

“

Big data are not a substitute for, but a supplement to traditional data collection and analysis.

”

The Parable of Google Flu
Lazer et al. 2014. *Science* 343 (6176): 1203-1205

Community-Driven Genome Curation: *Harnessing the Power of the Crowd*

Monica Munoz-Torres, PhD | @monimunozto

Phoenix Bioinformatics

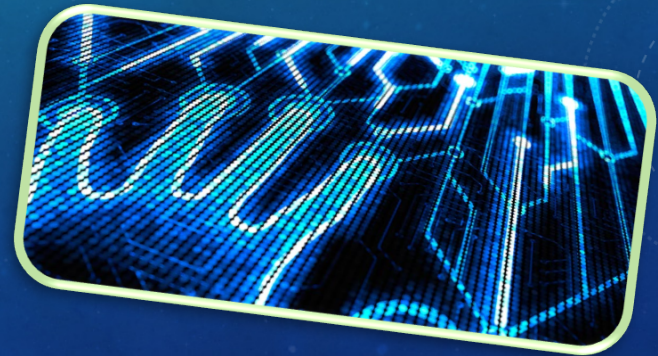
Formerly at Berkeley Bioinformatics Open-Source Projects, Berkeley Lab

Bioinformatics Workshop, CSHL | 29 November, 2017



Today...

Collaborative approaches to improving
genomic resources using software
tools that facilitate the curation process.

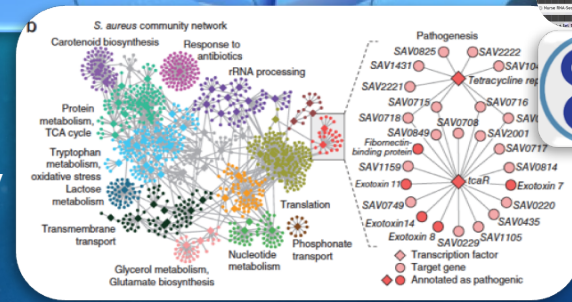


Moni Munoz-Torres, PhD

biocurator, genomicist, insect-biased

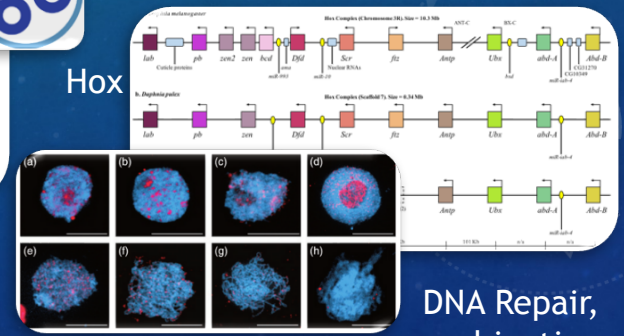


- Research, Teaching, Training
- Diversity & Inclusion in STEM
- Project Management & User advocacy

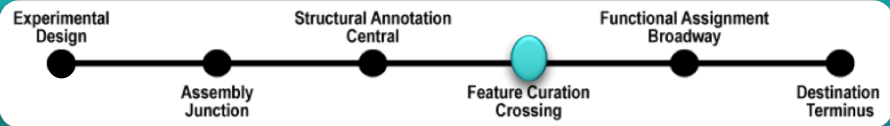


<http://GeneOntology.org>

<http://GenomeArchitect.org>



GENOME TRAIN



Sequencing

TTGAAAATTCTTCAAAAAGAGGGGAG
GTGATTACATACAATCGAGGTGCCTA
TTTGTCACTACATTTGCACCTATGTTT
GTAAGTTGATGAGAGAGAAAATGTGTG
TTTGCTAAACAAGTTTTATAAATAGTTG
AAATAATAGAAAACAATAAATGAATA*

Assembly

Automated Annotation

Input: genomic DNA sequence
GCCTGGAAAACCTCAACTTT...

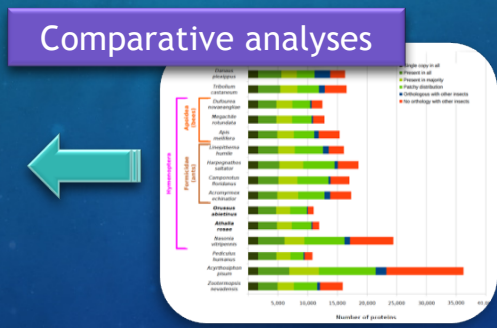
Gene model

Output: gene annotation

Experimental design, sampling

Manual Annotation

Synthesis & dissemination



Merged Gene Set

Extracting Knowledge from Data



“ NIH [seeks] fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability. ”

U.S. National Institutes of Health



“ Good health makes life better. We want to improve health for everyone by helping great ideas to thrive. Science and research expand knowledge by testing and investigating ideas. ”



Wellcome Trust



“

[The NSF aims to] promote and disseminate research, creating knowledge that is valuable to society, the economy, and politics.

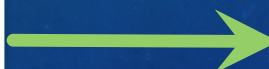
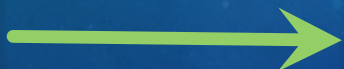
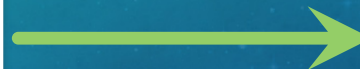
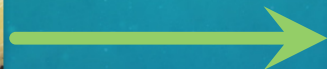
”

Swiss National Science Foundation

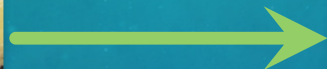
Good data are required

- Data must be infinitely shareable.
- Data reuse increases their value.
- The more data are reused, the more data they generate.
- Data are perishable.
- Combining data sets increases the value of individual data sets.
- The more accurate the information is, the more useful (and valuable!) it is.
- More data are not necessarily better.

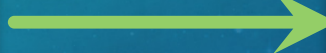
Scientific data



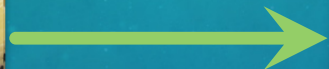
Curation



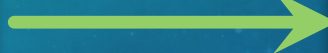
?



Curation

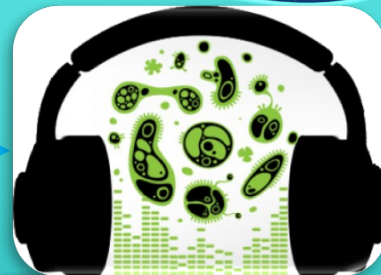
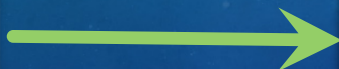


?

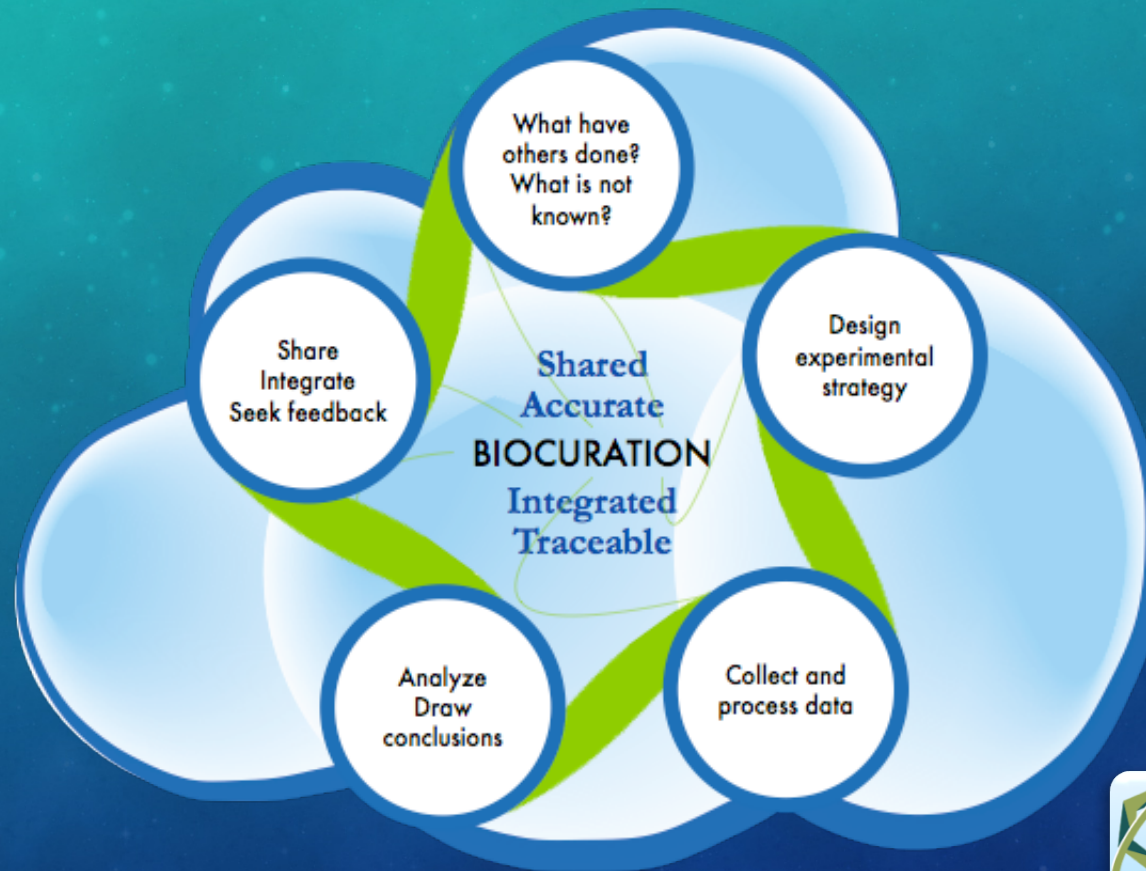


Biocurators

knowledge



Biocuration is a knowledge-extraction process



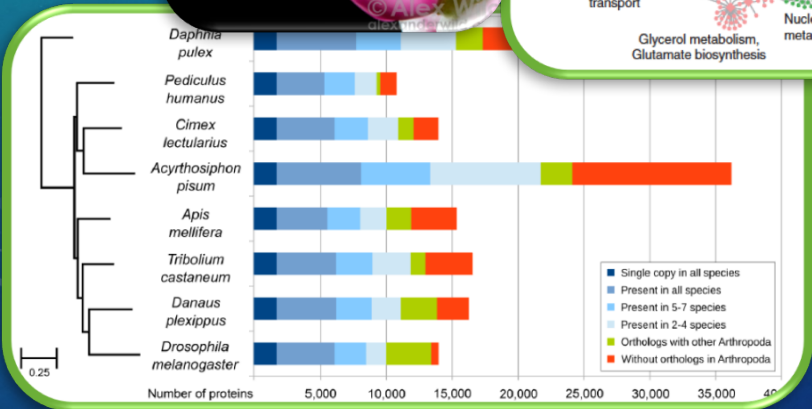
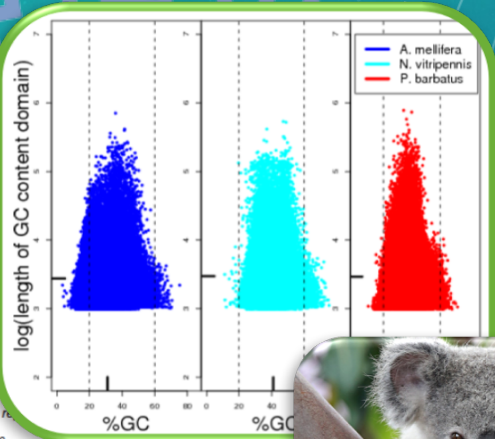
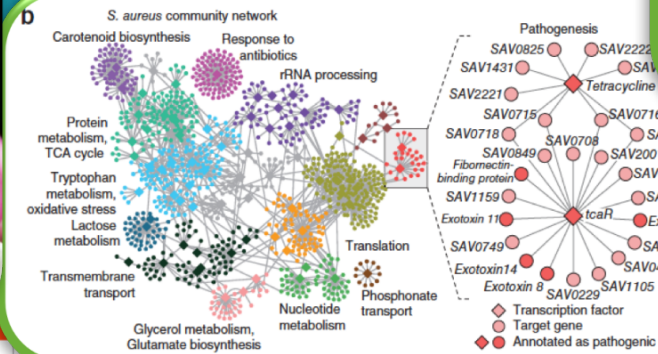
LIFE SCIENCES KNOWLEDGE
Figure is not yet public. Do not reproduce.



Genome Curation



Unlocking genomes



Good genes are required!

1. Generate gene models

- A few rounds of gene prediction.

2. Annotate gene models

- Function, expression patterns, metabolic network memberships.

3. Manually review them

- Structure & Function.

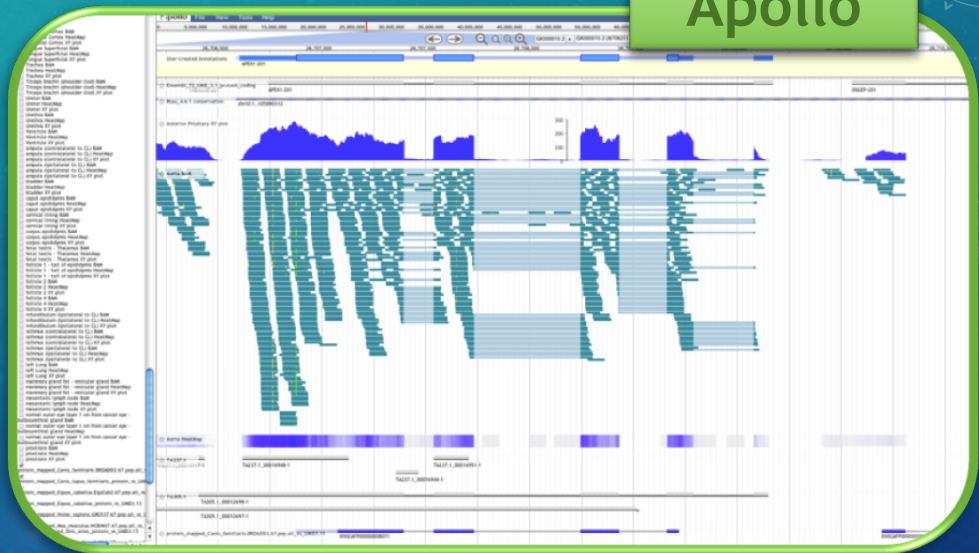


Curation improves quality

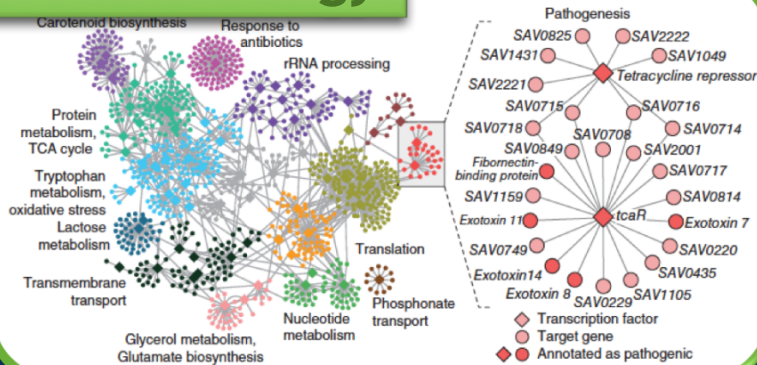


Apollo

Best representation of biology & removal of elements reflecting errors in automated analyses.



Gene Ontology



Functional assignments through comparative analysis using literature, databases, and experimental data.



Curation is valuable

- To make accurate orthology assessments
- To accurately annotate expanded / contracted gene families
- To identify novel genes, species-specific isoforms
- To efficiently take advantage of transcriptomic analyses



Curation is inherently collaborative



- It is impossible for a single individual to curate an entire genome with precise biological fidelity.



- Curators need second opinions and insights from colleagues with domain and gene family expertise.



Predicting & annotating gene structures



Gene Prediction & Gene Annotation

Identification and annotation of genomic elements:

- Primarily focuses on protein-coding genes.
- Also identifies RNAs (tRNA, rRNA, long and small non-coding RNAs (ncRNA)), regulatory motifs, repetitive elements, etc.
- Happens in 2 steps:
 - Computation phase
 - Annotation phase



Collaboratively curating gene structures

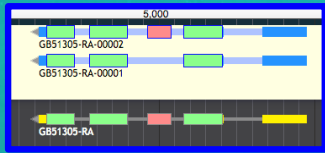


Apollo Genome Annotation Editor

Collaborative, instantaneous, web-based, built on top of JBrowse.

GenomeArchitect.org

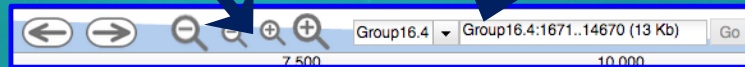
★ Color by CDS frame, toggle strands, set color scheme and highlights.



★ Query the genome using BLAT.

★ Navigation and zoom.

★ Search for a gene model or a scaffold.



★ Upload evidence files (GFF3, BAM, BigWig), add combination and sequence search tracks.

★ User-created annotations.

★ Stage and cell-type specific transcription data.

★ Admin

★ Annotator panel.

★ Evidence Tracks.

PROTEIN CODING, PSEUDOGENES, NCRNAS, REGULATORY ELEMENTS, VARIANTS, ETC.



Apollo

Right-click functionality

A screenshot of the 'Change annotation type' context menu in the Apollo genome browser. The menu is open, showing a list of annotation types. The 'pseudogene' option is highlighted. Other options include 'Delete', 'Merge', 'Split', 'Duplicate', 'Make Intron', 'Move to Opposite Strand', 'Set Translation Start', 'Set Translation End', 'Set Longest ORF', and 'Set Readthrough Stop Codon'. The background shows a portion of the genome browser interface with a track labeled 'gencito-3'.

- Change annotation type
 - gene
 - pseudogene**
 - rRNA
 - snRNA
 - snoRNA
 - tRNA
 - ncRNA
 - miRNA
 - repeat_region
 - transposable_element
- Delete
- Merge
- Split
- Duplicate
- Make Intron
- Move to Opposite Strand
- Set Translation Start
- Set Translation End
- Set Longest ORF
- Set Readthrough Stop Codon

A screenshot of the Apollo genome browser interface. The main window displays a genomic track with various annotations. A right-click context menu is open over a track labeled 'gencito-3'. The menu options include: 'Get Sequence', 'Get GFF3', 'Zoom to Base Level', 'Edit Information (alt-click)', 'Change annotation type', 'Delete', 'Merge', 'Split', 'Duplicate', 'Make Intron', 'Move to Opposite Strand', 'Set Translation Start', 'Set Translation End', 'Set Longest ORF', 'Set Readthrough Stop Codon', 'Set as 5' end', 'Set as 3' End', 'Set both Ends', 'Set to Downstream Splice Donor', 'Set to Upstream Splice Donor', 'Set to Downstream Splice Acceptor', 'Set to Upstream Splice Acceptor', 'Undo', 'Redo', and 'Show History'. The background shows a genomic track with various annotations, including 'gencito-3', 'GB51305-RA', 'gnl|AmeL_4.5|TA30.1_00044386-1', 'GB51304-RA', 'GB51307-RA', '|TA30.1_00005439-1', 'meL_4.5|TA30.1_00005439-2', and 'gnl|AmeL_4.5|TA30.1_000'. The interface includes a top navigation bar with 'Tools' and 'Help' menus, and a bottom status bar with the user name 'McMunozT@lbl.gov'.

Apollo



GenomeArchitect.org



Apollo

Export



Annotations Tracks Ref Sequence Organism Users Groups Admin

Search

Length Minimum

Export All Selected (2) None

GFF3 FASTA

Export

Honey2
2 exported
Type: GFF3
 GFF3 GFF3 with FASTA Export Annotations Close

Export

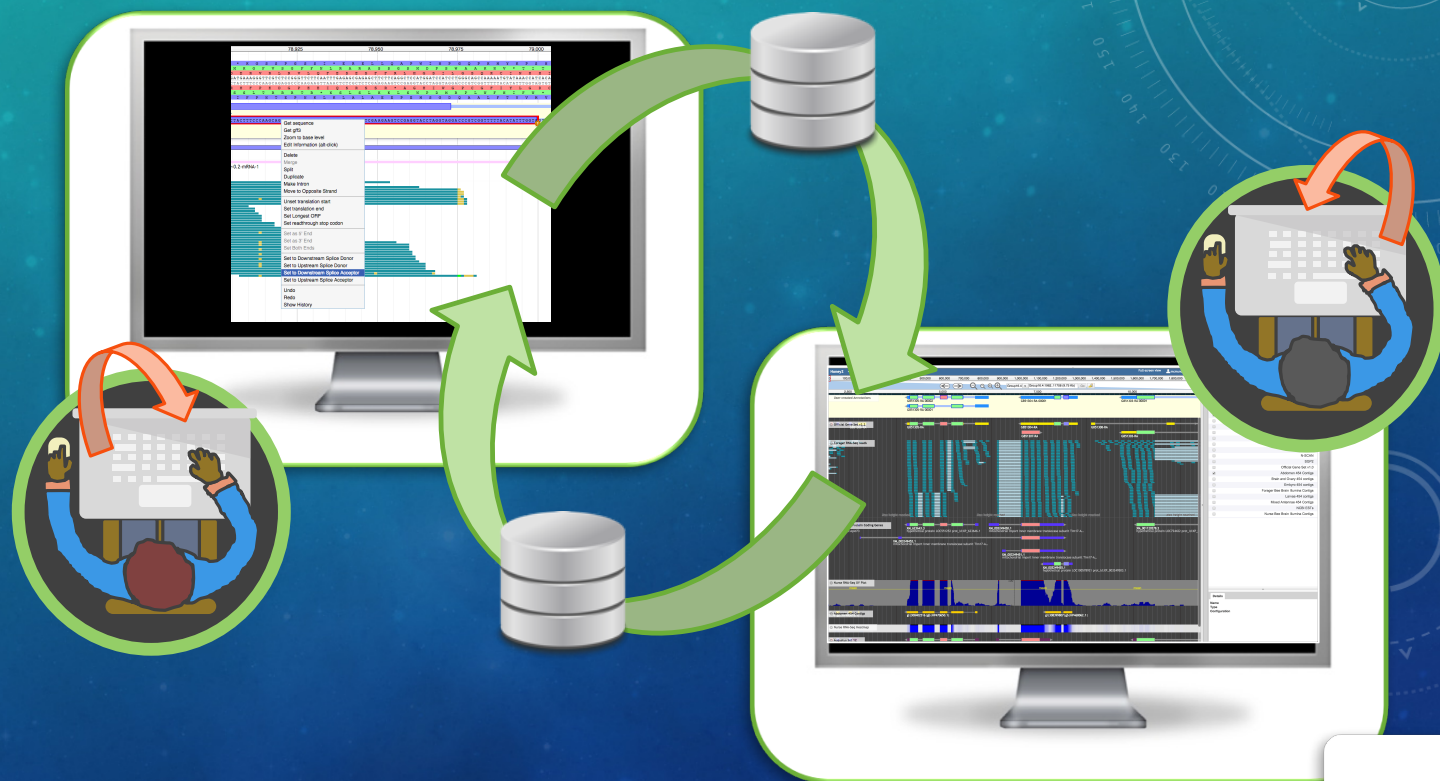
Honey2
2 exported
Type: FASTA
 Genomic cDNA CDS Peptide Export Annotations Close

	Length
	540
	560
GroupUn5044	564

Apollo

Collaboration in Real Time

Apollo





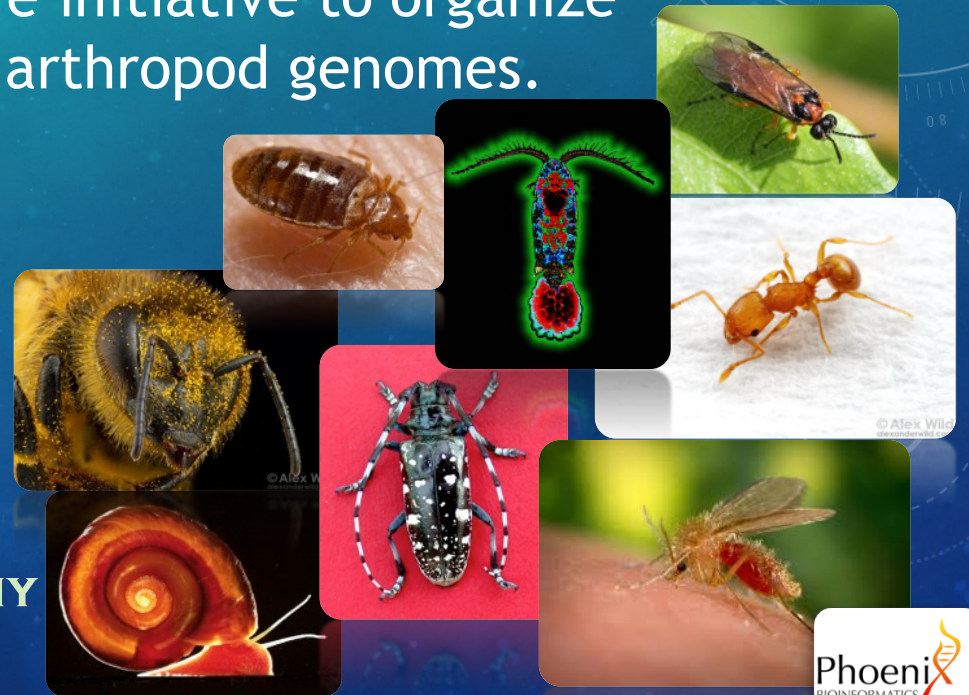
i5k - five thousand arthropod genomes

<http://i5k.github.io>



- Transformative, broad, & inclusive initiative to organize sequencing and analysis of 5,000 arthropod genomes.

- **WORLDWIDE AGRICULTURE**
- **FOOD SAFETY**
- **MEDICINE**
- **ENERGY PRODUCTION**
- **MODELS IN BIOLOGY**
- **MOST ECOSYSTEMS**
- **EVERY BRANCH OF THE PHYLOGENY**



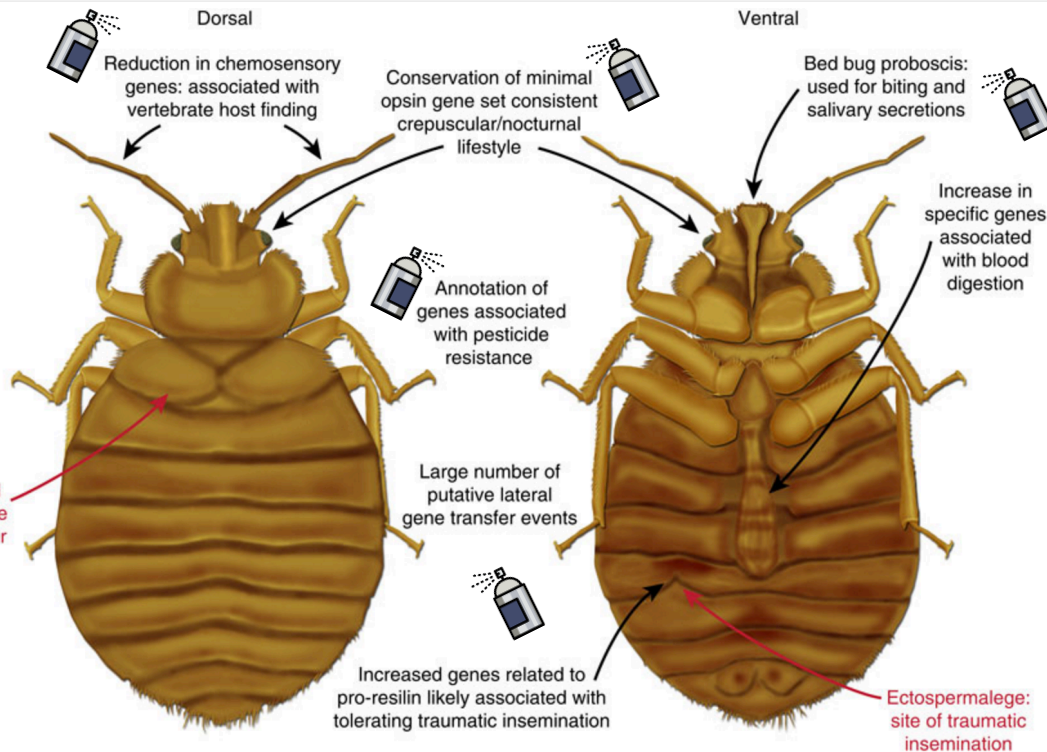
The bed bugs, they're back!

Benoit et al. (2015) *Nature Communications*. doi:10.1038/ncomms10165

- International Travel and Commerce
- Increased Insecticide Resistance

~80 Curators!

i5k.github.io



Red, general characteristics of bed bugs; black, key aspects identified and expanded by genome sequencing and manual curation.

- Timely resource for biology of human ectoparasites.
- Discovery of new targets for control.
- Common lab strain collected before introduction of pyrethroid insecticides.

- What triggered the current bed bug resurgence?
- Did bed bugs originate from one or multiple sources?
- Studies on mechanisms that hinder vertebrate pathogen survival & proliferation and transmission.

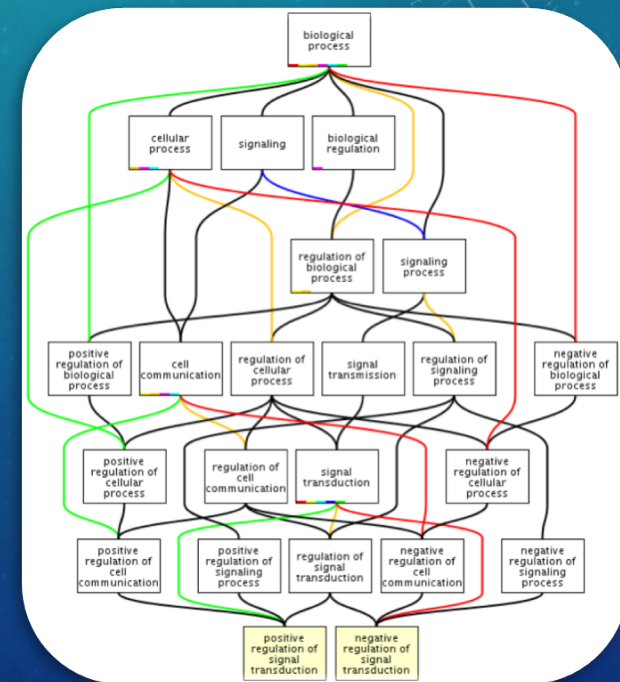


Annotating gene functions



Background in Ontologies

- Terms (classes) arranged in a graph
 - Entities such as genes annotated to terms
- Examples
 - GO
 - PO
 - TO
 - CHEBI



Language inconsistencies in biology

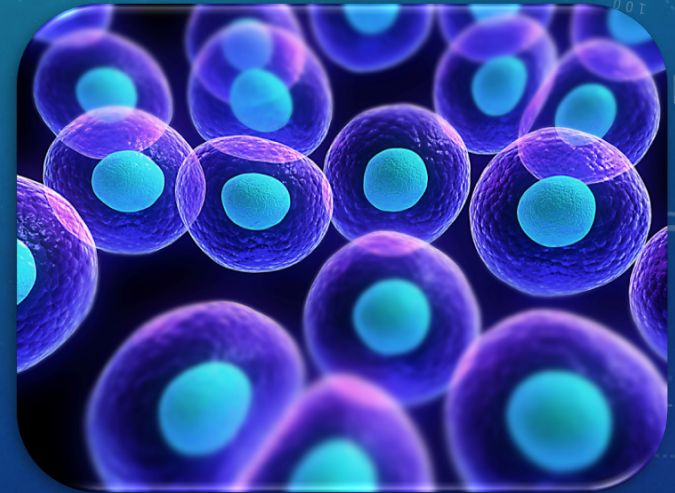
The same name for two different concepts

Cell

or



www.robertpaterson.com



www.biology.usf.edu

Language inconsistencies in biology

Two or more different names for the same concept

Eggplant

Aubergine



Brinjal

Melongene

The same is true for biological concepts:

- This makes comparisons difficult, specially when comparing across species or databases.

Growing number of available biological data

The screenshot shows the PubMed website interface. At the top, there are navigation links for 'NCBI Resources' and 'How To'. The search bar contains 'dna repair' and the 'PubMed' database is selected. Below the search bar, there are options to 'Create RSS', 'Create alert', and 'Advanced'. The main content area shows 'Search results' for 'Items: 1 to 20 of 81894'. A single result is visible, titled 'CITED2 silencing sensitizes cancer cells to cisplatin by inhibiting p53 trans-activation and chromatin relaxation on the ERCC1 DNA repair gene.' by Liu YC, Chang PY, Chao CC. The result is from 'Molecular Cell Biology' (Mol Cell Biol) 2015 Sep 17; pii: gkv934. [Epub ahead of print].

The 'Results by year' bar chart shows a steady increase in the number of publications from 2000 to 2015, with a significant spike in 2015.

Article types: Summary (selected), 20 per page, Sort by Most Recent

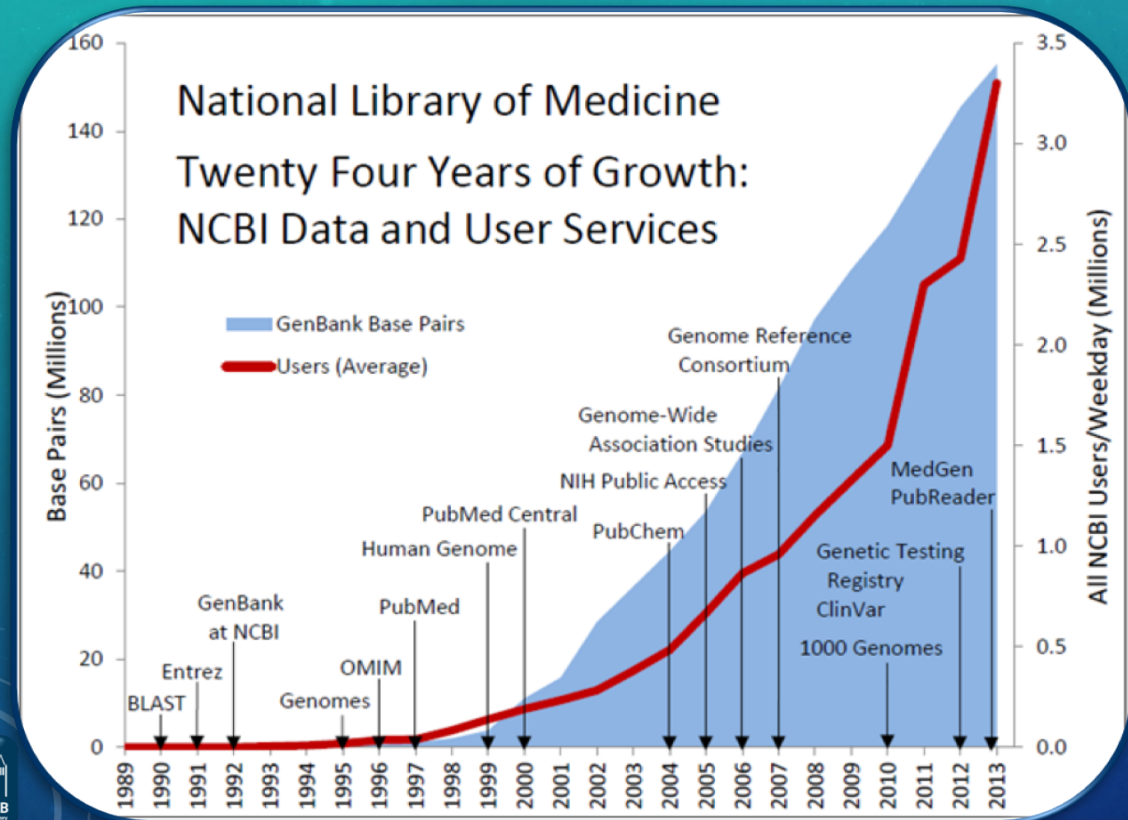
Text availability: Abstract, Free full text, Full text

PubMed

Related searches: dna repair review, cancer dna repair, dna repair genes, dna repair mechanisms, mitochondrial dna repair

<http://www.ncbi.nlm.nih.gov/pubmed/>

Growing number of available biological data still to come

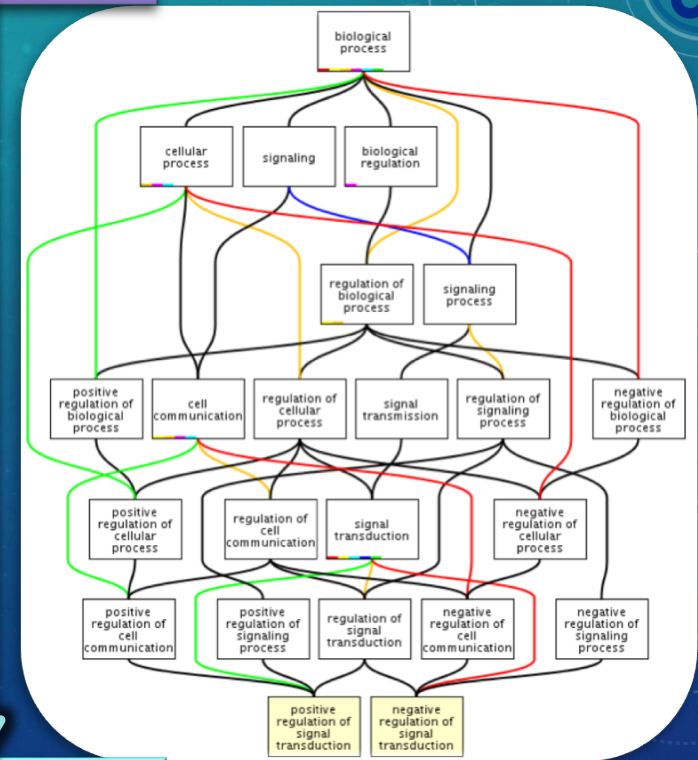


Information expansion in base pairs

Gene Ontology

- A way to capture biological knowledge for individual gene products in written and computable form.
- A set of concepts, and the relationships to each other, arranged as a (non-linear) hierarchy.
- **Gene Ontology Consortium (GOC):** generates and maintains software and databases used to assign function to genes of interest with GO.

Less specific concepts



More specific concepts

Gene Ontology

GeneOntology.org

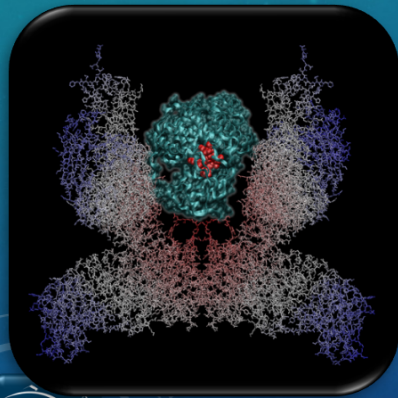


Terms (classes) arranged in a graph: molecular functions, biological processes, cellular locations, and the relationships connecting them all, in a species-independent manner.

1. Molecular Function

An elemental activity or task or job

- protein kinase activity
- insulin receptor activity

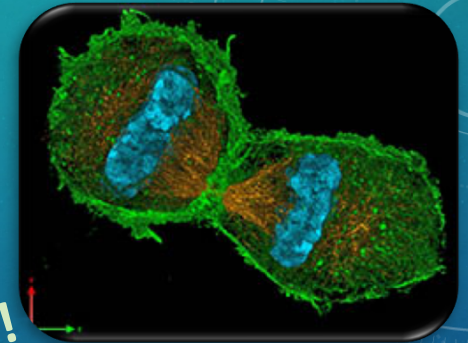


Insulin Receptor
Petrus et al, 2009, *ChemMedChem*
BERKELEY LAB
Lawrence Berkeley National Laboratory

2. Biological Process

A commonly recognized series of events

- cell division



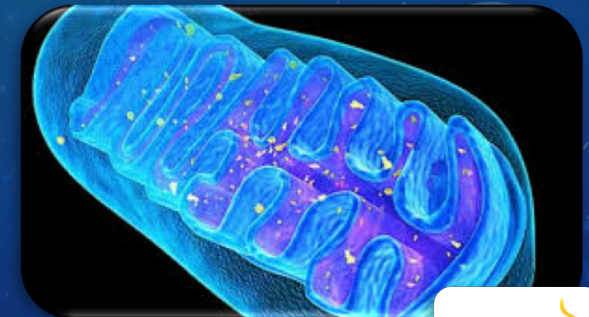
End of Telophase.
Lothar Schermelleh

~150 Contributors!

3. Cellular Component

Where a gene product is located

- mitochondria
- mitochondrial matrix
- mitochondrial inner membrane



Mitochondrion.
PaisekaScience Photo Library

A GO Annotation is:

... a statement that a gene product:

1. has a particular molecular function
or is involved in a particular biological process
or is located within a certain cellular component
2. as determined by a particular method
3. as described in a particular reference

Gene/product	Gene/product name	Qualifier	Direct annotation	Annotation extension	Assigned by	Taxon	Evidence	Evidence with	PANTHER family	Isoform	Reference	Date
<input type="checkbox"/> P02879	Ricin		protein binding		ParkinsonsUK-UCL	Ricinus communis	IPI	UniProtKB:Q8BJT9			PMID:24200403	20150609
<input type="checkbox"/> P02879	Ricin		protein binding	GO:0005515 (go to the term details page for protein binding)	UCL	Ricinus communis	IPI	UniProtKB:Q925U4			PMID:24200403	20150609
<input type="checkbox"/> P02879	Ricin		protein binding		UCL	Ricinus communis	IPI	UniProtKB:Q925U4			PMID:24200403	20150609



Anatomy of a GO term:

The diagram illustrates the structure of a GO term page. A central white box contains the term details, with purple callout boxes and yellow dashed arrows pointing to specific fields:

- Unique Identifier** points to the **Accession** field (GO:0000725).
- Term name** points to the **Name** field (recombinational repair).
- Ontology** points to the **Ontology** field (biological_process).
- Synonyms** points to the **Synonyms** field (None).
- Definition** points to the **Definition** field (A DNA repair process that involves the exchange, reciprocal or nonreciprocal, of genetic material between the broken DNA molecule and a homologous region of DNA. Source: GOC:elh).
- Cross-references** points to the **Related** section, which contains three links: "to all genes and gene products associated to recombinational repair.", "to all direct and indirect annotations to recombinational repair.", and "to all direct and indirect annotations download (limited to first 10,000) for recombinational repair."

Term Information

Accession GO:0000725

Name recombinational repair

Ontology biological_process

Synonyms None

Definition A DNA repair process that involves the exchange, reciprocal or nonreciprocal, of genetic material between the broken DNA molecule and a homologous region of DNA. *Source:* GOC:elh

Comment None

History See term [history for GO:0000725](#) at QuickGO

Subset gosubset_prok

Community [GN Add](#) usage comments for this term on the [GO Nucleus](#) wiki.

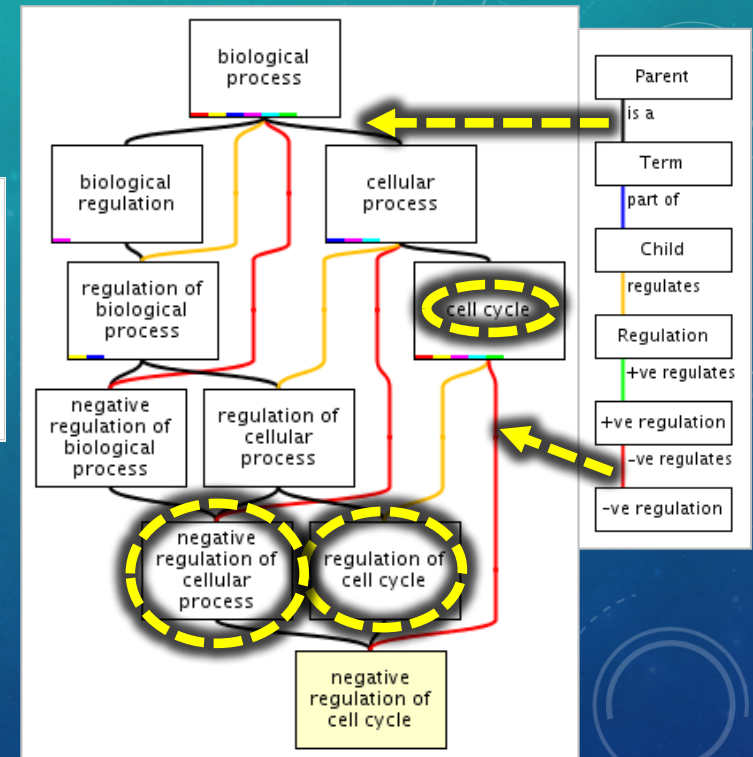
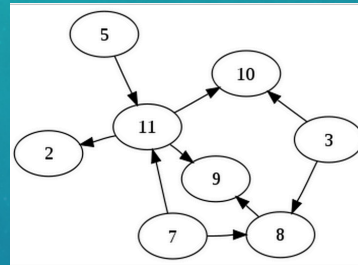
Related

- [Link](#) to all **genes and gene products** associated to recombinational repair.
- [Link](#) to all direct and indirect **annotations** to recombinational repair.
- [Link](#) to all direct and indirect **annotations download** (limited to first 10,000) for recombinational repair.

Feedback Contact the [GO Helpdesk](#) if you find mistakes or have concerns about the data you find here.

Ontology structure

- **Directed acyclic graph**
 - A term can have more than one 'parent'
 - A term can have more than one 'child'
- **Terms are linked by relationships**
 - is_a
 - part_of
 - regulates (& +/-)
 - has_part
 - occurs_in



These relationships allow for complex analysis of large datasets

Access to GO

1. GO Website <https://GeneOntology.org/>



The screenshot shows the Gene Ontology Consortium website. A yellow dashed arrow points from the 'Search GO data' section to the 'Download Annotations' page. The 'Download Annotations' page features a table titled 'Filtered Annotation File Downloads' with the following data:

Species/Database	Gene products annotated	Annotations	date	README	File
Leishmania major Sanger GeneDB	782	2204 (2204 non-IEA)	3/13/2015	README	gene_association.c (72 kb)
Plasmodium falciparum Sanger GeneDB	2373	6250 (6250 non-IEA)	3/10/2015	README	gene_association.c (187 kb)
Trypanosoma brucei Sanger GeneDB	3576	9047 (9047 non-IEA)	3/13/2015	README	gene_association.c (245 kb)

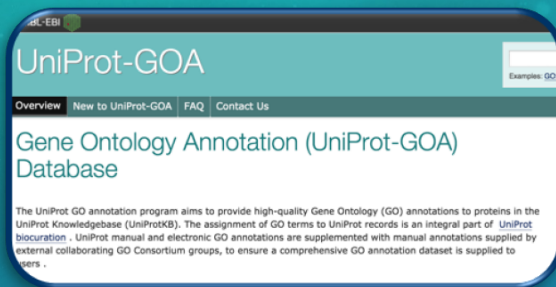
A second yellow dashed arrow points from the 'File' column of the table to the 'Download Annotations' page.

Annotation files store information about associations between a gene product and an ontology term.

There is one annotation file per species

Access to GO

2. UniProt-GOA



<http://www.ebi.ac.uk/GOA>

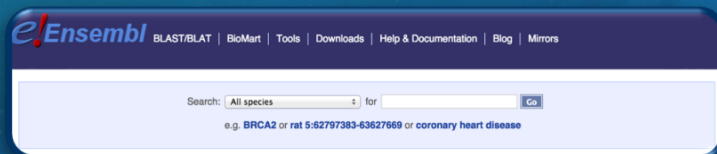


3. UniProtKB



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

4. Ensembl



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

5. NCBI Gene



<http://www.ncbi.nlm.nih.gov/gene/>



GO Browsers

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



The screenshot shows the AmiGO 2 web interface. At the top, there is a navigation bar with links for Home, Search, Tools & Resources, Help, Feedback, About, and AmiGO 1.8. The main content area is divided into several sections:

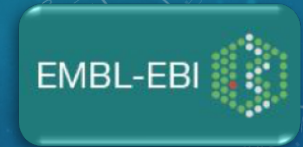
- AmiGO 2**: A central header with a dashed red box around it.
- Quick search**: A search bar with a "Search" button and a link for "more information on quick search".
- Grebe**: A section titled "Get Started with Grebe" featuring a grebe icon and a "Go »" button. A callout box explains: "Grebe: assisting interface for AmiGO2 with 'fill in the blank'." An arrow points from the callout to the Grebe section.
- Simple search**: A section titled "Advanced Search" featuring a magnifying glass icon and a "Search" button. A callout box explains: "Simple search: with auto-complete." An arrow points from the callout to the Simple search section.
- GOOSE**: A section featuring a goose icon and a "Go »" button. A callout box explains: "GOOSE: interrogate the GO database using SQL." An arrow points from the callout to the GOOSE section.
- Term Enrichment Service**: A section featuring a network diagram icon and a "Submit" button. A callout box explains: "Term Enrichment Service. Powered by PANTHER." An arrow points from the callout to the Term Enrichment Service section.
- Statistics**: A section featuring a bar chart icon and a "Go »" button.
- And Much More...**: A section featuring a wrench and screwdriver icon and a "Go »" button.



GO Browsers

<http://www.ebi.ac.uk/QuickGO/>

The screenshot shows the QuickGO web browser interface. At the top, there is a navigation bar with 'Services', 'Research', 'Training', and 'About us'. The main header features the 'QuickGO' logo and the tagline 'fast browser for Gene Ontology terms and annotations.' Below this, a search bar is prominently displayed with a 'Search!' button and icons for 'Web Services', 'Dataset', and 'Term Basket: 0'. A callout box points to the search bar with the text: 'Simple search: Search GO terms or proteins; with auto-complete.' To the left, a sidebar menu lists various resources like 'Help', 'Reference', 'FAQs', and 'Video tutorials'. A callout box points to this menu with the text: 'Find GO annotation sets: With filters to generate annotation subsets.' The main content area includes sections for 'Search and Filter GO annotation sets', 'Investigate GO slims', and 'View the history of changes to GO'. A callout box points to the 'Investigate GO slims' section with the text: 'Investigate GO Slims: Focused visualization for certain GO subsets.' At the bottom, there is a footer with contact information and a '2ms' indicator.



Functional Annotation

Attaching metadata to structural annotations for the purpose of assigning a particular function.

- Assignments do not necessarily have to be supported by your own experimental data.
- Sequence similarity approaches must be informed and validated by evolutionary theory, not just a score value.

Functional Annotation

Assembly:

- Genome
- Transcriptome

Predicted
DNA / Protein
Sequences

Examples:

- InterProScan
- PANTHER Term Enrichment Tool
- JAMp

Compare against
databases

Relying solely on
sequence similarity

Building protein and domain
profiles, then running
sequence similarity analysis

Enrichment analysis
(expression data)

Functional assignments

Gene Ontology, Uberon anatomy (or stage) Ontology, Chemical Entities of Biological Interest, Kyoto Encyclopedia of Genes and Genomes (KEGG), OrthoDB, Benchmarking Universal Single-Copy Orthologs (BUSCO), etc.



Term Enrichment

<http://geneontology.org/page/go-enrichment-analysis>

For MODs

Powered by



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



Gene Ontology Consortium

Home Documentation Download

Search GO data

terms and gene products

Search

Enrichment analysis (beta)

Your gene IDs here...

biological process

H. sapiens Submit

Advanced options
Powered by PANTHER

<http://GeneOntology.org/>

Term Enrichment Service

AmiGO 2

Your genes here...

biological process

H. sapiens Submit

Powered by PANTHER

Advanced »

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

Term Enrichment Service

Information about Term Enrichment Services

Current documentation and discussion of term enrichment with the GO can be found [here](#).

The Remote Term Enrichment tool is a basic querying and viewing application for services that conform to [TERP](#).

This is a work in progress, so your feedback is appreciated.

Gene IDs

Gene IDs...

Species

Species

H. sapiens

Ontology

Ontology

biological process

Correction

Use Bonferroni correction
 Use no correction

Resource

PANTHER

Results viewer

PANTHER

Submit

Species

Ontology

Correction

Data source, visualization

“Term Enrichment”

If not working with MODs



The screenshot shows the InterProScan web interface. At the top, it says "EBI" and "InterPro Protein sequence analysis & classification". There are navigation tabs for "Home", "Search", "Release notes", "Download", "About InterPro", "Help", and "Contact". Below these are two search options: "By sequence" (selected) and "By domain architecture". The main heading is "InterProScan sequence search". The text below explains that the form allows scanning a protein sequence against the InterPro protein site and provides instructions on FASTA format. A text input field is present with the placeholder "Analyse your protein sequence". At the bottom, there are "Advanced options" and a "Search" button.



- **InterProScan**
Predicts GO terms based on detected domains using our mapping file InterPro-2-GO, one sequence at a time.

<http://geneontology.org/page/download-mappings>

<http://www.ebi.ac.uk/interpro/search/sequence-search>

Download	
InterProScan	
Name	
InterProScan	5.14-53.0



Term Enrichment at The *Arabidopsis* Information Resource (TAIR)

<https://arabidopsis.org>



The screenshot shows the TAIR website homepage. At the top, there is a navigation bar with links for Home, Help, Contact, About Us, Subscribe, Login, and Register. Below this is a secondary navigation bar with buttons for Search, Browse, Tools, Portals, Download, Submit, News, and ABRC Stocks. The main content area features a large banner for NAASC (The North American Arabidopsis Steering Committee) with the text "User survey to gather input on ICAR2020". To the right of the banner, there are sections for "Breaking News" and "Featured Paper".

The Arabidopsis Information Resource

NAASC
The North American Arabidopsis Steering Committee
User survey to gather input on ICAR2020

Breaking News

NAASC community survey
[Nov 6, 2017]
The North American Arabidopsis Steering Committee (NAASC) is soliciting community feedback on the 2020 ICAR meeting. Please contribute your opinions by filling out the [survey](#).

New stocks available from ABRC
[Oct 18, 2017]
HALO-tagged transcription factors for DAP-Seq to identify transcription factor binding sites donated by Joe Ecker (CD4-92).

Featured Paper
[Oct 17, 2017]
Waese, J., et al., (2017) ePlant: Visualizing and Exploring Multiple Levels of Data for Hypothesis Generation in Plant Biology
DOI: [10.1105/tpc.17.00073](https://doi.org/10.1105/tpc.17.00073)

12th public release of TAIR@Phoenix data
[Oct 2, 2017]
12th public release of data curated under TAIR's subscription-based funding model. Files contain new publications, annotations, gene symbols and other data through September 30, 2016.

Mark your calendars
[Oct 2, 2017]

About TAIR

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every week from the latest published research literature and community data submissions. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed and DNA resources of *Arabidopsis thaliana* and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.

Phoenix Bioinformatics TAIR is located at Phoenix Bioinformatics and funded by subscriptions.



Identify terms over or under-represented in a set of genes: in this list of genes, are there any functional classes that are found more often than expected when compared with the reference list?

<https://arabidopsis.org>



Search Browse **Tools** Portals Download Submit News All

GO Term Enrichment for Plants

Statistical Over/Under Representation (powered by PANTHER)

Use this tool to identify Gene Ontology terms that are over or under-represented in a set of genes (for example from co-expression or RNAseq data). The data are sent to the PANTHER Classification System which contains up to date GO annotation data for Arabidopsis and other plant species. the advanced setting if you want to change parameters or explore PANTHER's other tools for a sets of genes. [Help]

Enter a list of valid identifiers, separated by newline. [Try a sample gene list](#)

```
Solyc04g011550.2.1  
Solyc06g068130.2.1  
Solyc06g036170.1.1  
Solyc01g087750.2.1  
Solyc01g112260.2.1  
Solyc12g098890.1.1  
Solyc03g005090.2.1  
Solyc05g005550.2.1  
Solyc03g098020.2.1  
Solyc02g067690.2.1
```

Choose Organism


- ✓ Arabidopsis thaliana
- False Brome (Brachypodium distachyon)
- Chlamydomonas reinhardtii
- Soybean (Glycine max)
- Rice (Oryza sativa)
- Moss (Physcomitrella patens)
- Poplar (Populus trichocarpa)
- Tomato (Solanum lycopersicum)**
- Sorghum (Sorghum bicolor)
- Grape (Vitis vinifera)



Also, capture experimental data about gene functions (GO) and expression (Plant Ontology) using TOAST: The *Arabidopsis* Annotation Submission Tool

<https://arabidopsis.org>



 **tair Online Annotation Submission Tool** Clear Form Logout

Need help? See our [video tutorial](#)

Please specify a Pubmed ID or a DOI for the article.

Pubmed ID (for example, 21051552)

Digital Object Identifier (DOI) (for example, 10.1104/pp.110.166546)

Article ID No article, please enter a DOI or PubMed ID

Use TAIR gene search to find the AGI locus name for your gene symbol.

Locus Name	Symbol	Symbol Full Name
<input type="text" value="AT2G23380"/>	<input type="text" value="CLF"/>	<input type="text" value="CURLY LEAF"/>

You must enter at least one annotation below to be able to submit annotations.

- Molecular Function Annotations**

Molecular Function	Method
<input type="text" value="protein kinase activity"/>	<input type="text" value="enzyme assays"/>
<input type="text"/>	<input type="text" value="Please choose a method."/>
- Biological Process Annotations**

Biological Process	Method
<input type="text" value="seed development"/>	<input type="text" value="biochemical/chemical analysis"/>
<input type="text"/>	<input type="text" value="Please choose a method."/>
- Subcellular Localization Annotations**

Subcellular Localization	Method
<input type="text" value="plasma membrane"/>	<input type="text" value="localization of GFP/YFP fusion protein"/>
<input type="text"/>	<input type="text" value="Please choose a method."/>
- Expression Annotations**

Expression	Method
<input type="text" value="hypocotyl"/>	<input type="text" value="transcript levels (e.g. RT-PCR)"/>
<input type="text"/>	<input type="text" value="Please choose a method."/>
- Interacting Partner Annotations**

Interacting Partner	Method
<input type="text" value="At1g01040"/>	<input type="text" value="enzyme assays"/>
<input type="text"/>	<input type="text" value="Please choose a method."/>
- Other Comments**

Contributing to GO



Ad hoc curation efforts

A handful of annotations from manual efforts.



Annotations



Large sets of annotations



Ontology

terms, relationships, etc.



genes, proteins



articles

<http://geneontology.org/page/contributing-go>



Noctua



Collaboratively curating gene functions



GO Annotation

http://amigo.geneontology.org/amigo/gene_product/UniProtKB:P20719



Total annotations: 33; showing: 1-25
Results count

«First <Prev Next> Last» [Download \(up to 100000\)](#)



<input type="checkbox"/>	Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		RNA polymerase II core promoter proximal region sequence-specific DNA binding		NTNU_SB	Homo sapiens	IDA		family not named pthr24326		PMID:10879542
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding		NTNU_SB	Homo sapiens	IDA		family not named pthr24326		PMID:10879542
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		respiratory system process		Ensembl	Homo sapiens	IEA	UniProtKB:P09021 ensembl:ENSMUSP00000039012	family not named pthr24326		GO_REF:0000107
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		DNA binding		UniProt	Homo sapiens	IDA		family not named pthr24326		PMID:8657138
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		transcription factor activity, sequence-specific DNA binding	has_direct_input UniProtKB:P04637	UniProt	Homo sapiens	IDA		family not named pthr24326		PMID:10879542
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		transcription factor activity, sequence-specific DNA binding		UniProt	Homo sapiens	IDA		family not named pthr24326		PMID:16756717
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		protein binding		UniProt	Homo sapiens	IPI	UniProtKB:Q15672	family not named pthr24326		PMID:15545268
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		nucleus		UniProt	Homo sapiens	IDA		family not named pthr24326		PMID:15545268
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		transcription from RNA polymerase II promoter		GOC	Homo sapiens	IEA	GO:0001077	family not named pthr24326		GO_REF:0000108
<input type="checkbox"/>	HOXA5	Homeobox protein Hox-A5		anterior/posterior pattern specification		Ensembl	Homo sapiens	IEA	UniProtKB:P09021 ensembl:ENSMUSP00000039012	family not named pthr24326		GO_REF:0000107

GO Annotation

http://amigo.geneontology.org/amigo/gene_product/UniProtKB:P35453



Total annotations: 22; showing: 1-10
Results count

«First <Prev Next> Last» [Download \(up to 100000\)](#)



<input type="checkbox"/>	Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		RNA polymerase II core promoter proximal region sequence-specific DNA binding		Ensembl	Homo sapiens	IEA	UniProtKB:P70217 ensembl:ENSMUSP00000001872	family not named pthr24326		GO_REF:0000107
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding		Ensembl	Homo sapiens	IEA	UniProtKB:P70217 ensembl:ENSMUSP00000001872	family not named pthr24326		GO_REF:0000107
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding		UniProt	Homo sapiens	IMP		family not named pthr24326		PMID:24789103
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		skeletal system development		Ensembl	Homo sapiens	IEA	UniProtKB:P70217 ensembl:ENSMUSP00000001872	family not named pthr24326		GO_REF:0000107
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		DNA binding		UniProt	Homo sapiens	IDA		family not named pthr24326		PMID:26581570
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		chromatin binding		Ensembl	Homo sapiens	IEA	UniProtKB:P70217 ensembl:ENSMUSP00000001872	family not named pthr24326		GO_REF:0000107
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		transcription factor activity, sequence-specific DNA binding		UniProt	Homo sapiens	ISS	UniProtKB:P70217	family not named pthr24326		GO_REF:0000024
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		nucleus		HPA	Homo sapiens	IDA		family not named pthr24326		GO_REF:0000052
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		regulation of transcription, DNA-templated		PINC	Homo sapiens	TAS		family not named pthr24326		PMID:9207113
<input type="checkbox"/>	HOXD13	Homeobox protein Hox-D13		transcription from RNA polymerase II promoter		PINC	Homo sapiens	TAS		family not named pthr24326		PMID:8614804

Strengths of classic GO annotation

- Simplicity
- Basically just ‘tagging’ genes with terms
- You can do GO annotation (sort of) in a spreadsheet
- Easy to process computationally
- 100s of tools

Limitations of classic GO annotation

- Classic GO annotations don't describe how genes work together
- Every annotation is independent
- Limited ability to use other ontologies:
 - E.g. PO for plant anatomy and cell types
 - TO for traits

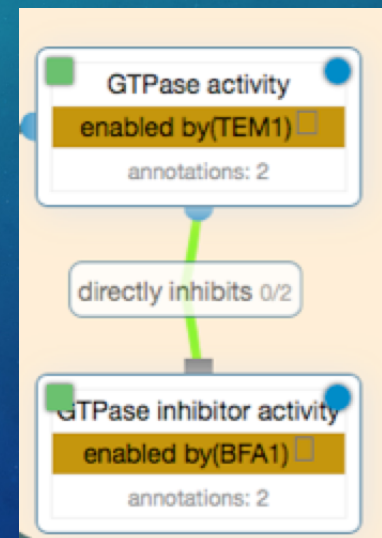
Solution:

Linked expressions in the Gene Ontology (LEGO)

Classic Annotation

Gene	Term	Evidence
TEM1	GTPase activity	IDA
	...	
BFA1	GTPase inhibitor activity	IDA
	...	

LEGO



A data model for causal ontology annotations: “LEGO”

Activity
GO:nnnnnnn

What: <molecule>

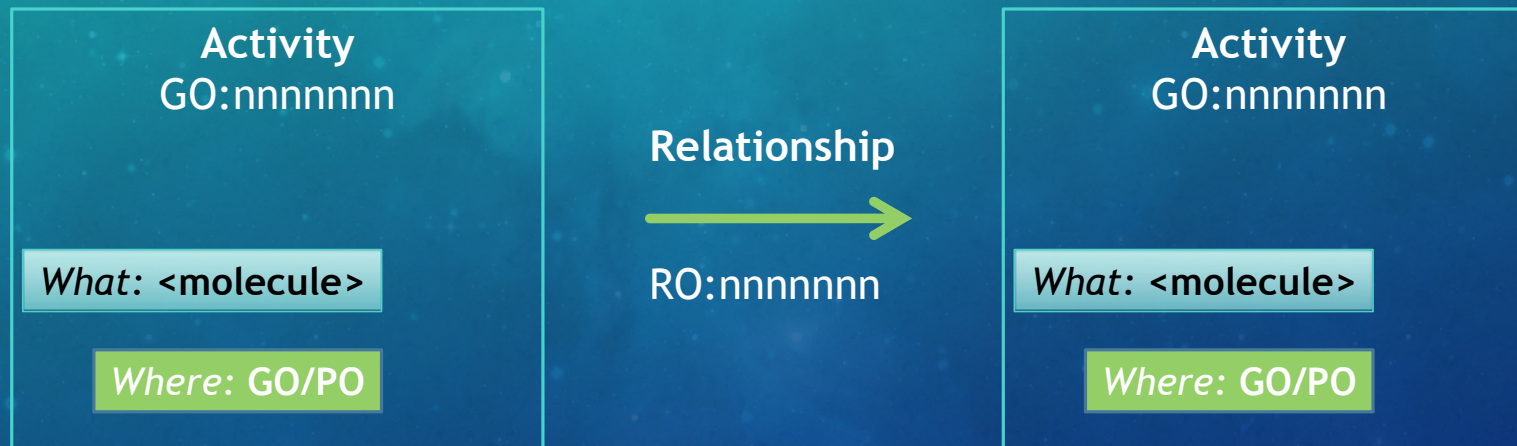
A data model for causal ontology annotations: “LEGO”

Activity
GO:nnnnnnn

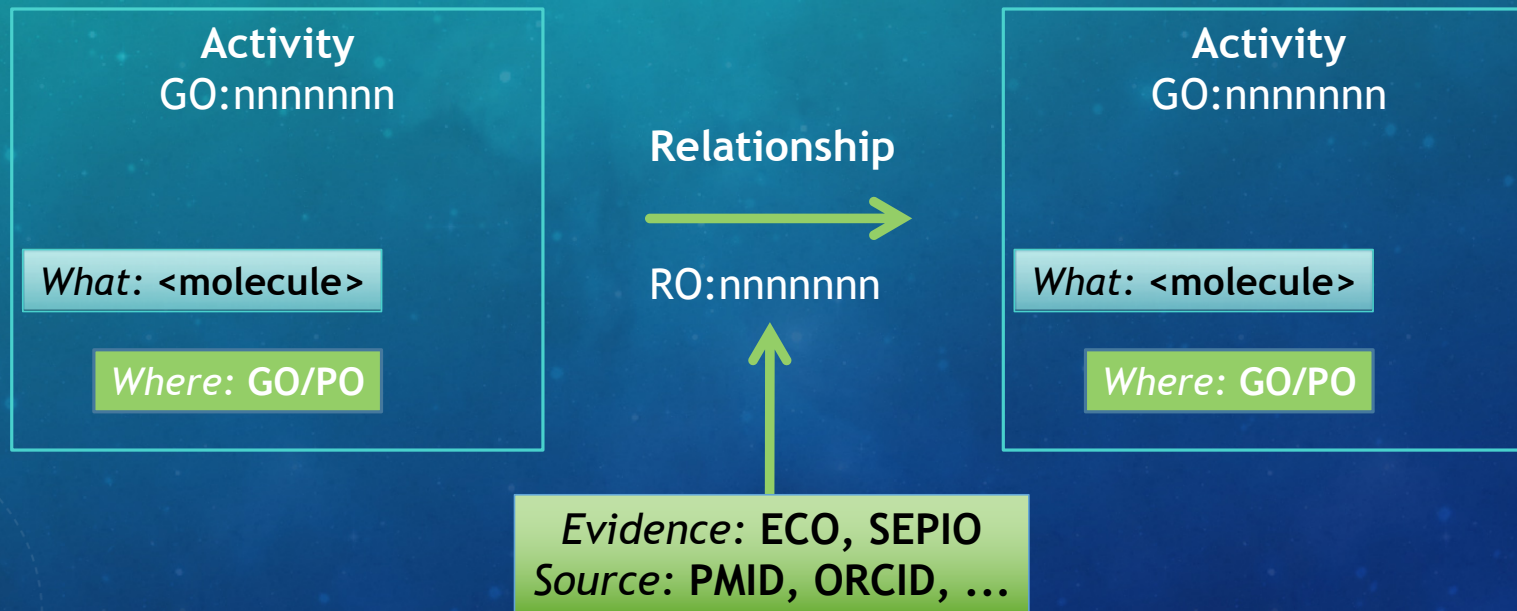
What: <molecule>

Where: GO/PO

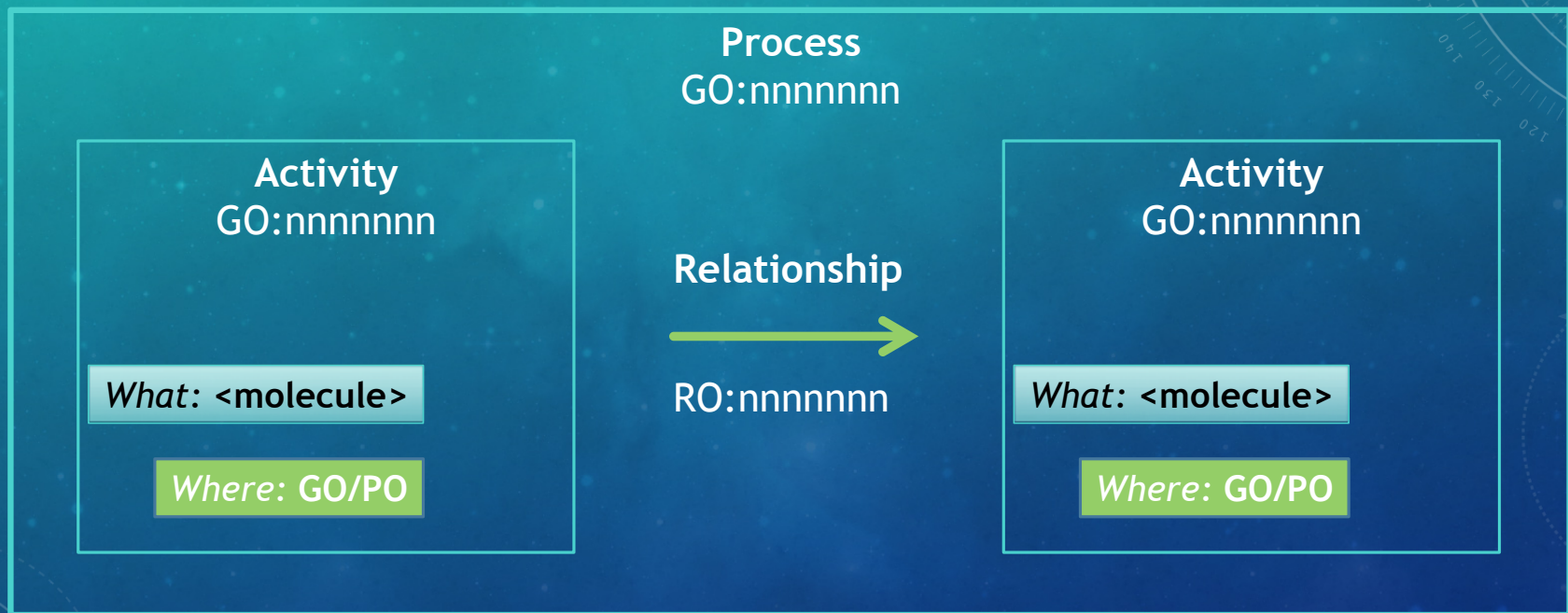
A data model for causal ontology annotations: “LEGO”



A data model for causal ontology annotations: “LEGO”



A data model for causal ontology annotations: “LEGO”



A data model for causal ontology annotations: “LEGO”

GTPase inhibitor activity
GO:0005095

What: **BFA1** S000003814

Where: spindle pole
GO:0000922

GTPase activity
GO:0003924

What: **TEM1** S000004529

Where: spindle pole
GO:0000922

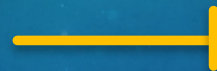
A data model for causal ontology annotations: “LEGO”

Exit from mitosis
GO:0010458

GTPase inhibitor activity
GO:0005095

What: BFA1 S000003814

Where: spindle pole
GO:0000922



GTPase activity
GO:0003924

What: TEM1 S000004529

Where: spindle pole
GO:0000922

Add annoton

enabled by

molecular_function

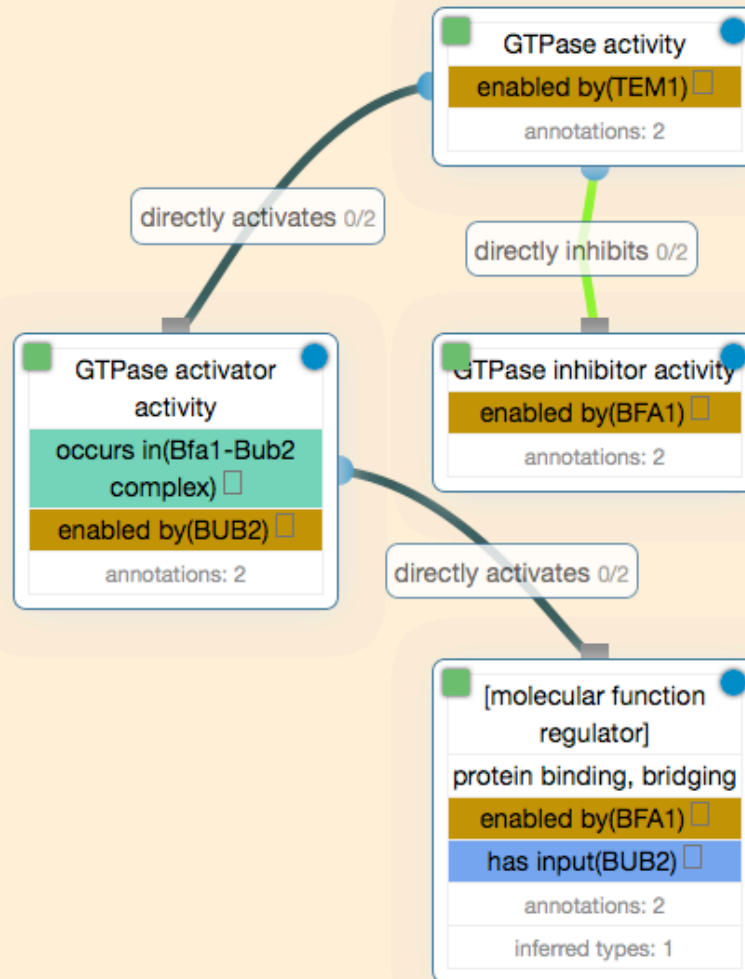
biological_process

cellular_component

Add

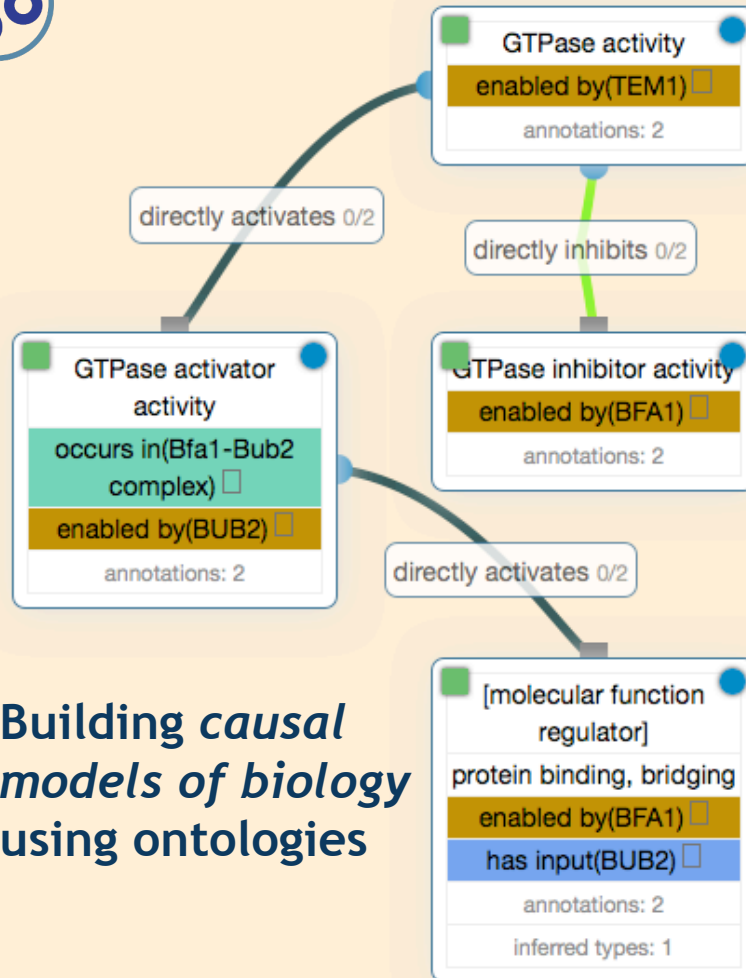
Add function

Add process



<http://noctua.berkeleybop.org/>

Noctua



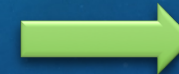
Building *causal models of biology* using ontologies

<http://noctua.berkeleybop.org/>

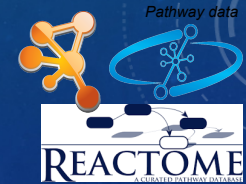
Collaborative Editing!



RDF/OWL Semantic Representation



-Reasoning
-Linked data



Gene sets



Noctua: Current Workflow

- GO curators create models using Noctua
- A lossy version of this is exported as GAF/GPAD
 - All existing GO tools and databases can use and display this
 - Does not include full precision of Noctua model
- The complete version is stored in a graph database
 - Can be browsed using AmiGO (alpha)
 - Can be exported as cytoscape etc.
 - Awaiting the next generation of analytic tools....!
- Upcoming improvements.

Improving the quality of annotated genomes requires:

- 1) Increasing researchers' efficiency by providing a suite of integrated curation tools, and
- 2) Increasing the effective population of researchers by providing universally accessible tools.

SHARING THE LOAD OF
DATA ANALYSIS DISTILLS
VALUABLE KNOWLEDGE!



Lessons learned:

Plan from the start



❖ Experimental design:

- which are the most interesting, “burning” biological matters to explore?
- what questions do we wish to answer?
- which data do we need to answer these questions?
- what are the best strategies to obtain these data?

❖ Data capture:

- Agree on minimum standards for efficiently capturing data, even before data capture begins.
 - *Cultural shift*: Not so common to rely on big genome centers. Not all data come from large repositories, but mostly from GFF3s (from sequencing centers & individual labs).

❖ Long-term housing for these data; storage & dissemination:

- where will the data live?
- who will pay for resources?
- large investment in hardware, and longer-term investment on maintenance.



Lessons learned:

Plan from the start



- ❖ **Link to and/or add to existent databases:**
Find your 'home' database, AND / or deposit to large public resource (NCBI, DDBJ, Ensembl, etc.)
- ❖ **Seek IT support at hosting institution(s) now:**
Engage reliable and responsive systems administrators.



For your attention, Thank You.



Collaborators

Berkeley Bioinformatics Open-Source Projects, Environmental Genomics & Systems Biology, Lawrence Berkeley National Laboratory

[Chris Mungall](#)

[Suzanna Lewis](#)

Seth Carbon (Noctua / AmiGO)

Nathan Dunn (Apollo)

- Ian Holmes, Eric Yao, UC Berkeley (JBrowse)
- Chris Elisk, Deepak Unni, U of Missouri (Apollo)
- Paul Thomas, USC (Noctua)
- Monica Poelchau, USDA/NAL (Apollo)
- Gene Ontology Consortium
- i5k Community

Funding

- Work for GOC is supported by NIH grant 5U41HG002273-14 from NHGRI.
- Apollo is supported by NIH grants 5R01GM080203 from NIGMS, and 5R01HG004483 from NHGRI.
- BBOP is also supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231
- TAIR is a Flexible Access Database, supported by the scientific community.

berkeleybop.org



Berkeley
UNIVERSITY OF CALIFORNIA



Once upon a time...



Undisclosed

