Bioevo seminars

The true story behind the annotation of a pathway

Giovanni Dall'Olio,
IBE (UPF-CEXS)

Summary of the talk

- We have recently published two works:
  - Dall'Olio GM, Jassal B, Montanucci L, Gagneux P, Bertranpetit J, Laayouni H. **The annotation of the Asparagine N-linked Glycosylation pathway in the Reactome Database**. Glycobiology. 2011 Jan 2. PubMed PMID: 21199820.
  - Dall'Olio GM, Bertranpetit J, Laayouni H. **The annotation and the usage of scientific databases could be improved with public issue tracker software**. Database (Oxford). 2010 Dec 23;2010:baq035. Print 2010. PubMed PMID: 21186182; PubMed Central PMCID: PMC3011984.
- One is about the annotation of a pathway in a database, the other about reporting errors to databases

What can you learn from this talk

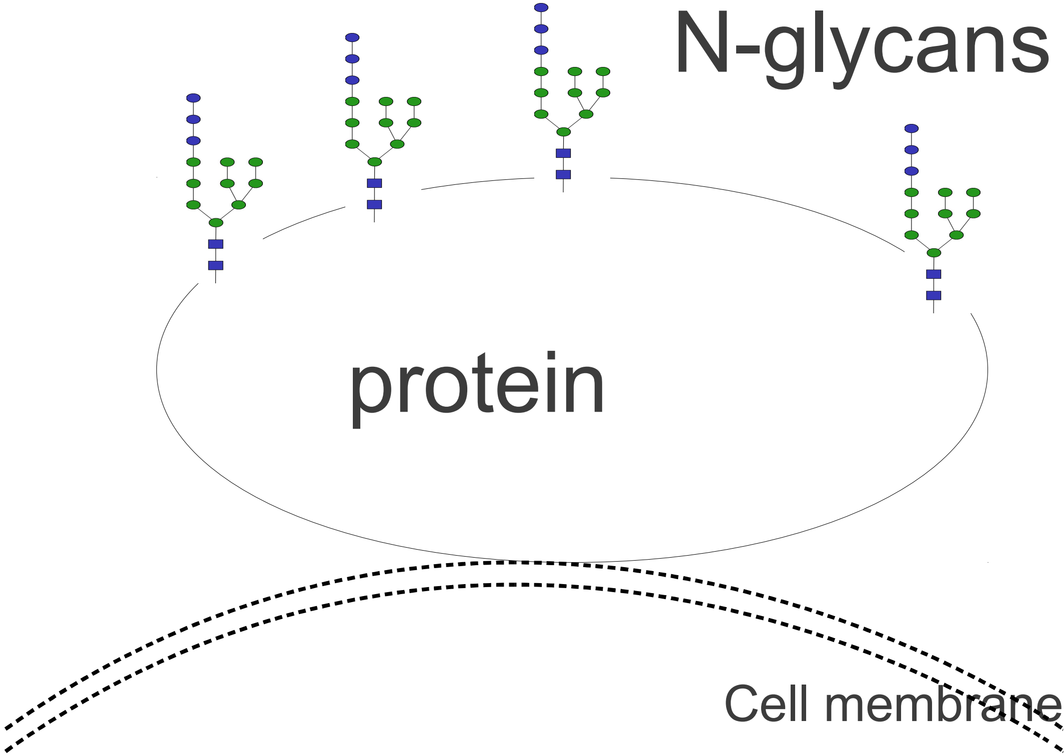
- The data annotated in scientific databases is not perfect, and can contain errors. Even when is correct, it can have multiple interpretations.

What can you learn from this talk

- The data annotated in scientific databases is not perfect, and can contain errors. Even when is correct, it can have multiple interpretations.
- Errors don't get fixed by themselves, and problems don't get solved alone. When you find something wrong, it is your duty to report it.

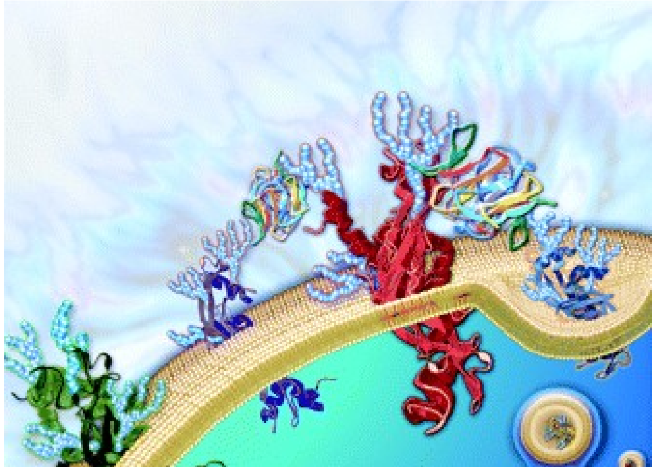
What is N-Glycosylation?

- One of the most important forms of protein modification
- A complex sugar composed by 14 units is attached to a protein, and later modified.



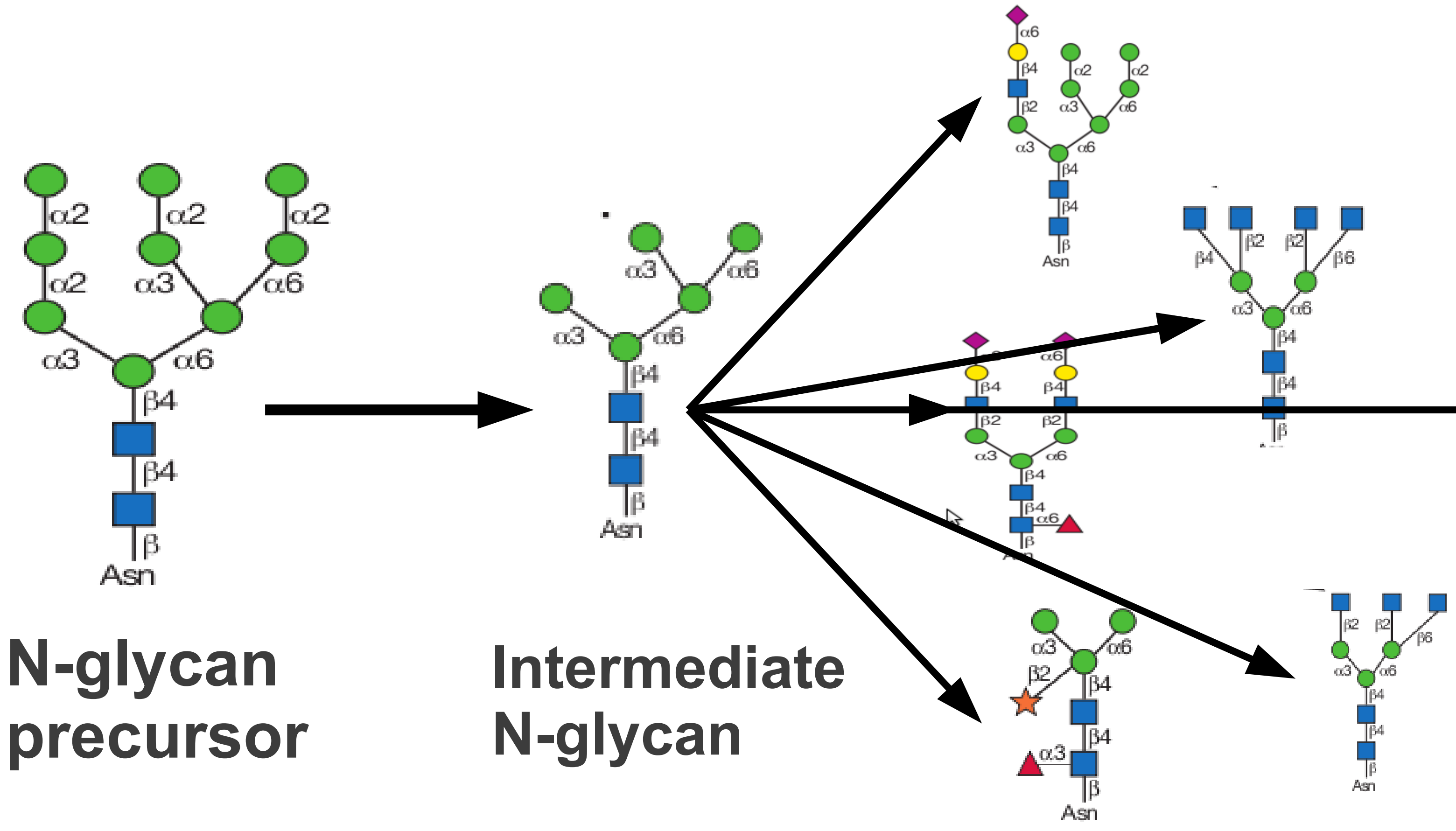
What is N-Glycosylation? (II)

- The surface of a cell is usually covered by N-glycosylated proteins
- It enhances solubility and is required for the proper folding.



N-Glycosylation – how does it works

- A common N-Glycan precursor is attached to a nascent protein, and then modified



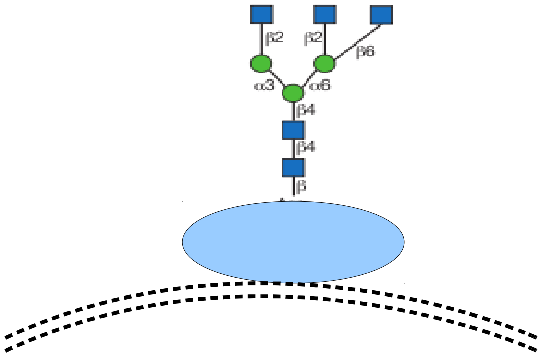
N-glycan precursor

Intermediate N-glycan

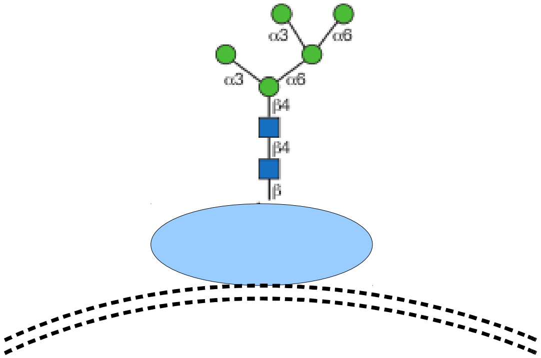
Millions of different N-glycans

Advanced N-Glycosylation

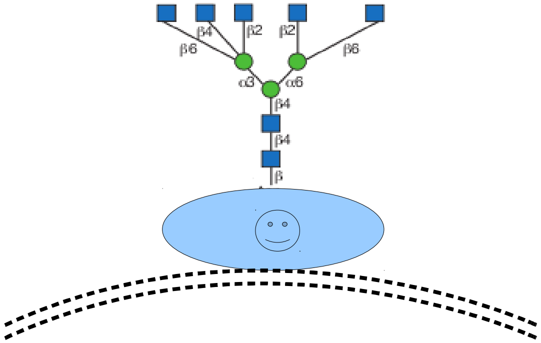
- The same protein can have different N-Glycosylation on different tissues and times



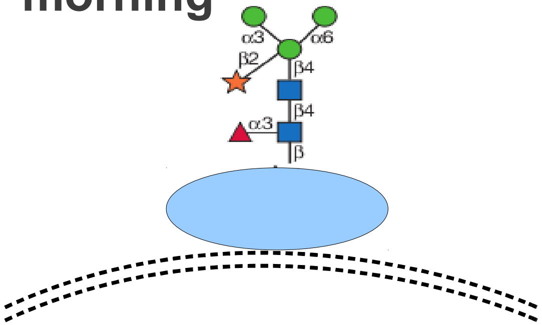
Protein A in the morning



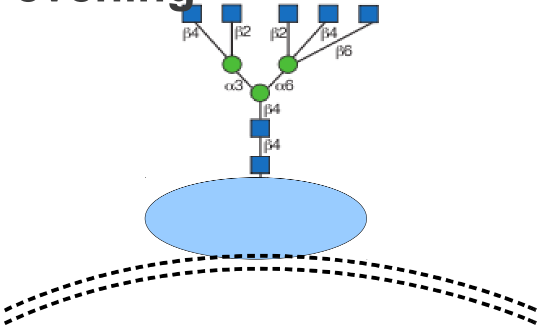
Protein A in the evening



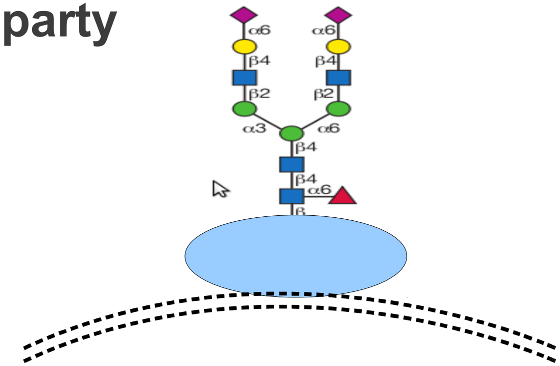
Protein A during a party



Protein A on an eritrocite



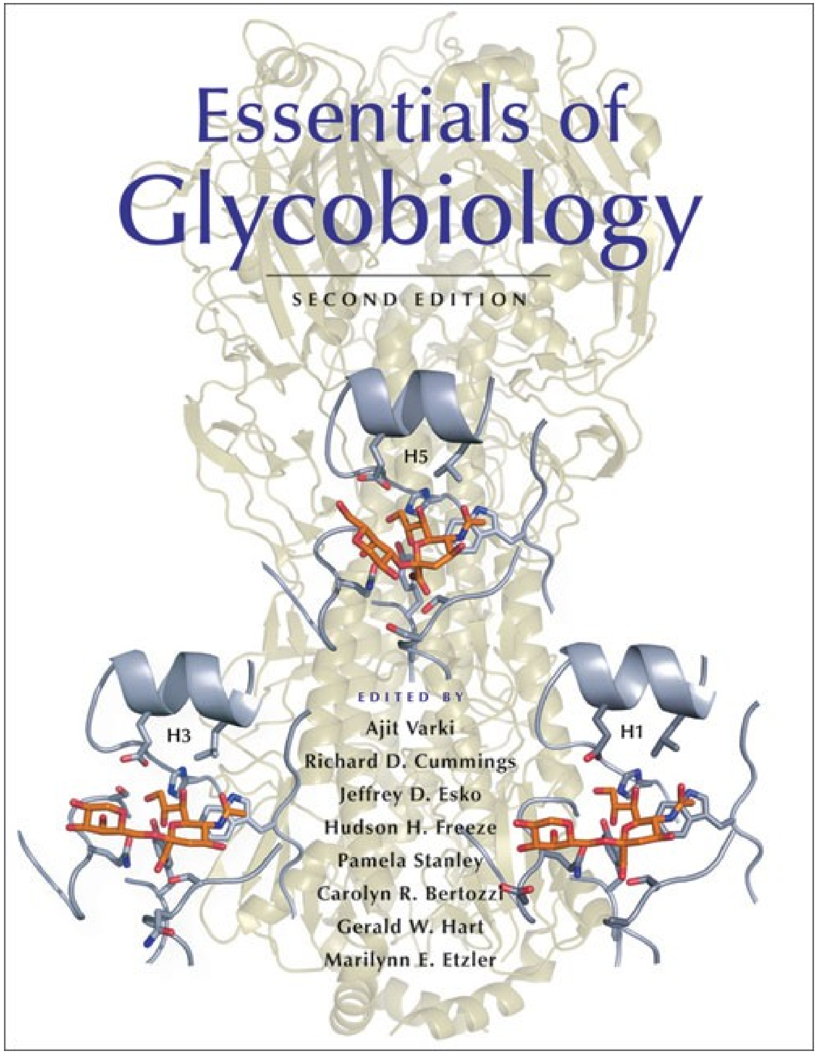
Protein A on a epithelial cell



Protein A on a dead cell

N-Glycosylation is a text-book example

- The first part of this pathway was characterized in the 1980's by random mutagenesis in yeast (the ALG mutants)
- N-glycosylation is important to biotechnologists because of its implications for drugs biosynthesis



Essentials of Glycobiology

SECOND EDITION

EDITED BY
Ajit Varki
Richard D. Cummings
Jeffrey D. Esko
Hudson H. Freeze
Pamela Stanley
Carolyn R. Bertozzi
Gerald W. Hart
Marilynn E. Etzler

Annotating the pathway

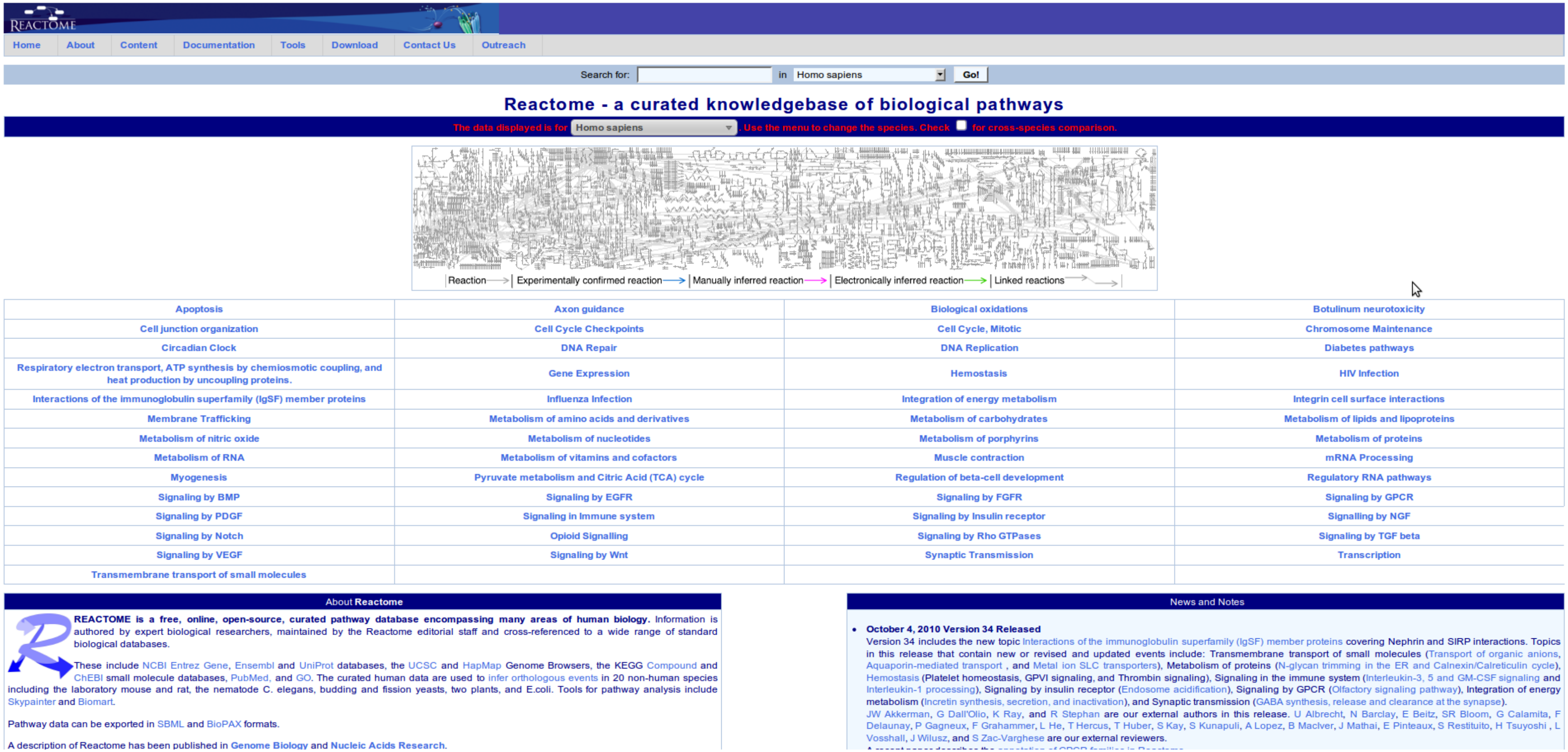
- My PhD thesis is about selection within the genes of N-Glycosylation, so I had to study the biology behind it

Annotating the pathway

- My PhD thesis is about selection within the genes of N-Glycosylation, so I had to study the biology behind it
- I was not happy with the current annotation of the pathway in other databases, so I decided to make a new annotation by myself, and publish it on a public database

Reactome

- A database for biological relevant pathways
  - TCA cycle, apoptosis, telomerases, influenza...



Reactome - a curated knowledgebase of biological pathways

The data displayed is for Homo sapiens. Use the menu to change the species. Check ☐ for cross-species comparison.

Apoptosis	Axon guidance	Biological oxidations	Botulinum neurotoxicity
Cell junction organization	Cell Cycle Checkpoints	Cell Cycle, Mitotic	Chromosome Maintenance
Circadian Clock	DNA Repair	DNA Replication	Diabetes pathways
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	Gene Expression	Hemostasis	HIV Infection
Interactions of the immunoglobulin superfamily (IgSF) member proteins	Influenza Infection	Integration of energy metabolism	Integrin cell surface interactions
Membrane Trafficking	Metabolism of amino acids and derivatives	Metabolism of carbohydrates	Metabolism of lipids and lipoproteins
Metabolism of nitric oxide	Metabolism of nucleotides	Metabolism of porphyrins	Metabolism of proteins
Metabolism of RNA	Metabolism of vitamins and cofactors	Muscle contraction	mRNA Processing
Myogenesis	Pyruvate metabolism and Citric Acid (TCA) cycle	Regulation of beta-cell development	Regulatory RNA pathways
Signaling by BMP	Signaling by EGFR	Signaling by FGFR	Signaling by GPCR
Signaling by PDGF	Signaling in Immune system	Signaling by Insulin receptor	Signalling by NGF
Signaling by Notch	Opioid Signalling	Signaling by Rho GTPases	Signaling by TGF beta
Signaling by VEGF	Signaling by Wnt	Synaptic Transmission	Transcription
Transmembrane transport of small molecules			

About Reactome

REACTOME is a free, online, open-source, curated pathway database encompassing many areas of human biology. Information is authored by expert biological researchers, maintained by the Reactome editorial staff and cross-referenced to a wide range of standard biological databases.

These include NCBI Entrez Gene, Ensembl and UniProt databases, the UCSC and HapMap Genome Browsers, the KEGG Compound and ChEBI small molecule databases, PubMed, and GO. The curated human data are used to infer orthologous events in 20 non-human species including the laboratory mouse and rat, the nematode C. elegans, budding and fission yeasts, two plants, and E.coli. Tools for pathway analysis include Skypainter and Biomart.

Pathway data can be exported in SBML and BioPAX formats.

A description of Reactome has been published in Genome Biology and Nucleic Acids Research.

News and Notes

- **October 4, 2010 Version 34 Released**
  Version 34 includes the new topic Interactions of the immunoglobulin superfamily (IgSF) member proteins covering Nephrin and SIRP interactions. Topics in this release that contain new or revised and updated events include: Transmembrane transport of small molecules (Transport of organic anions, Aquaporin-mediated transport , and Metal ion SLC transporters), Metabolism of proteins (N-glycan trimming in the ER and Calnexin/Calreticulin cycle), Hemostasis (Platelet homeostasis, GPVI signaling, and Thrombin signaling), Signaling in the immune system (Interleukin-3, 5 and GM-CSF signaling and Interleukin-1 processing), Signaling by insulin receptor (Endosome acidification), Signaling by GPCR (Olfactory signaling pathway), Integration of energy metabolism (Incretin synthesis, secretion, and inactivation), and Synaptic transmission (GABA synthesis, release and clearance at the synapse).
  JW Akkerman, G Dall'Olio, K Ray, and R Stephan are our external authors in this release. U Albrecht, N Barclay, E Beitz, SR Bloom, G Calamita, F Delaunay, P Gagneux, F Grahammer, L He, T Hercus, T Huber, S Kay, S Kunapuli, A Lopez, B MacIver, J Mathai, E Pinteaux, S Restituito, H Tsuyoshi , L Vosshall, J Wilusz, and S Zac-Varghese are our external reviewers.

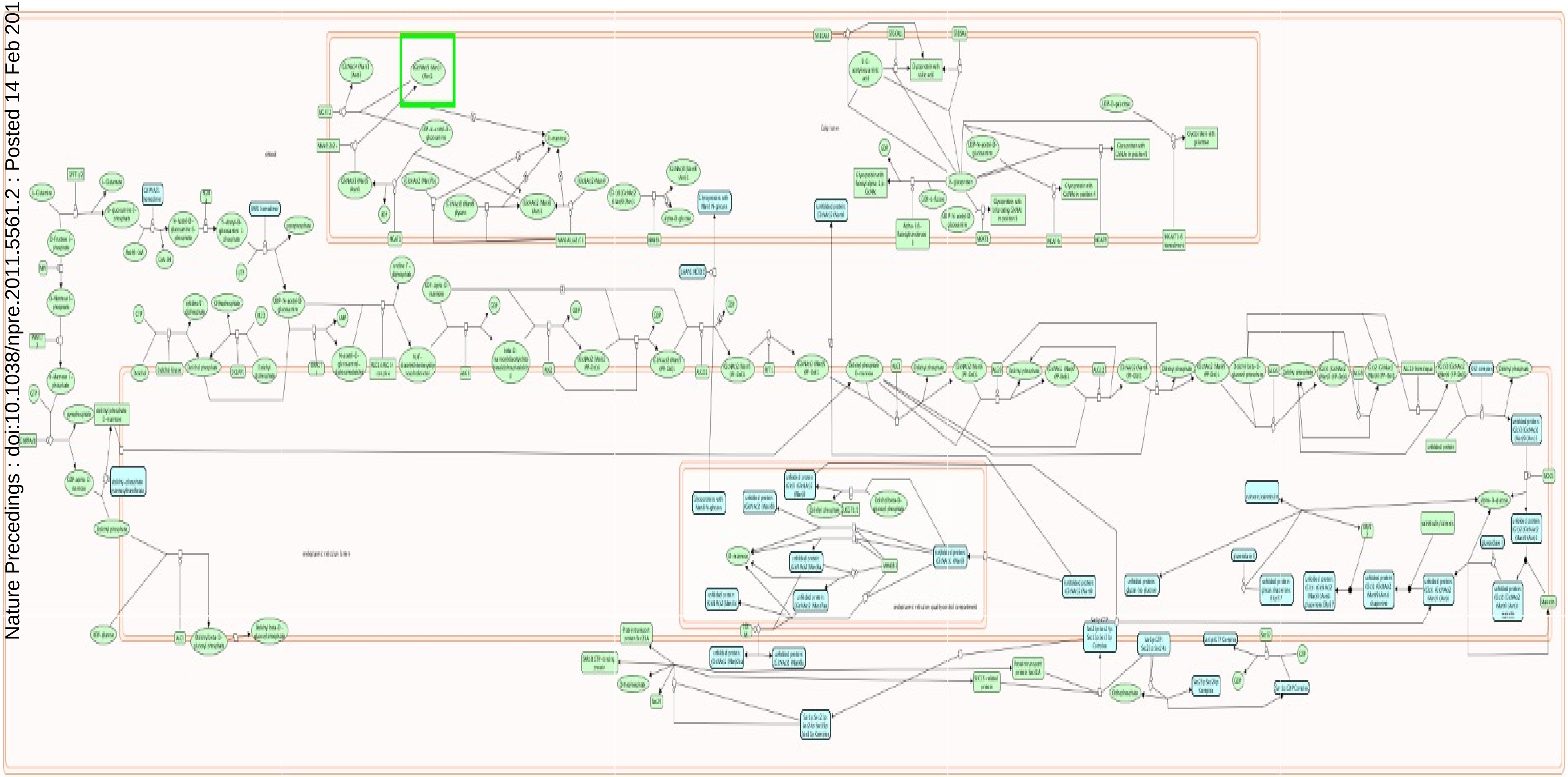
Reactome - advantages

- Open-source approach. Anyone can contribute to a pathway, and all contributions are public and transparent
- Annotates a greater number of informations for each reaction
- Does not artificially distinguish between metabolic/interaction pathways (uses GO terms instead)

The process of annotation in Reactome

- Annotating a pathway in Reactome means that for each reaction, you have to details like:
    - Description and name of the reaction
    - Genes involved
    - GO localization terms for input, outputs and enzymes
    - References to experimental evidences

The process of annotation in Reactome



Our N-Glycosylation pathway

The process of annotation in Reactome

- Annotating a pathway in Reactome means that for each reaction, you have to provide details like:
  - Description and name of the reaction
  - Genes involved
  - GO localization terms for input, outputs and enzymes
  - References to experimental evidences
- Each reaction must pass a peer-reviewed process before being included in Reactome

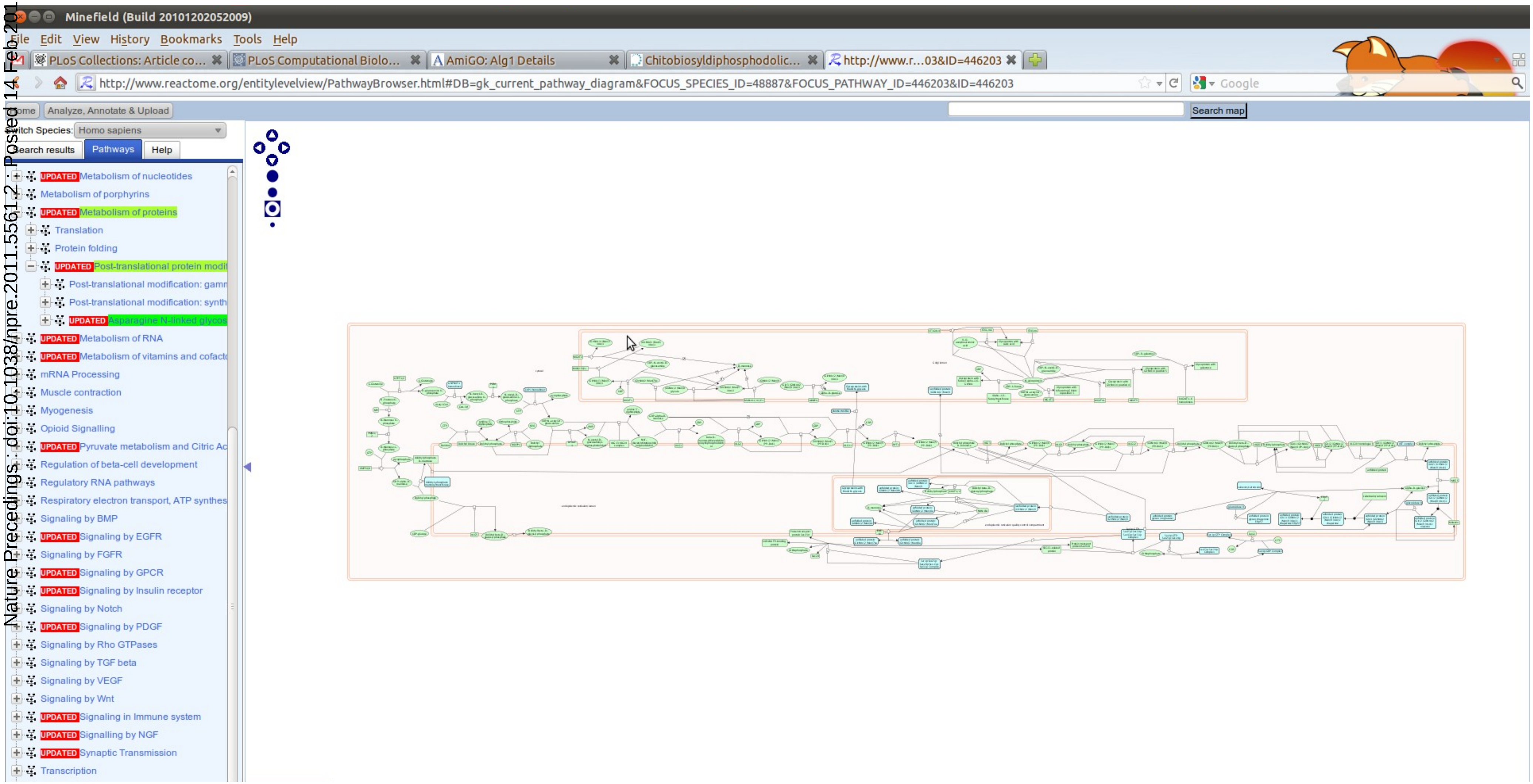
N-Glycosylation on Reactome final result

- It tooks about 6 months of work to annotate the pathway
  - ~100 reactions
  - lot of literature read
  - problems with entries in other databases
  - technical implementation problems
  - problems, problems, problems, time wasted

Is it worth to annotate a pathway manually as I did?

- If you are a PhD student: probably yes, and at the beginning
- Pros:
  - Helps you having good ideas
  - Recognize how much you can trust database annotations
  - Can be published
- Cons:
  - Takes time
  - Not feasible for larger scale studies

The Reactome pathway, complete



http://tinyurl.com/nglyco-reactome

N-Glycosylation in other databases

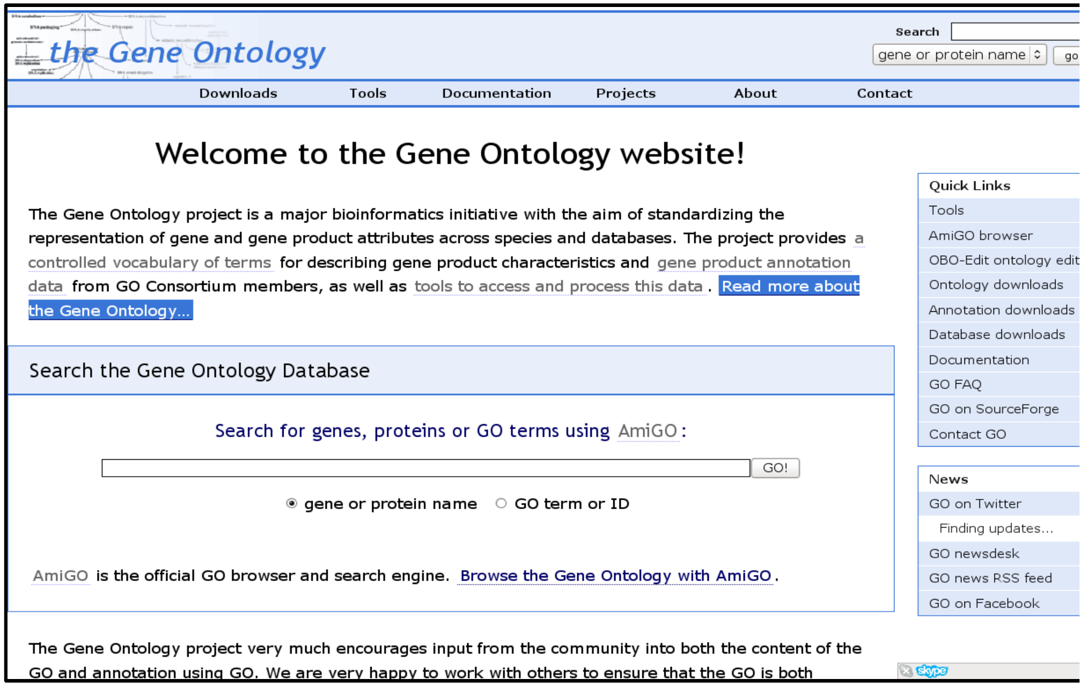
- Historically, this pathway is well characterized and is a textbook example of a metabolic pathway
- It is a good example to compare its annotations in different databases

N-Glycosylation in other databases

- Historically, this pathway is well characterized and is a textbook example of a metabolic pathway
- It is a good example to compare its annotations in different databases
- If you dedicate 6 months to read tons of literature about what you are studying, which kind of errors do you expect to find in databases?

GeneOntology (GO)

- GO is an ontology of terms to describe the function, localization of a gene
- A dictionary designed to reduce the problem of synonyms and unclear terminology



Welcome to the Gene Ontology website!

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. Read more about the Gene Ontology...

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

● gene or protein name ○ GO term or ID

AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.

The Gene Ontology project very much encourages input from the community into both the content of the GO and annotation using GO. We are very happy to work with others to ensure that the GO is both

What is a GO annotation?

- A GO annotation is the association between a GO term and a gene, protein or event
  - *The gene ALG1 is Integral to the membrane (GO:32130)*
  - *The protein XYZ has peroxidase activity (GO:03123)*

GO annotations

- GO annotations are frequently used in the scientific literature
  - *This set of genes is enriched of a specific GO term*
  - *We subdivided the pathway in GO modules and studied selection within them*
- However, which are the procedures to associate a GO term with a gene?
- Which are the possible sources of error in a GO association?

The GO term assignment

- In order to assign a GO term to a protein in a species, there it must be an experimental evidence for the association in that species

The GO term assignment

- In order to assign a GO term to a protein in a species, there it must be an experimental evidence for the association in that species

➡ Lots of false negatives in GO!

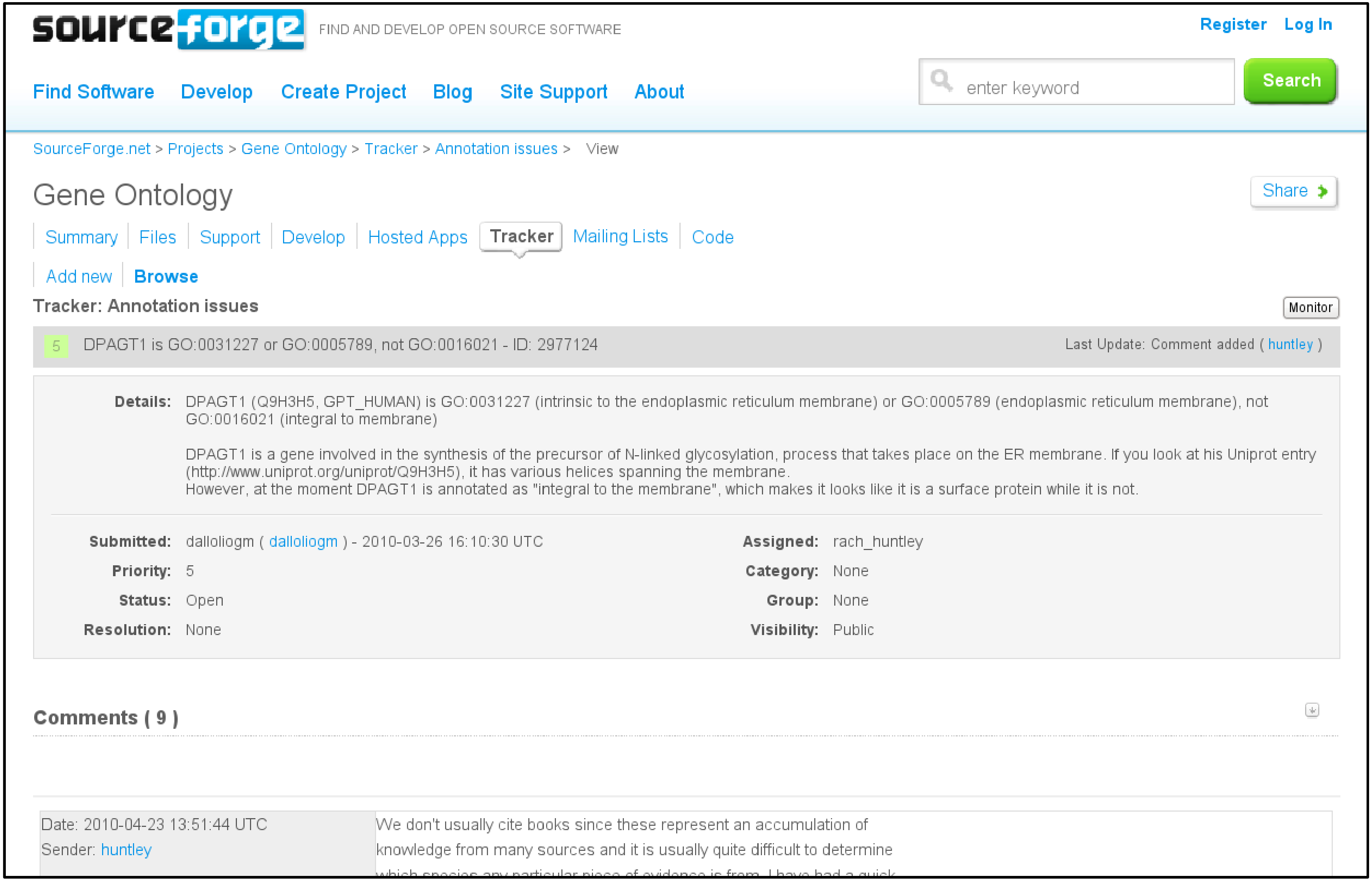
Example of false negative in GO

- We proposed the association of DPAGT1 with the term 'Integral to the ER membrane' in human
  - All the literature assumes that this association exists
  - This association is annotated in Kegg for Yeast

Example of false positive in GO

- We proposed the association of DPAGT1 with the term 'Integral to the ER membrane' in human
  - All the literature assumes that this association exists
  - This association is annotated in Kegg for Yeast
- The proposal has been rejected because there it were no experimental evidence in human
  - Probably there it won't never be any

DPAGT1 on GO issue tracker



Gene Ontology

Tracker: Annotation issues

5 DPAGT1 is GO:0031227 or GO:0005789, not GO:0016021 - ID: 2977124

Last Update: Comment added (huntley)

Details: DPAGT1 (Q9H3H5, GPT_HUMAN) is GO:0031227 (intrinsic to the endoplasmic reticulum membrane) or GO:0005789 (endoplasmic reticulum membrane), not GO:0016021 (integral to membrane)

DPAGT1 is a gene involved in the synthesis of the precursor of N-linked glycosylation, process that takes place on the ER membrane. If you look at his Uniprot entry (http://www.uniprot.org/uniprot/Q9H3H5), it has various helices spanning the membrane.
However, at the moment DPAGT1 is annotated as "integral to the membrane", which makes it looks like it is a surface protein while it is not.

Submitted:	dalloliogm (dalloliogm) - 2010-03-26 16:10:30 UTC	**Assigned:**	rach_huntley
Priority:	5	**Category:**	None
Status:	Open	**Group:**	None
Resolution:	None	**Visibility:**	Public

Comments (9)

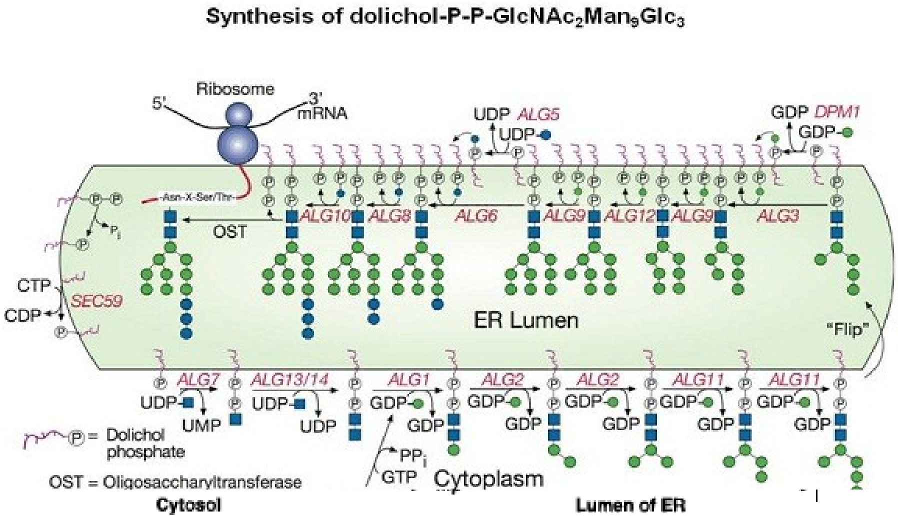
Date: 2010-04-23 13:51:44 UTC
Sender: huntley

We don't usually cite books since these represent an accumulation of knowledge from many sources and it is usually quite difficult to determine

Ambigous interpretation of the term *N-Glycosylation* in GO

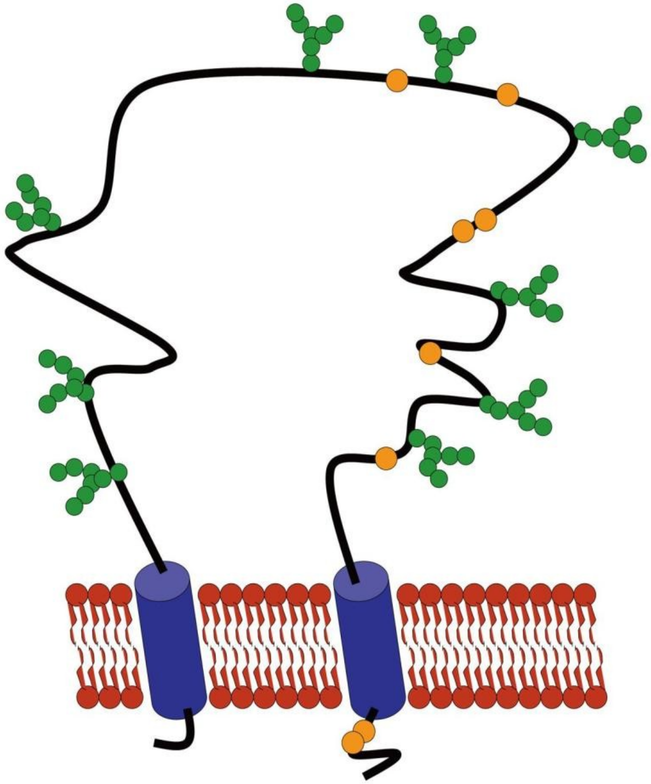
- We found a big mistake with the term 'N-Glycosylation'
- This was being associated to two classes of genes: those which are glycosylated, and those which participate to the process of glycosylation

Ambigous interpretation of the term *N-Glycosylation* in GO



N-Glycosylation pathway

Essentials of Glycobiology, 3rd edition



N-Glycosylated protein

World J Gastroenterol. 2010 December 21; 16(47): 5916-5924.

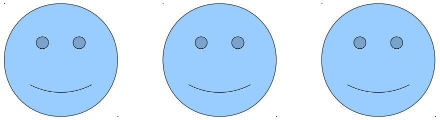
merged

Summary of GO annotations for N-glycosylation

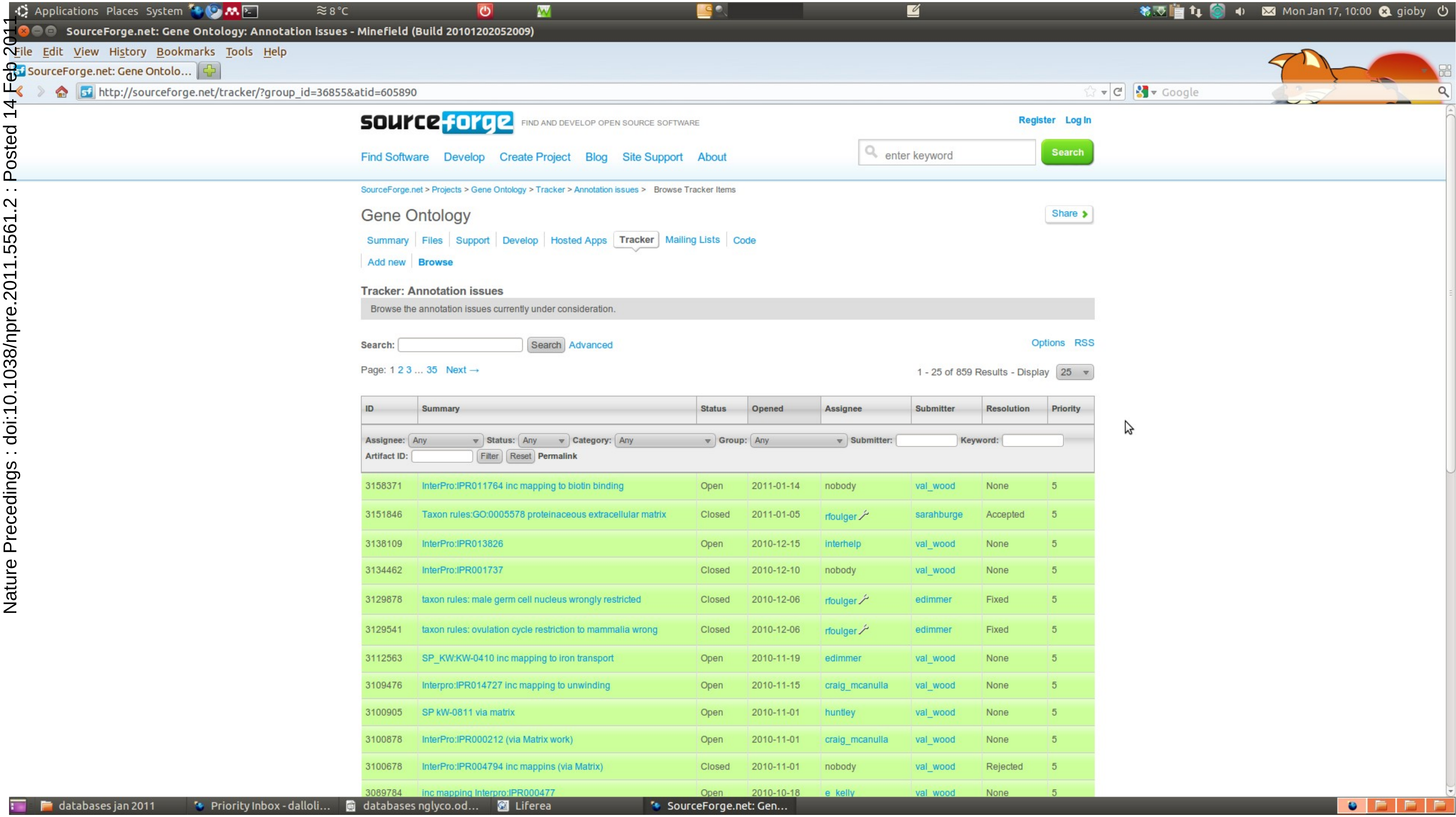
- ~80 correct GO classification
- ~50 new gene-term association proposed
- 10 genes had a clearly wrong classification
- 2 new GO terms proposed:
  - Endoplasmic reticulum quality control compartment
  - Intrinsic to the lumenal side of the endoplasmic reticulum :-)

The GO issue tracker

- All the errors on GO elements are tracked on a publicly accessible online application, the *issue tracker*
- From there, you can report new errors, or check whether there are reported cases for the terms you are using



The GO tracker



SourceForge.net: Gene Ontology: Annotation issues - Minefield (Build 20101202052009)

http://sourceforge.net/tracker/?group_id=36855&atid=605890

SourceForge.net > Projects > Gene Ontology > Tracker > Annotation issues > Browse Tracker Items

Gene Ontology

Tracker: Annotation issues

Browse the annotation issues currently under consideration.

Page: 1 2 3 ... 35 Next →

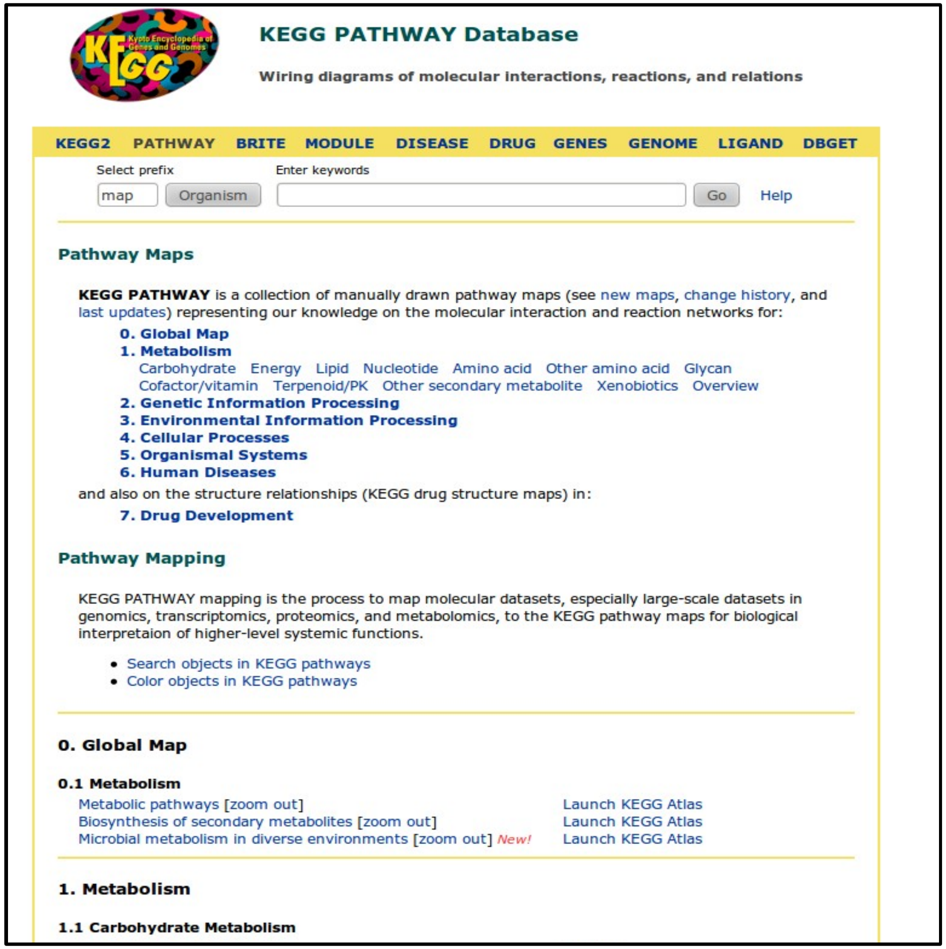
1 - 25 of 859 Results - Display 25

Assignee: Any Status: Any Category: Any Group: Any Submitter: Keyword:
Artifact ID: Filter Reset Permalink

ID	Summary	Status	Opened	Assignee	Submitter	Resolution	Priority
3158371	InterPro:IPR011764 inc mapping to biotin binding	Open	2011-01-14	nobody	val_wood	None	5
3151846	Taxon rules:GO:0005578 proteinaceous extracellular matrix	Closed	2011-01-05	rfoulger	sarahburge	Accepted	5
3138109	InterPro:IPR013826	Open	2010-12-15	interhelp	val_wood	None	5
3134462	InterPro:IPR001737	Closed	2010-12-10	nobody	val_wood	None	5
3129878	taxon rules: male germ cell nucleus wrongly restricted	Closed	2010-12-06	rfoulger	edimmer	Fixed	5
3129541	taxon rules: ovulation cycle restriction to mammalia wrong	Closed	2010-12-06	rfoulger	edimmer	Fixed	5
3112563	SP_KW:KW-0410 inc mapping to iron transport	Open	2010-11-19	edimmer	val_wood	None	5
3109476	Interpro:IPR014727 inc mapping to unwinding	Open	2010-11-15	craig_mcanulla	val_wood	None	5
3100905	SP kW-0811 via matrix	Open	2010-11-01	huntley	val_wood	None	5
3100878	InterPro:IPR000212 (via Matrix work)	Open	2010-11-01	craig_mcanulla	val_wood	None	5
3100678	InterPro:IPR004794 inc mappins (via Matrix)	Closed	2010-11-01	nobody	val_wood	Rejected	5
3089784	inc mapping Interpro:IPR000477	Open	2010-10-18	e_kelly	val_wood	None	5

KEGG/Pathways

- A database of Pathways, curated by experts in the field and manually drawn



KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions, and relations

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps (see new maps, change history, and last updates) representing our knowledge on the molecular interaction and reaction networks for:

- **0. Global Map**
- **1. Metabolism**
  - Carbohydrate Energy Lipid Nucleotide Amino acid Other amino acid Glycan
  - Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Overview
- **2. Genetic Information Processing**
- **3. Environmental Information Processing**
- **4. Cellular Processes**
- **5. Organismal Systems**
- **6. Human Diseases**

and also on the structure relationships (KEGG drug structure maps) in:

- **7. Drug Development**

Pathway Mapping

KEGG PATHWAY mapping is the process to map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretaion of higher-level systemic functions.

- Search objects in KEGG pathways
- Color objects in KEGG pathways

0. Global Map

0.1 Metabolism

Metabolic pathways [zoom out]	Launch KEGG Atlas
Biosynthesis of secondary metabolites [zoom out]	Launch KEGG Atlas
Microbial metabolism in diverse environments [zoom out] *New!*	Launch KEGG Atlas

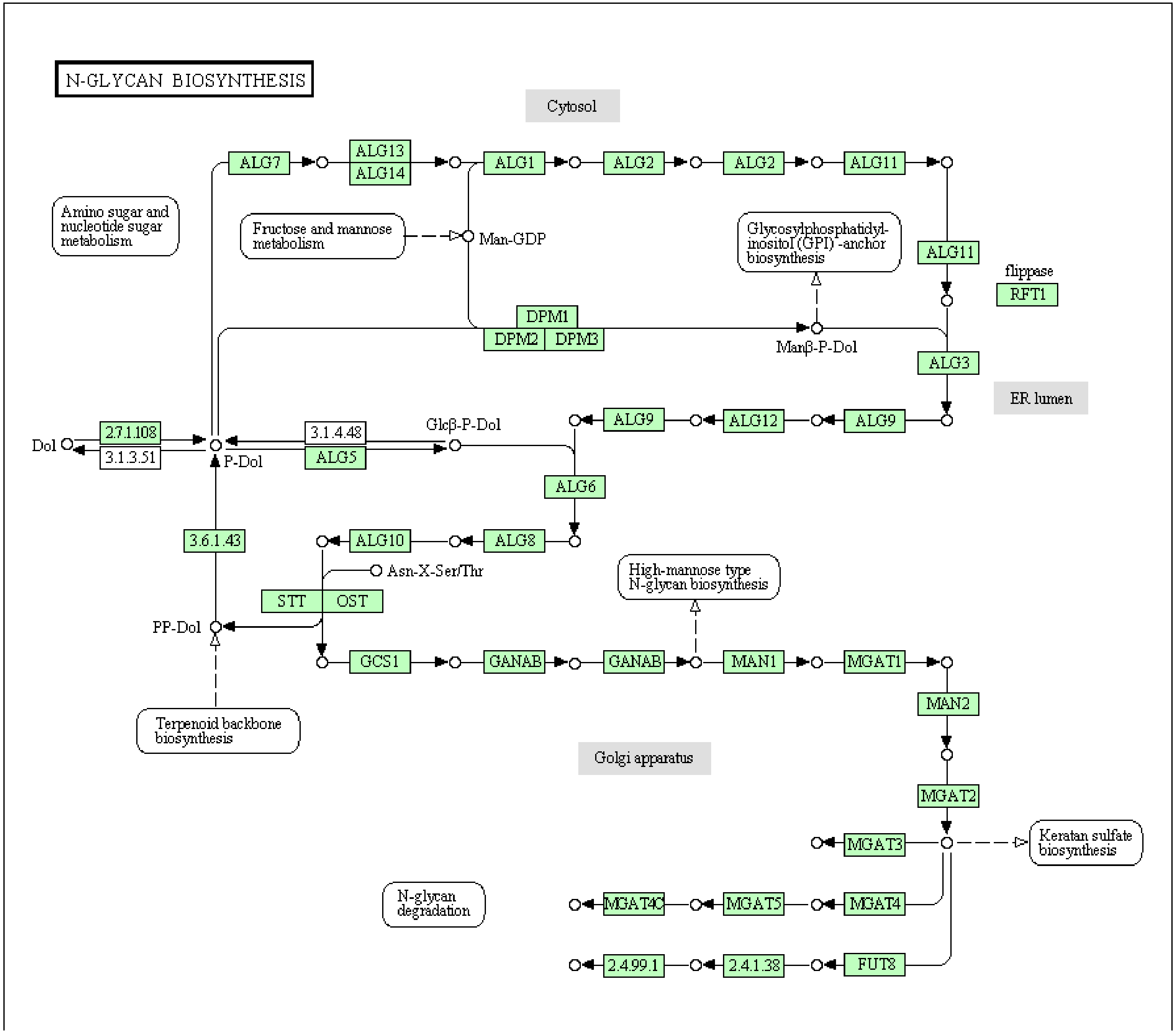
1. Metabolism

1.1 Carbohydrate Metabolism

KEGG/Pathway

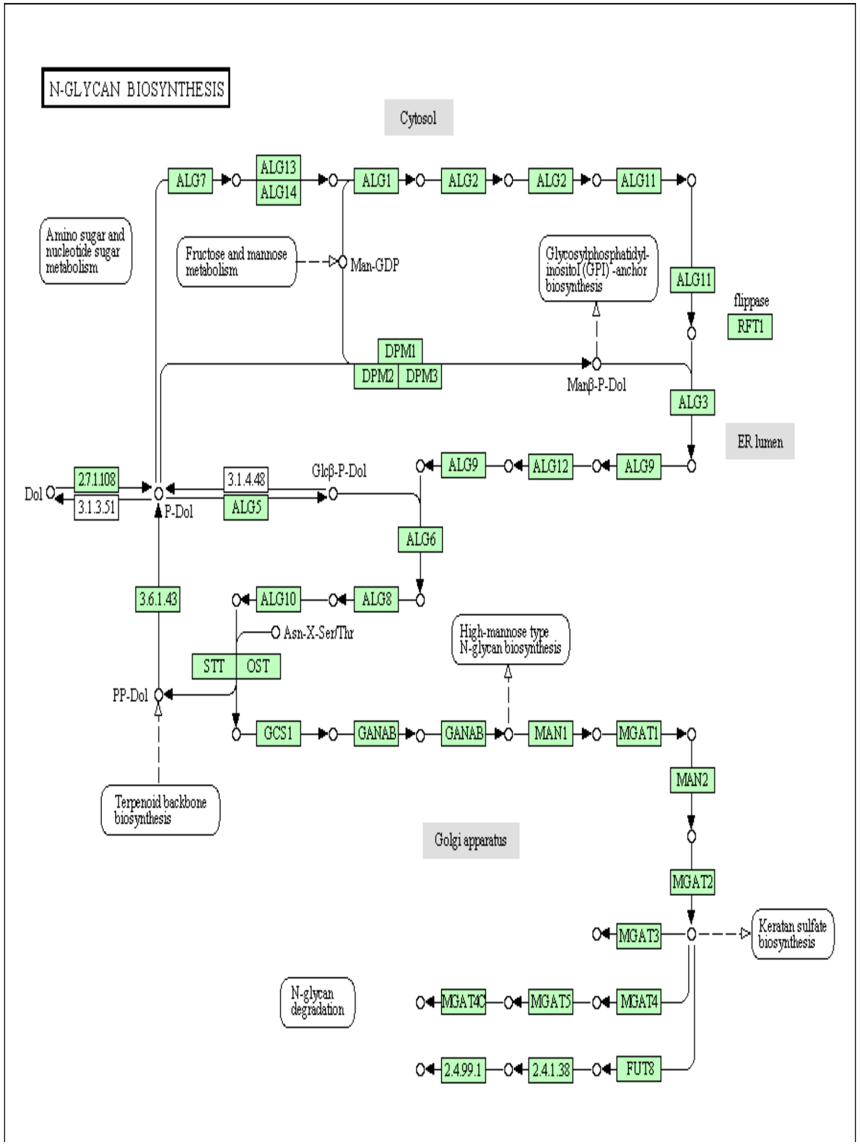
- KEGG/Pathways is the most renowned database for biological pathways
- Pros:
  - Lot of pathways annotated
- Cons:
  - Authors of the pathways are unknown, and no references are given for each reaction. Difficult to ask for clarification

N-Glycosylation on KEGG

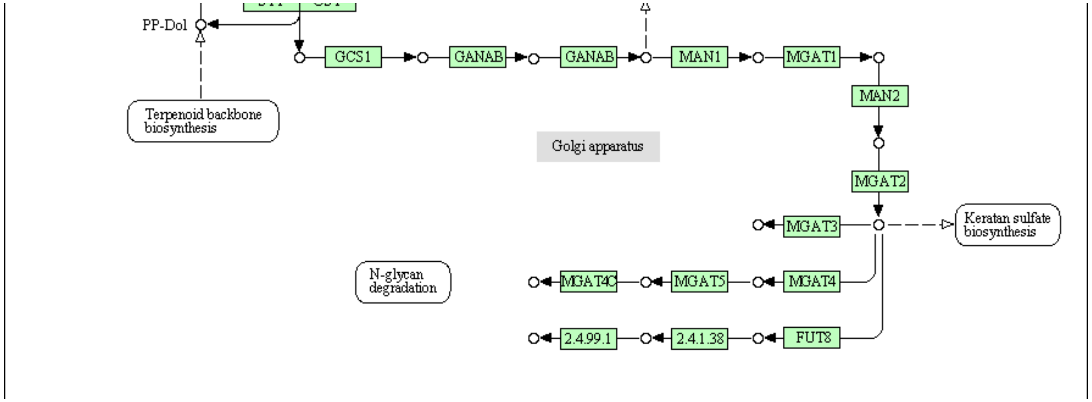


KEGG/Pathways

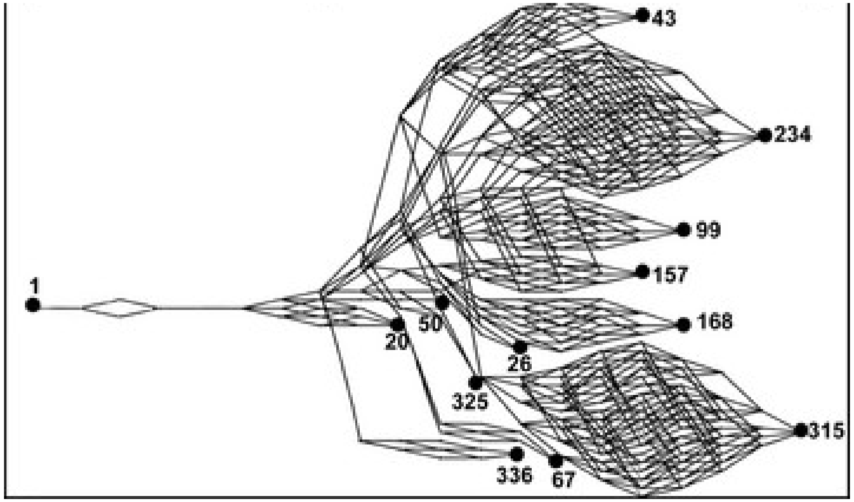
- KEGG/pathways diagrams show a simplified version of the pathway
- These diagrams are not wrong, but they are there for visualization only.
- Easy to interpret them erroneously



This is how the latter part of the pathway really looks like:



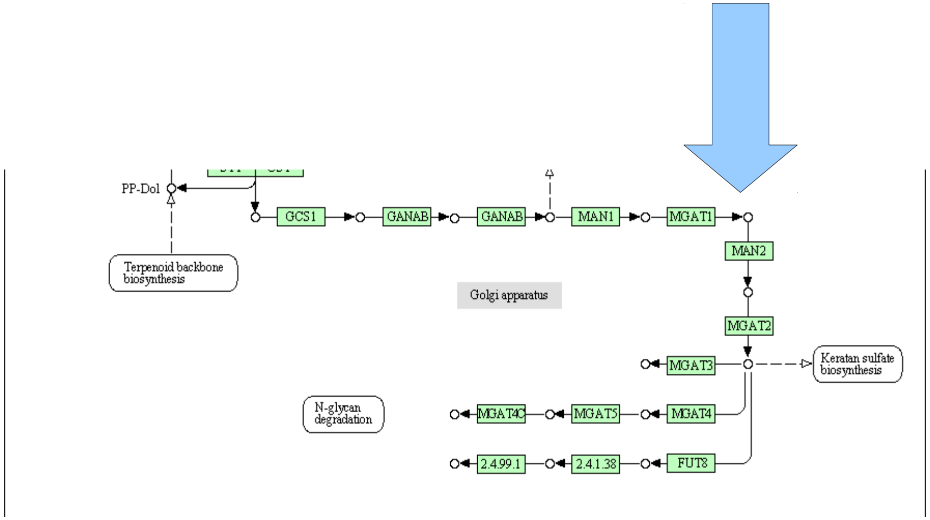
Advanced N-Glycosylation in KEGG



Real represenation of advanced N-Glycosylation

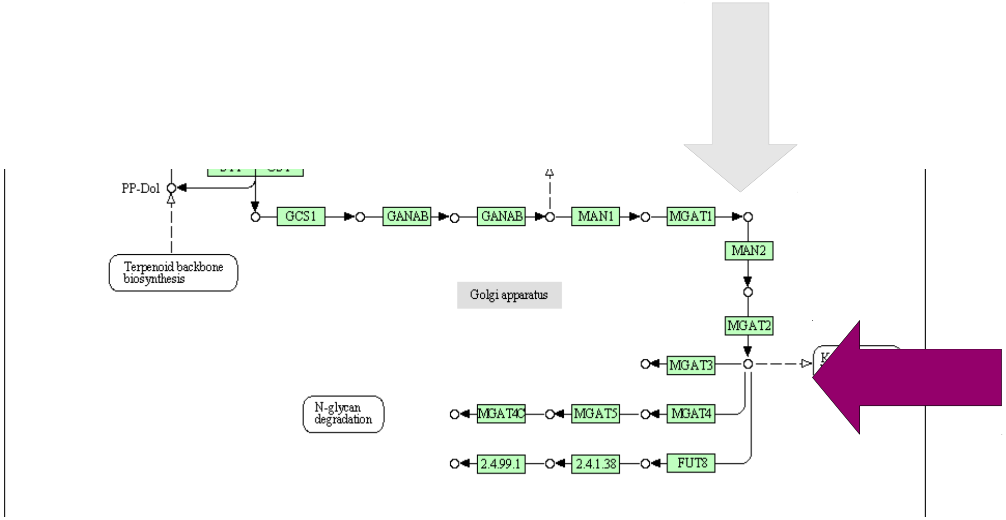
Why it is important to have references for reactions

- For our knowledge of the literature, there it should be a trifurcation at this point



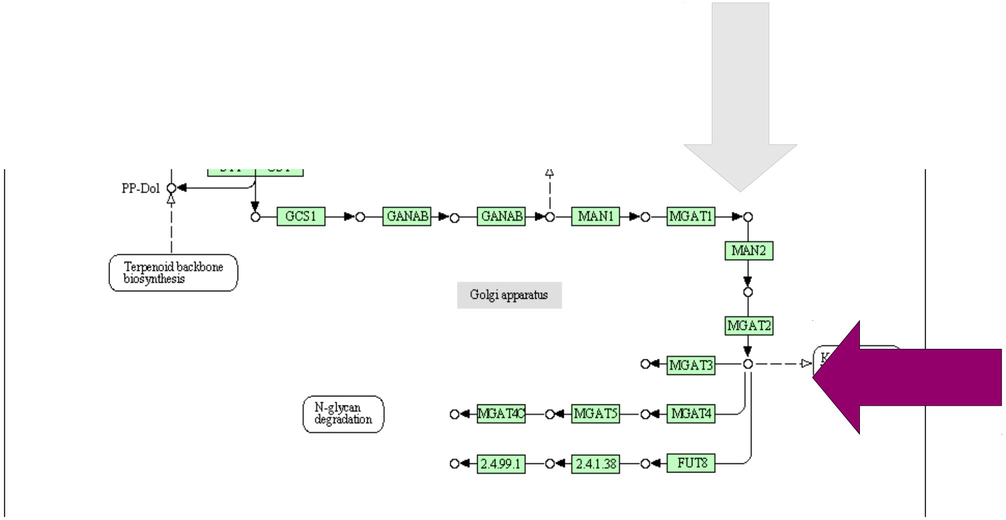
Why it is important to have references for reactions

- For our knowledge of the literature, there it should be a trifurcation at this point
- Instead, it is shown here:



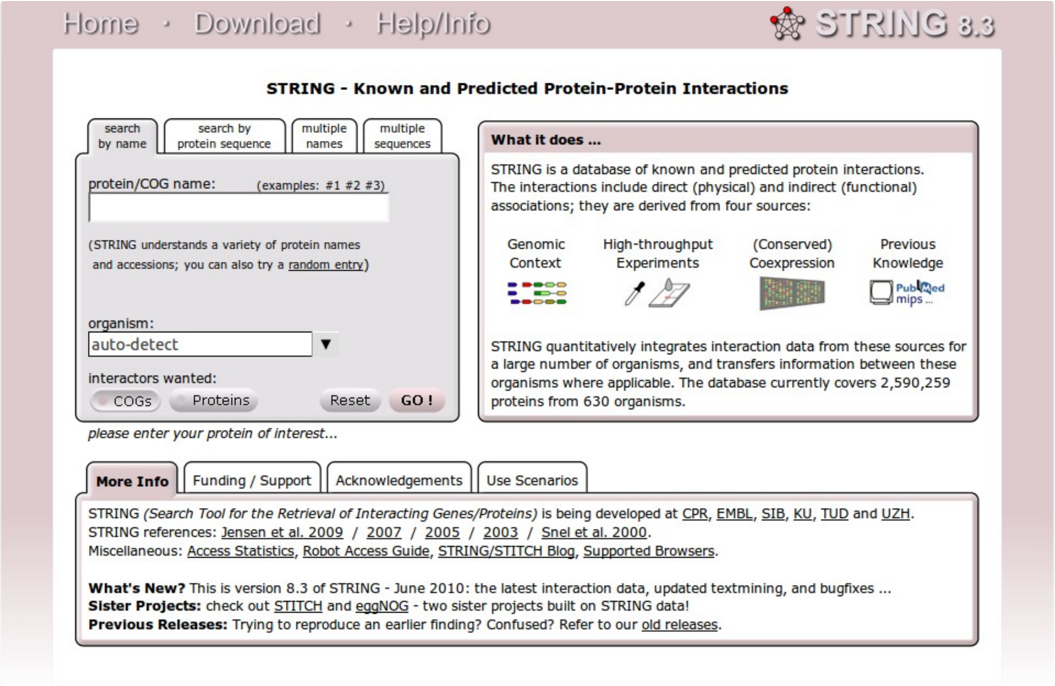
Why it is important to have references for reactions

- For our knowledge of the literature, there it should be a triple bifurcation at this point
- Instead, it is shown here:
- Without references, it is impossible to know why the annotator put the bifurcation there



String

- Database of interactions among proteins
- Collects informations from different resources and merge them together (example of meta-database)



STRING - Known and Predicted Protein-Protein Interactions

(STRING understands a variety of protein names and accessions; you can also try a random entry)

please enter your protein of interest...

STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:

STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 2,590,259 proteins from 630 organisms.

STRING *(Search Tool for the Retrieval of Interacting Genes/Proteins)* is being developed at CPR, EMBL, SIB, KU, TUD and UZH.
STRING references: Jensen et al. 2009 / 2007 / 2005 / 2003 / Snel et al. 2000.
Miscellaneous: Access Statistics, Robot Access Guide, STRING/STITCH Blog, Supported Browsers.

What's New? This is version 8.3 of STRING - June 2010: the latest interaction data, updated textmining, and bugfixes ...
Sister Projects: check out STITCH and eggNOG - two sister projects built on STRING data!
Previous Releases: Trying to reproduce an earlier finding? Confused? Refer to our old releases.

False positives in interactions databases

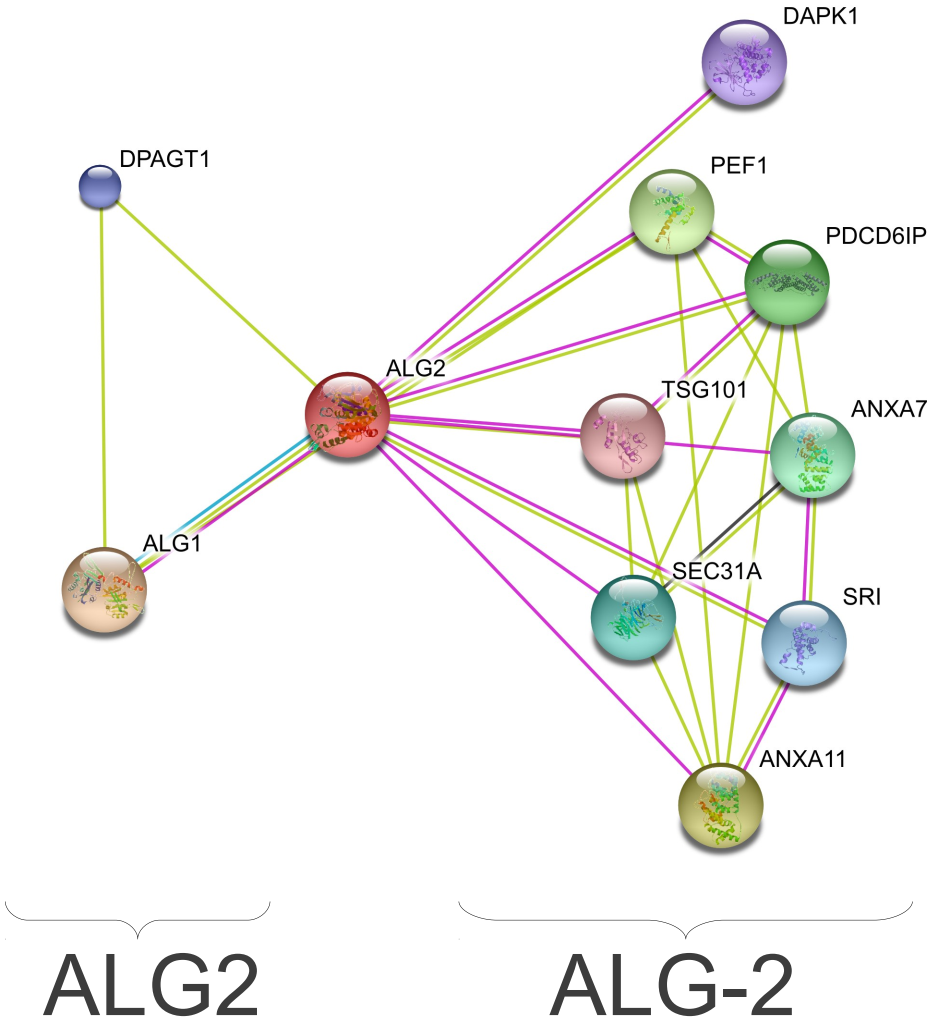
- Interaction databases are usually full of false positives and negatives
  - Yeast-two hybrids and similar
  - Automatic processing of scientific papers

String – the pathway is not there

- Most of the interactions in N-Glycosylation were not found in String
  - Latest release has improved
- The same term *interaction* is ambiguous

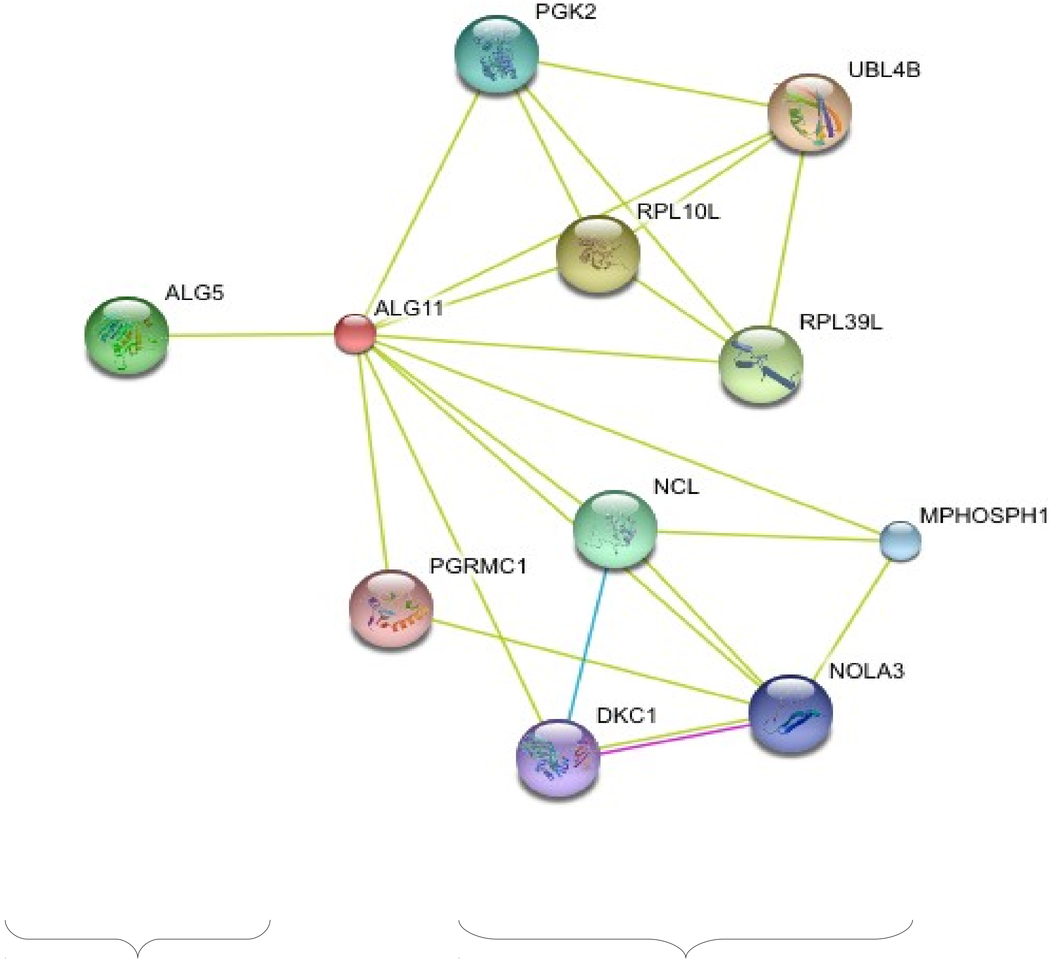
String, the ALG2 case

- The biggest issue we found in String was due to two genes with similar names
- There are two genes with the symbol ALG2:
  - ALG2 (Asparagine Linked Glycosylation 2)
  - ALG-2 (Apoptosis Linked Gene – 2)
- In string, these two were confused



The ALG11 case

- Similar case for ALG11
- Was confused with ALG11, a Ribonucleosomal protein



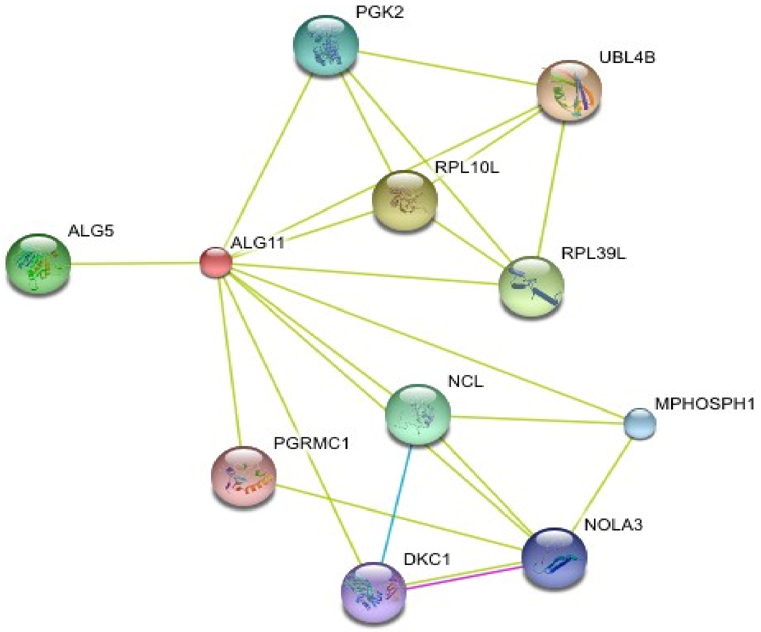
ALG11
(N-Glycosylation)

ALG11
(Ribonucleosomal protein)

What's wrong with String

- The main problem with String is not that there are many false positives/negatives,
- rather, is that there is no way to annotate or report wrong interactions

- It would be good if I could tell other users about these false positives



Uniprot

- Uniprot is a database for annotations of proteins
- The standard resource to use when looking for information on a protein

Annotations in Uniprot

- We have found very few errors in the annotations in Uniprot
- Mostly they were due to outdated names or minor imprecisions
- Every Uniprot entry has a 'Send me feedback' button

So, I have found an error in an annotation – now what?

- Ignore the error

So, I have found an error in an annotation – now what?

- Ignore the error
    - Wrong because the error remains there and can affect other people's work

So, I have found an error in an annotation – now what?

- Ignore the error
    - Wrong because the error remains there and can affect other people's work
- Use a different database

So, I have found an error in an annotation – now what?

- Ignore the error
  - Wrong because the error remains there and can affect other people's work
- Use a different database
  - Even worst than the previous: doesn't fix the error and waste times

So, I have found an error in an annotation – now what?

- Ignore the error
  - Wrong because the error remains there and can affect other people's work
- Use a different database
  - Even worst than the previous: doesn't fix the error and waste times
- Report the error

So, I have found an error in an annotation – now what?

- Ignore the error
  - Wrong because the error remains there and can affect other people's work
- Use a different database
  - Even worst than the previous: doesn't fix the error and waste times
- Report the error
  - This is the best approach, however not everybody knows how to do it, and it requires time

Current state of reporting errors in the scientific community

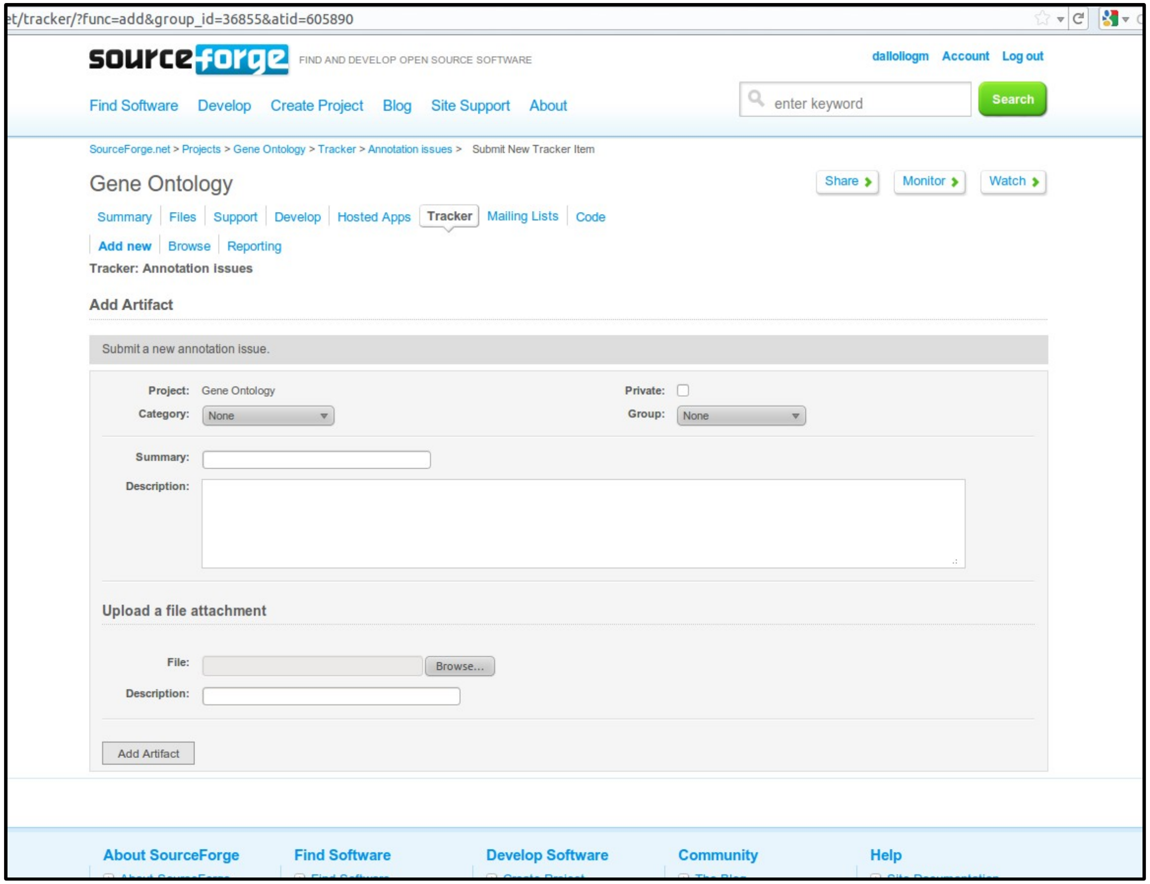
- Unfortunately, reporting errors is not encouraged in the modern scientific literature
  - Time consuming, but not acknowledged
  - Few people know how to do it
  - Made difficult by lack of transparency

Issue trackers

- Usually the best way to report an error is by using an issue tracker software
- When you find an error in a database, always check whether there is a public issue tracker
- If not, things get more difficult

How to report an error in GO

- Check the instructions
- Verify the error has not been reported yet
- Go to the tracker Home page, and click on 'Report new issue'
- Describe the problem with a good title and references



et/tracker/?func=add&group_id=36855&atid=605890

SourceForge.net > Projects > Gene Ontology > Tracker > Annotation issues > Submit New Tracker Item

Gene Ontology

Tracker: Annotation Issues

Add Artifact

Submit a new annotation issue.

Project: Gene Ontology

Category: None

Private: ☐

Group: None

Summary:

Description:

Upload a file attachment

File:

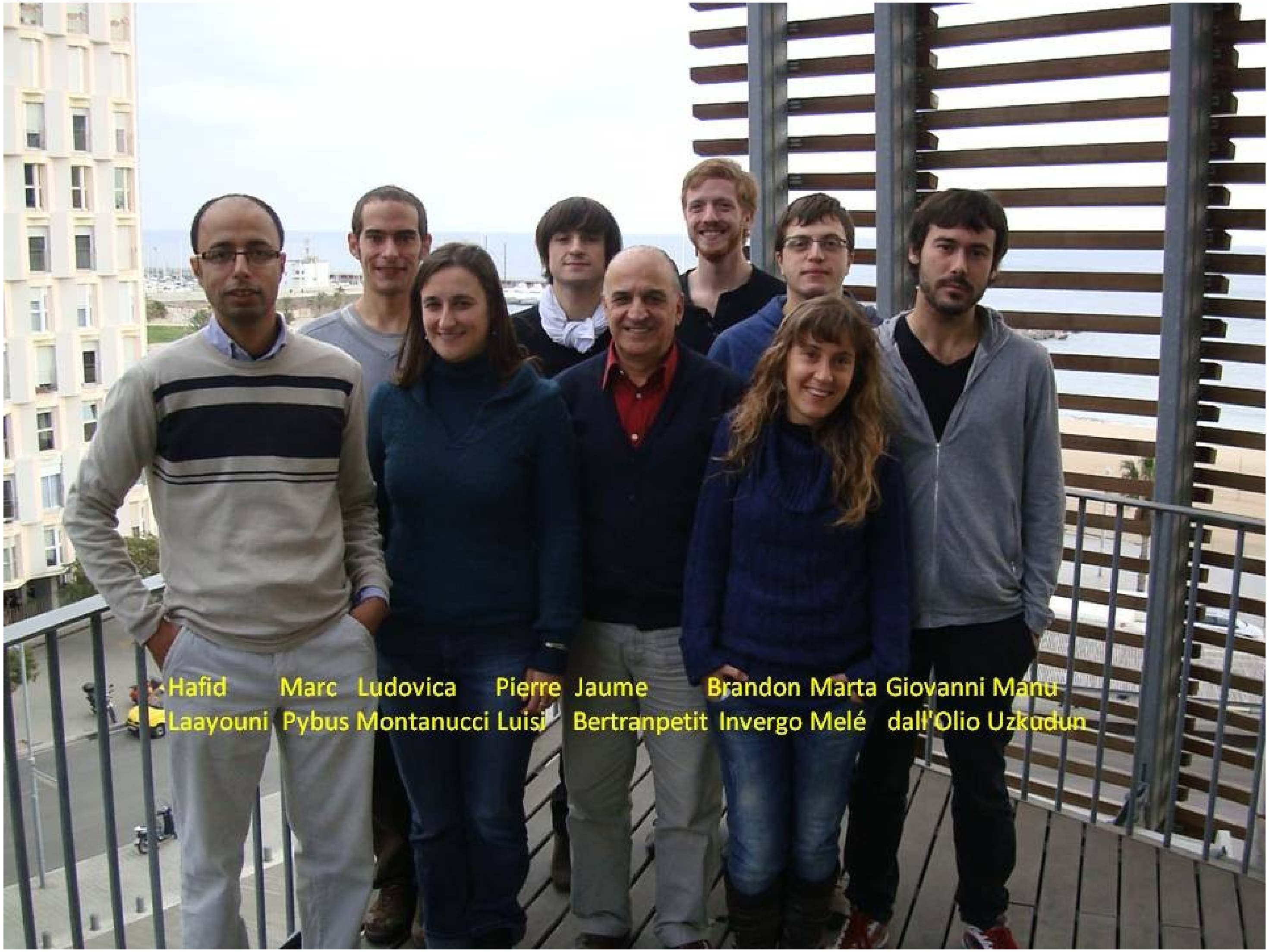
Description:

Take-home messages

- Before using any data from a scientific database:
  - Check whether they have a public issue tracker and check whether there are errors reported about your data
  - Read the literature!

Agradecimientos

(Thank you, grazie, moltes gracies)



GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

iBe CSIC UPF
INSTITUT de BIOLOGIA EVOLUTIVA

Thanks also to:

- Martin Sikora
- Kevin Keys
- Anna Bauer-Mehren from IMIM
- Everybody in BioEvo!

http://www.ibe.upf-csic.es/ibe/research/research-groups/bertranpetit.html