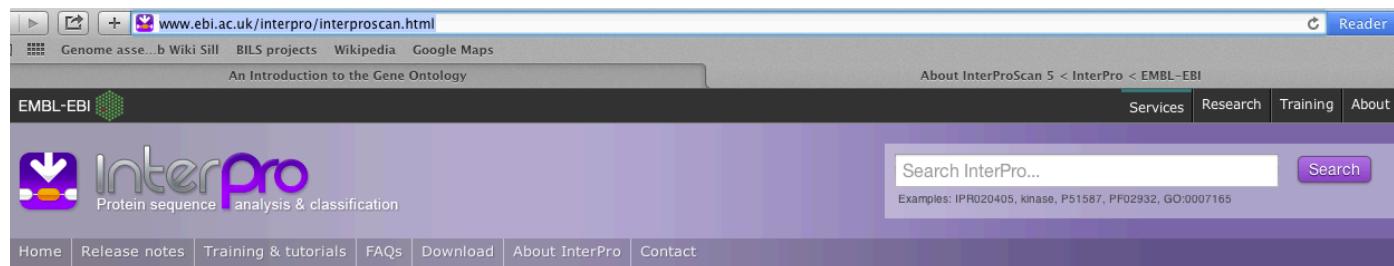
“InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites.

To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium.”

<https://www.ebi.ac.uk/interpro/about.html>



- Annotate the sequences functionally using Interproscan



The screenshot shows the EBI InterPro homepage. At the top, there's a navigation bar with links like 'Home', 'Release notes', 'Training & tutorials', 'FAQs', 'Download', 'About InterPro', and 'Contact'. Below the navigation is a search bar with placeholder text 'Search InterPro...' and a 'Search' button. To the right of the search bar, it says 'Examples: IPR020405, kinase, P51587, PF02932, GO:0007165'. The main content area features the 'InterPro' logo and the tagline 'Protein sequence analysis & classification'. There are also links for 'Services', 'Research', 'Training', and 'About us'.

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.

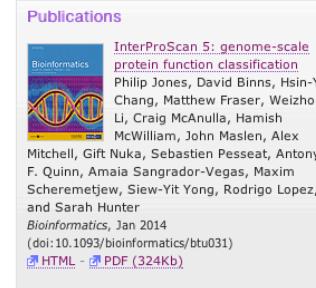
The software is available:

- 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, Giff Nuka, Sébastien Pesquet, 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

Contents and coverage of InterPro 62.0

InterPro protein matches are now calculated for all UniProtKB and UniParc proteins. The following statistics are for all UniProtKB proteins.

InterPro release 62.0 contains [29930](#) entries (last entry: [IPR034768](#)), representing:

 Family (19869)

 Domain (8868)

 Repeat (282)

 Sites

 ↳ Active site (132)

 ↳ Binding site (76)

 ↳ Conserved site (686)

 ↳ PTM (17)



Member database information

Signature database	Version	Signatures*	Integrated signatures**
CATH-Gene3D	4.1.0	2737	1198
CDD	3.14	11273	1526
HAMAP	201701.18	2160	2160
PANTHER	11.1	91538	5923
Pfam	30.0	16306	15710
PIRSF	3.01	3285	3222
PRINTS	42.0	2106	1986
ProDom	2006.1	1894	1131
PROSITE patterns	20.132	1309	1289
PROSITE profiles	20.132	1174	1142
SFLD	2	480	146
SMART	7.1	1312	1265
SUPERFAMILY	1.75	2019	1461
TIGRFAMs	15.0	4488	4450

* 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

Other sequence features

Coils Phobius SignalP TMHMM

Interproscan

Sequence database	Version	Count	Count of proteins matching	
			any signature	integrated signatures
UniProtKB	2017_03	80758400	71118703 (88.1%)	64919649 (80.4%)
UniProtKB/TrEMBL	2017_03	80204459	70576370 (88.0%)	64384952 (80.3%)
UniProtKB/Swiss-Prot	2017_03	553941	542333 (97.9%)	534697 (96.5%)

InterPro2GO

Total number of GO terms mapped to InterPro entries - 32178

Not integrated signatures = signature not yet curated or do not reach InterPro's standards for integration

pathway information available as well:

- KEGG
- MetaCyc
- Reactome
- UniPathway

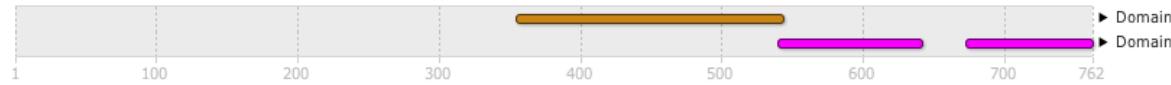
Interproscan results



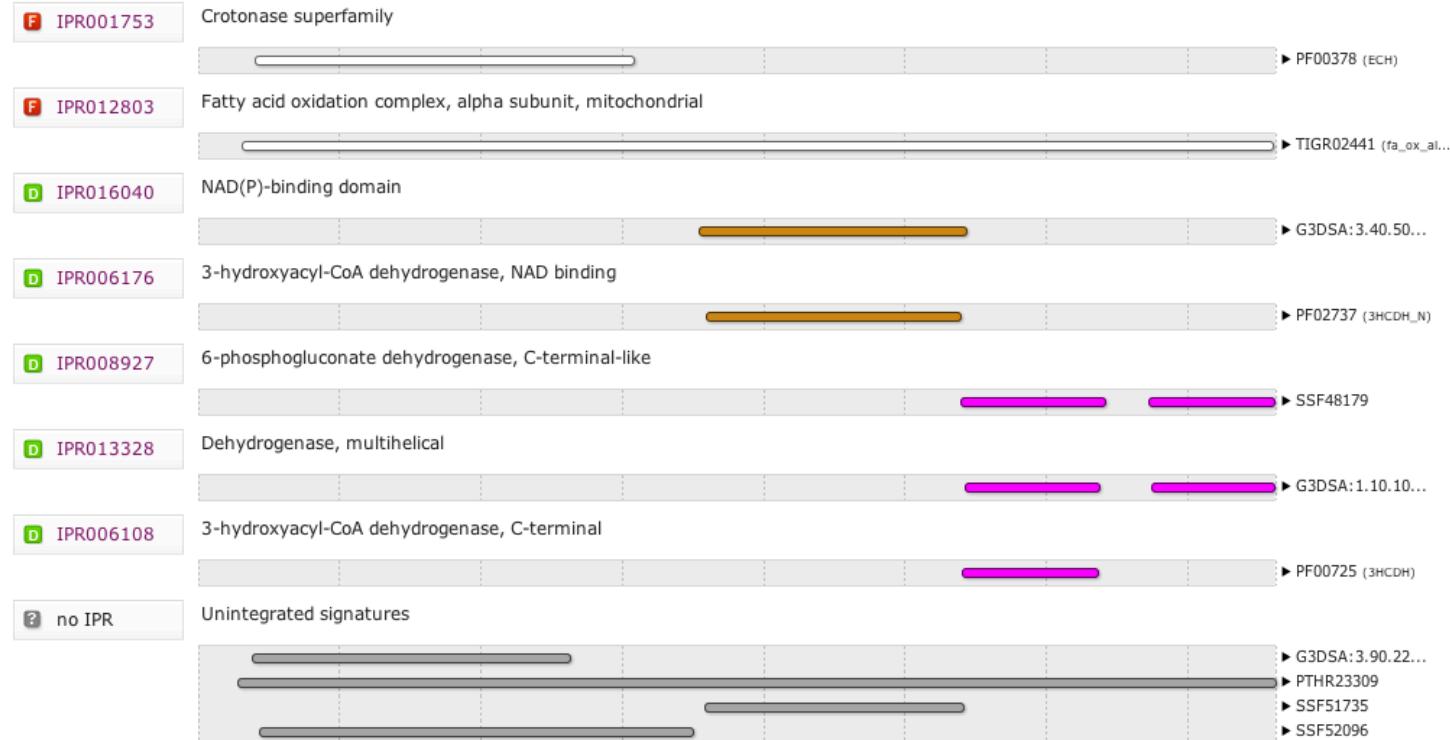
Protein family membership

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

Domains and repeats



Detailed signature matches



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 42 481
4.19E-50 T 31-10-2015 IPR016196 Major facilitator superfamily domain, general substrate transporter

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

Scripts exist to merge the interproscan-results to the structural annotation gff file

Trinotate

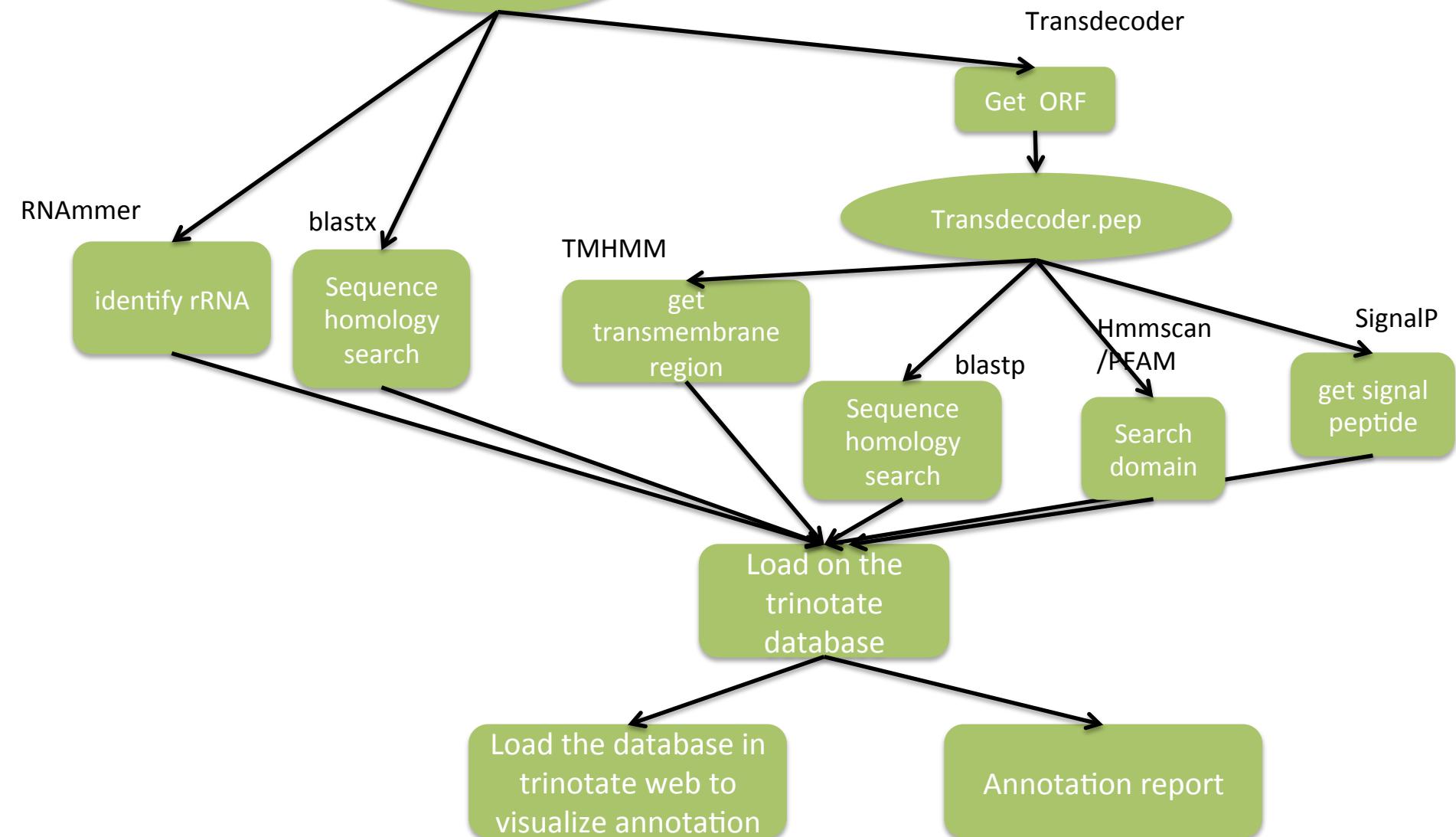
Trinotate

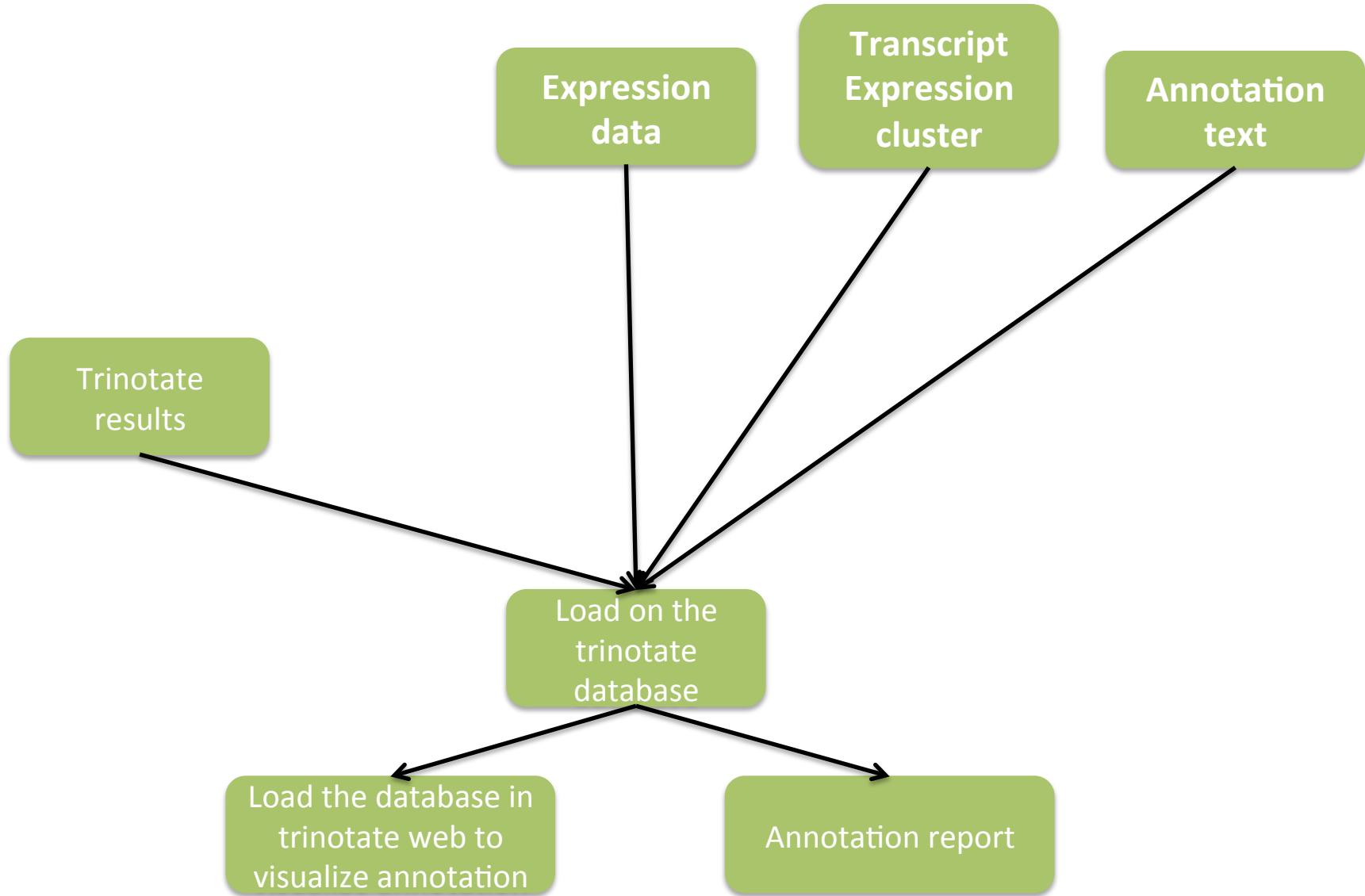
- Trinotate is a suite tools that was created to annotate specifically Trinity output
- Can also work with any fasta file if suitable inputs are available
- Now exists in pipeline



Get/create the SQLite database

Trinotate retrieve uniprot and pfam database
that will be needed later
It also create a trinotate database that will be populate later





Trinotate output

- Can create a report file (tabulated file)
- Number of columns depends on what you integrate in your database, if you integrate more blast or expression data you will have more columns

```
#gene_id      transcript_id  sprot_Top_BLASTX_hit  RNAMMER prot_id prot_coords
sprot_Top_BLASTP_hit  Pfam  SignalP TmHMM  eggnog Kegg  gene_ontology_blast
gene_ontology_pfam  transcript  peptide

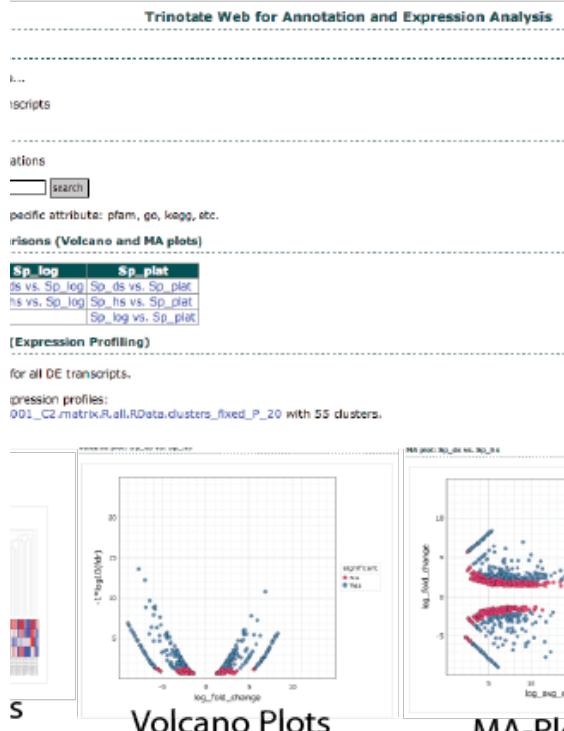
TRINITY_DN6975_c0_g2
TRINITY_DN6975_c0_g2_i1
tr|B4R0X8|B4R0X8_DROSI^tr|B4R0X8|B4R0X8_DROSI^Q:559-92,H:1-156^100%ID^E:8.42e-94^.^.

.
.
TRINITY_DN6975_c0_g2_i1.p1
89-664[-]
tr|B4R0X8|B4R0X8_DROSI^tr|B4R0X8|B4R0X8_DROSI^Q:36-191,H:1-156^100%ID^E:4.89e-111^.^.
PF03066.15^Nucleoplasmin^Nucleoplasmin/nucleophosmin domain^41-147^E:9e-28
.
.
.
```

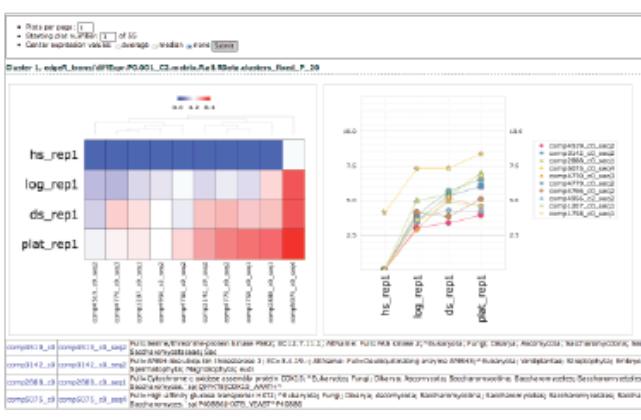
Trinotate output

trinotateWeb

tateWeb Entry Point

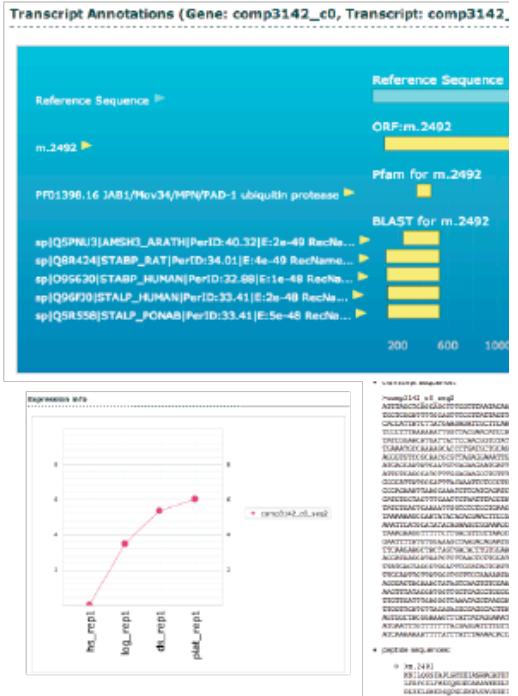


Clustered Expression Profiles



Very Early Release and
Just Scratching the Surface

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



Transc
Protei



4. Conclusion

- Functional annotation found
/!\ Transmission of error from databases !
Experimental check is good !
- Hypothetical protein / Uncharacterized protein
=> depends largely on conventional experiments.

Knowing the function is not enough: Chimp and human => 98% similarity

=> Knowledge of other parameters useful (pathway, positional and temporal regulation of genes)

THE END

<https://github.com/NBISweden/GAAS>

