“

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

BERKELEY LAB

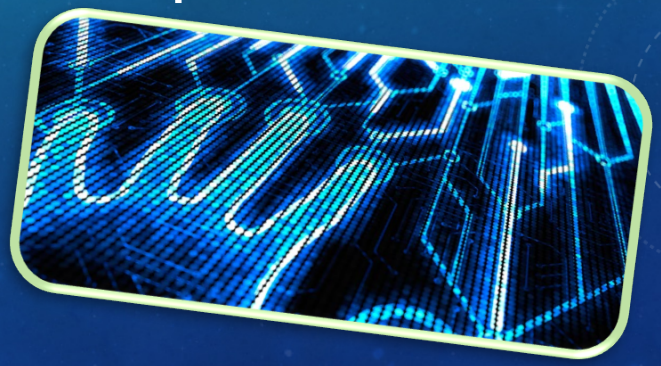
U.S. DEPARTMENT OF ENERGY

Berkeley UNIVERSITY OF CALIFORNIA

Phoenix BIOINFORMATICS

Today…

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

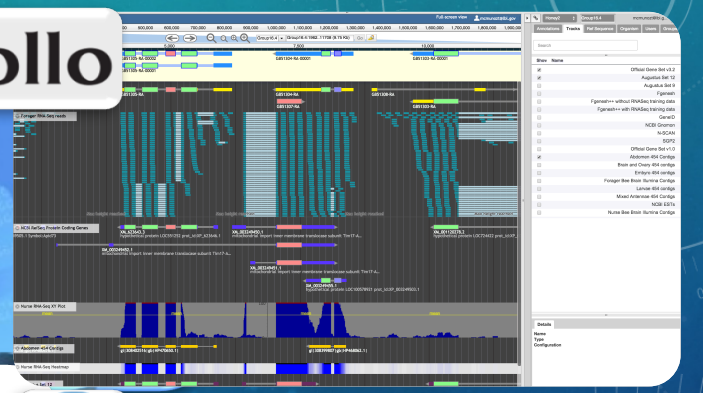


Moni Munoz-Torres, PhD

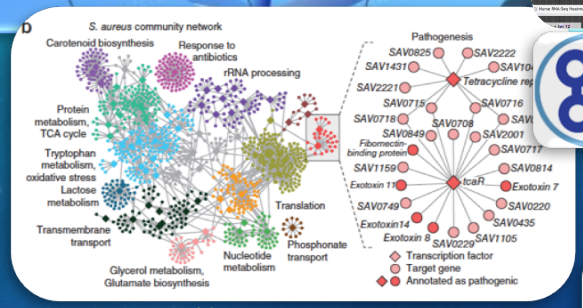
biocurator, genomicist, insect-biased

@monimunozto

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

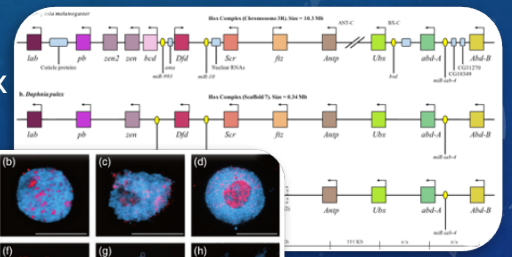


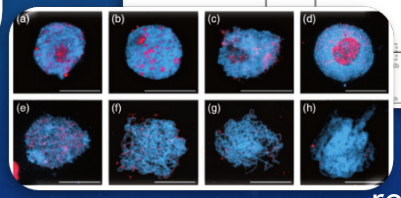
http://GenomeArchitect.org



http://GeneOntology.org

Hox





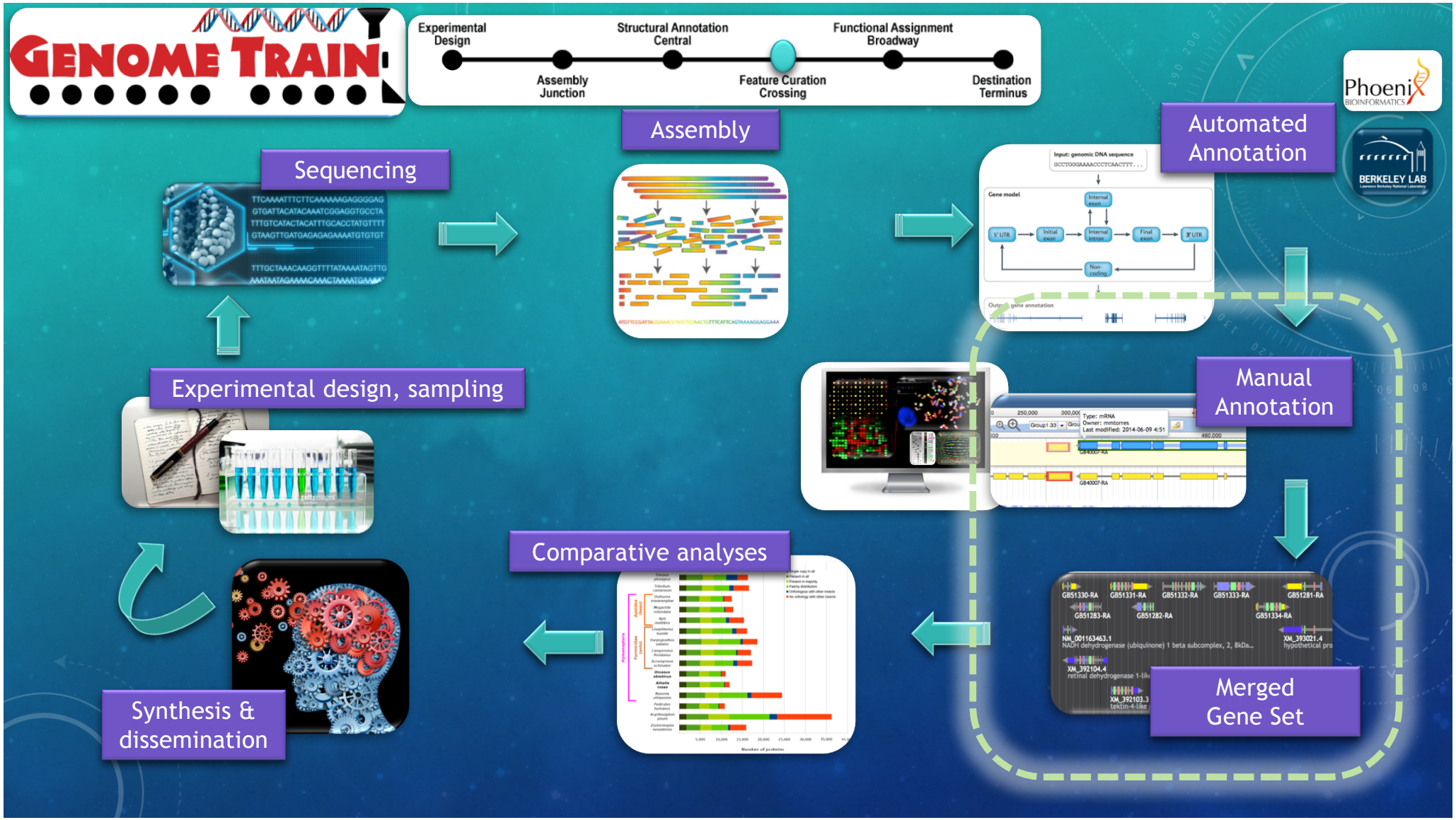
DNA Repair, recombination

Phoenix BIOINFORMATICS

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY

Berkeley UNIVERSITY OF CALIFORNIA


GENOME TRAIN
Experimental Design
Assembly Junction
Structural Annotation Central
Feature Curation Crossing
Functional Assignment Broadway
Destination Terminus
Phoenix BIOINFORMATICS
BERKELEY LAB
Sequencing
Assembly
Automated Annotation
Manual Annotation
Merged Gene Set
Comparative analyses
Experimental design, sampling
Synthesis & dissemination

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

?

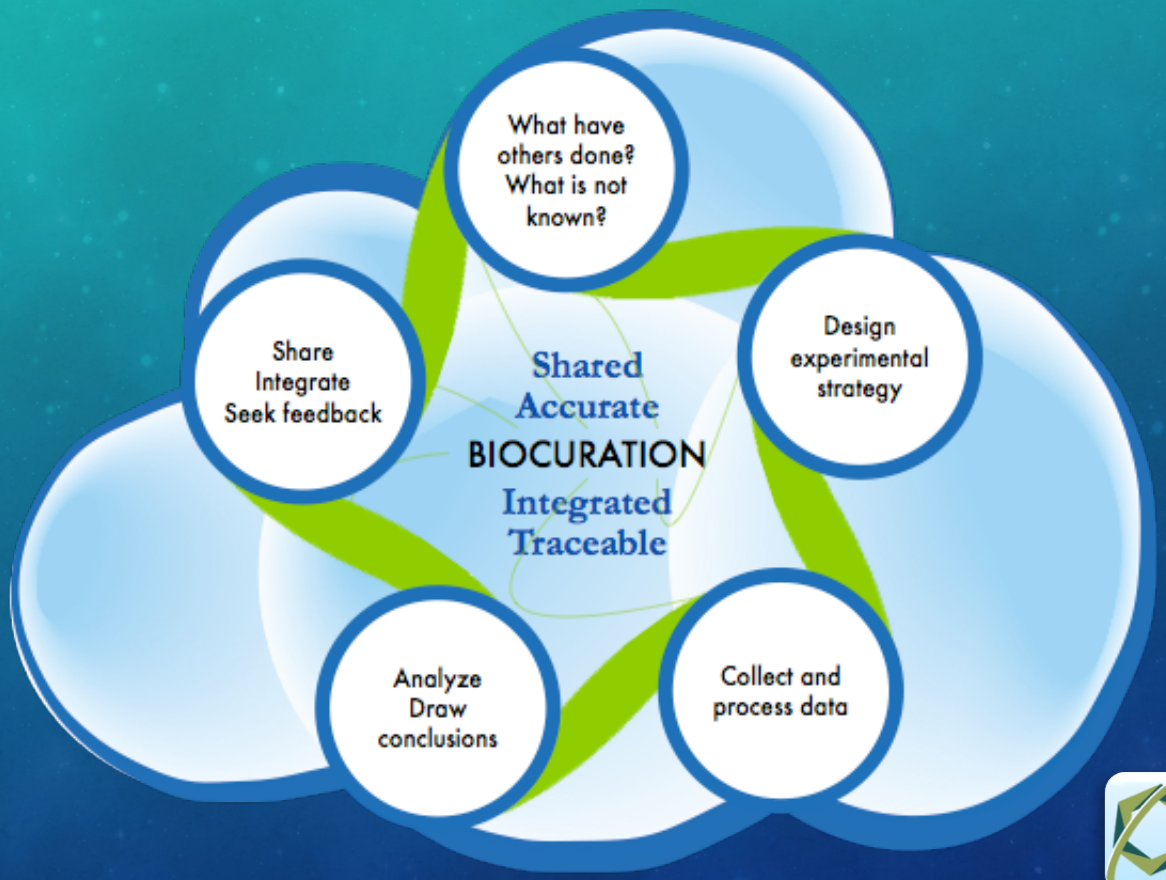
Curation

?

Biocurators

knowledge

Biocuration is a knowledge-extraction process



LIFE SCIENCES KNOWLEDGE

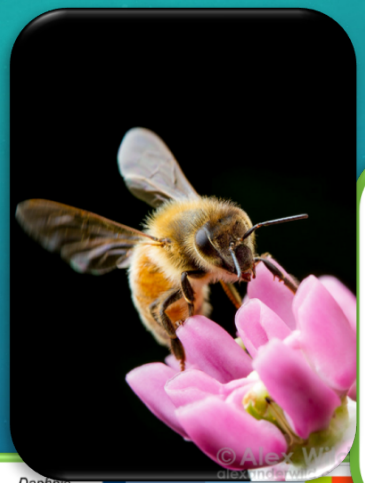
Figure is not yet public. Do not reproduce.

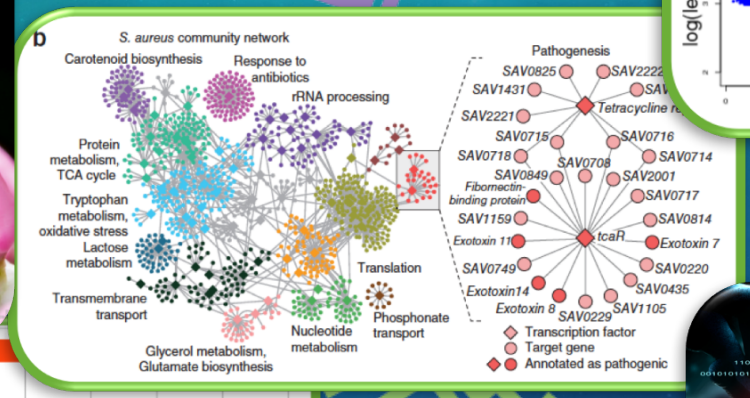
ISB International Society for Biocuration

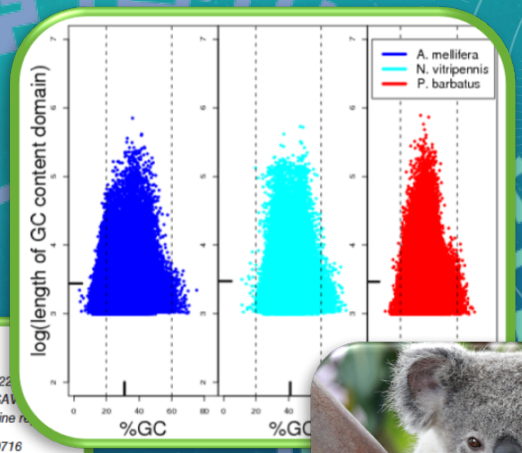
biocuration.org

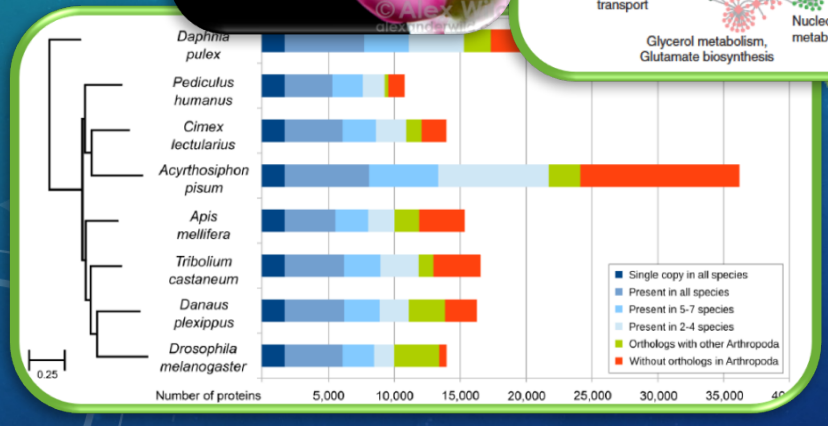
Genome Curation

Unlocking genomes









Marbach et al. 2011. *Nature Methods* | Shutterstock.com | Alexander Wild

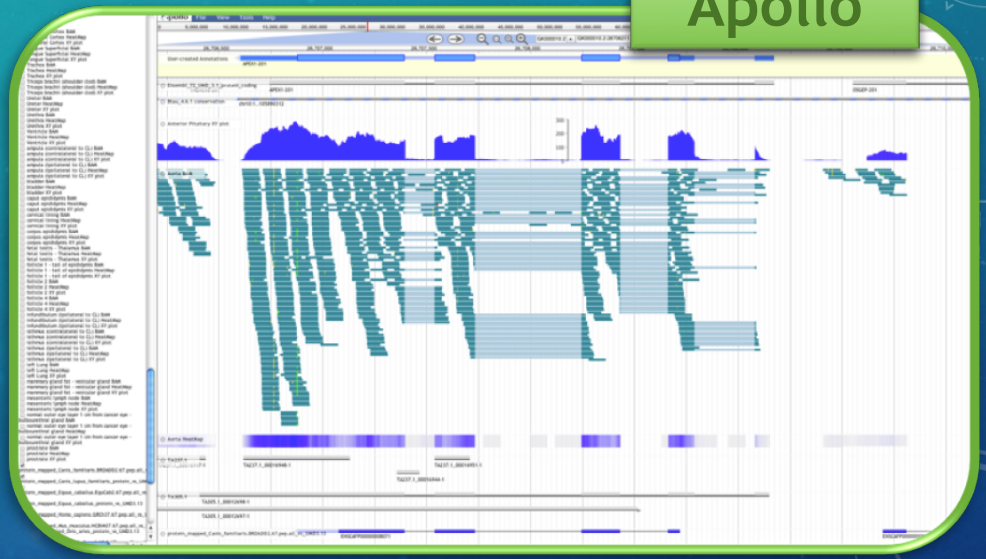
Good genes are required!

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.

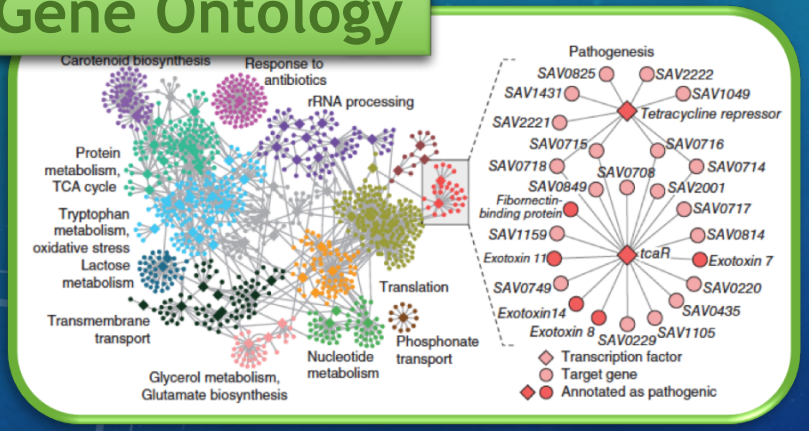
Curation improves quality

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

Apollo



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.

SCALE

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

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

BERKELEY LAB

Phoenix BIOINFORMATICS

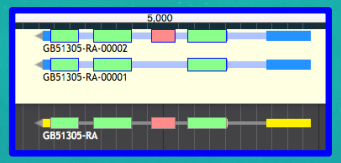
Collaboratively curating gene structures

Apollo Genome Annotation Editor

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

GenomeArchitect.org

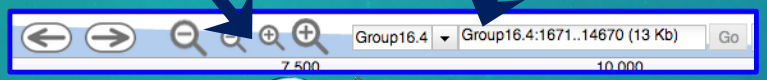
Color by CDS frame, toggle strands, set color scheme and highlights.



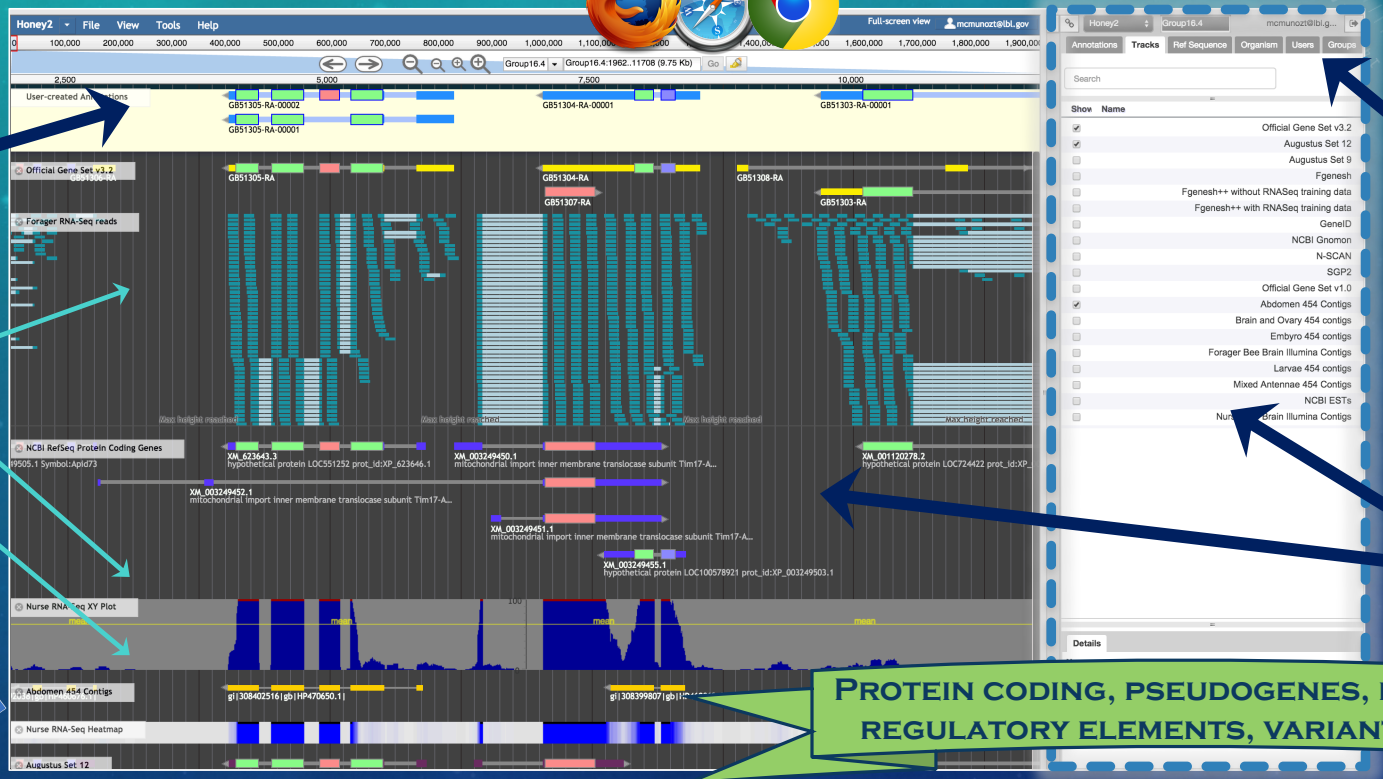
Query the genome using **BLAT.**

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.

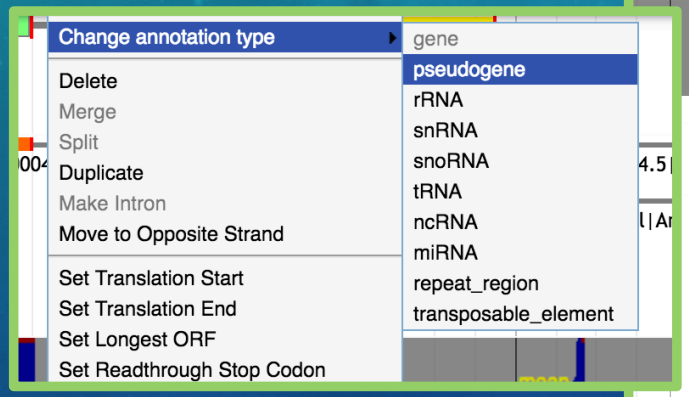
Apollo JBrowse

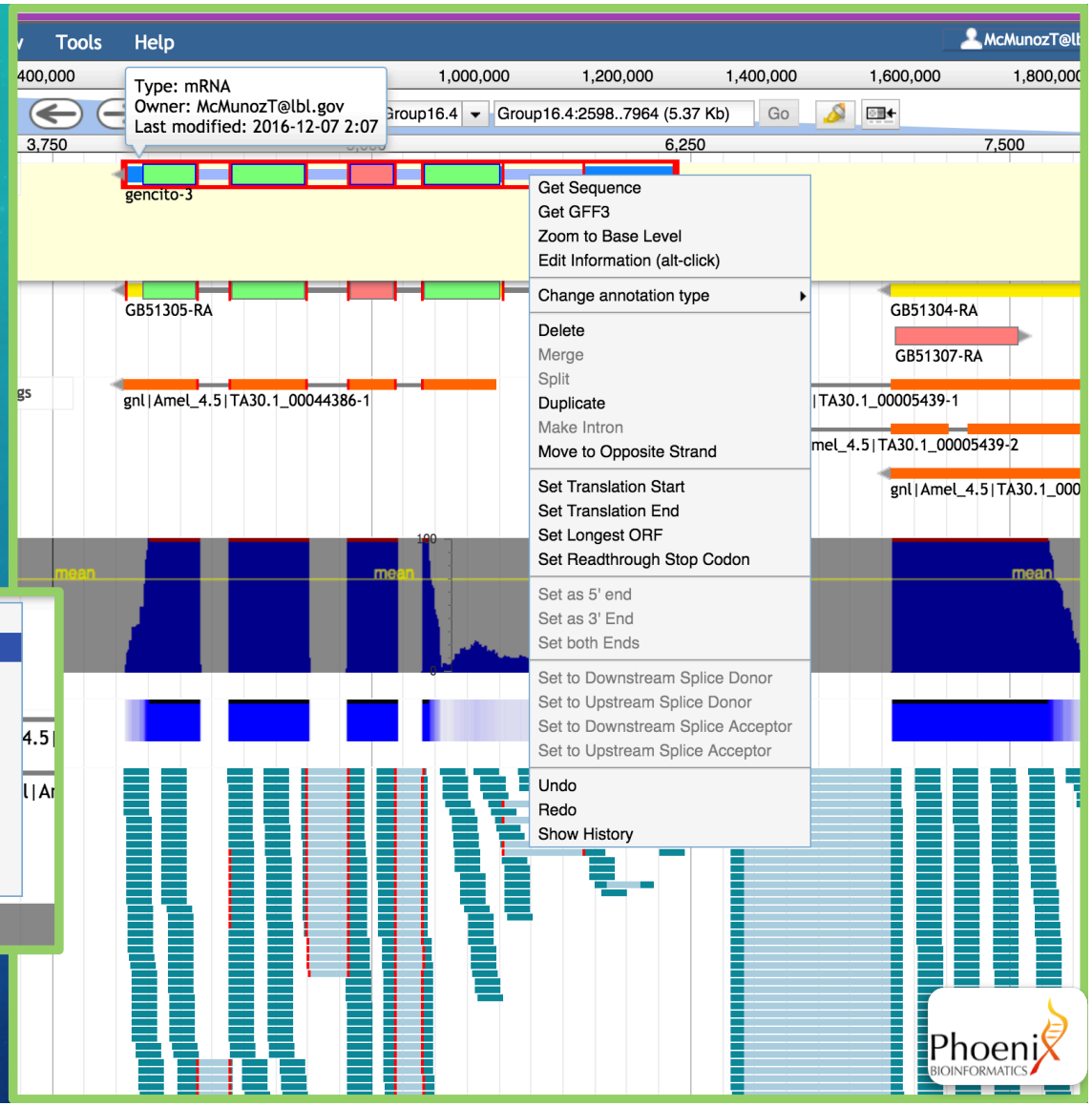
BERKELEY LAB

Phoenix BIOINFORMATICS

Apollo

Right-click functionality





Apollo

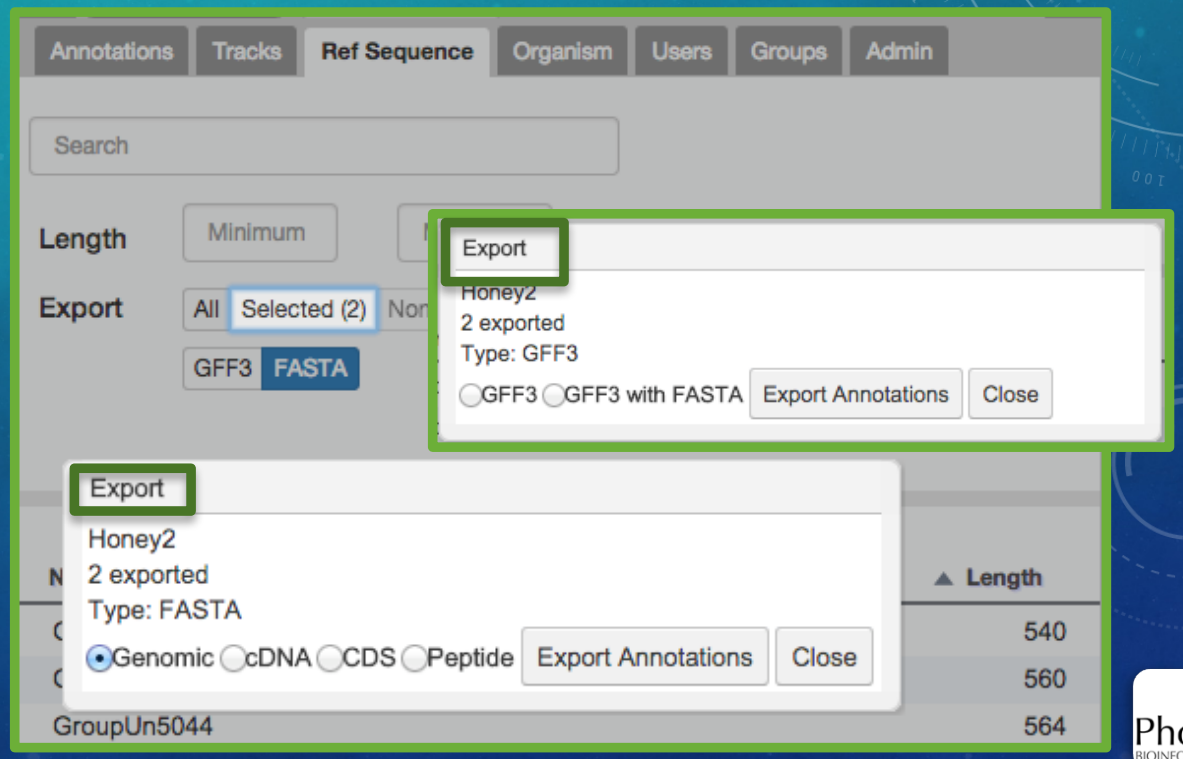
BERKELEY LAB

GenomeArchitect.org

Phoenix BIOINFORMATICS

Apollo

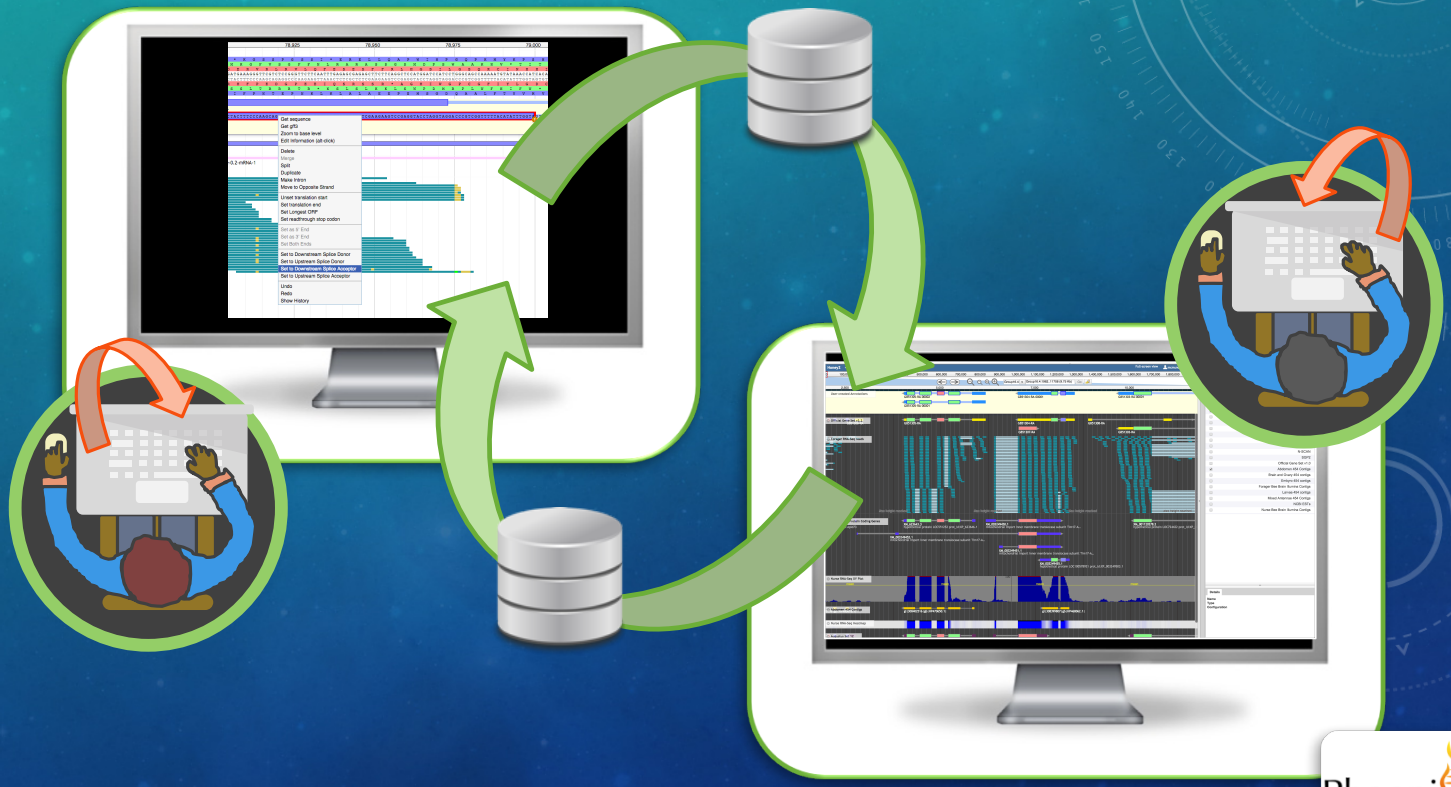
Export



GenomeArchitect.org

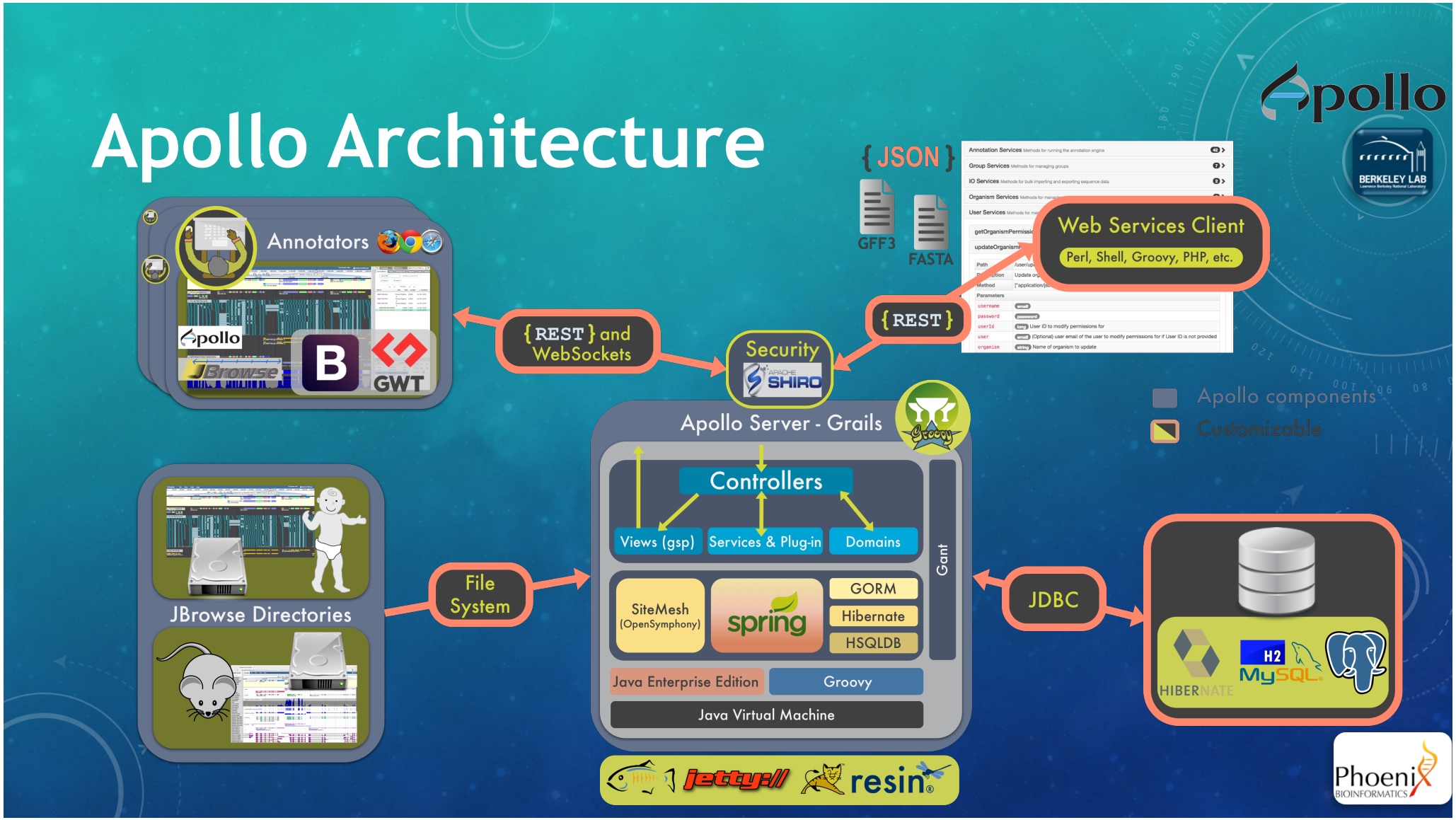
Apollo

Collaboration in Real Time



GenomeArchitect.org

Apollo Architecture

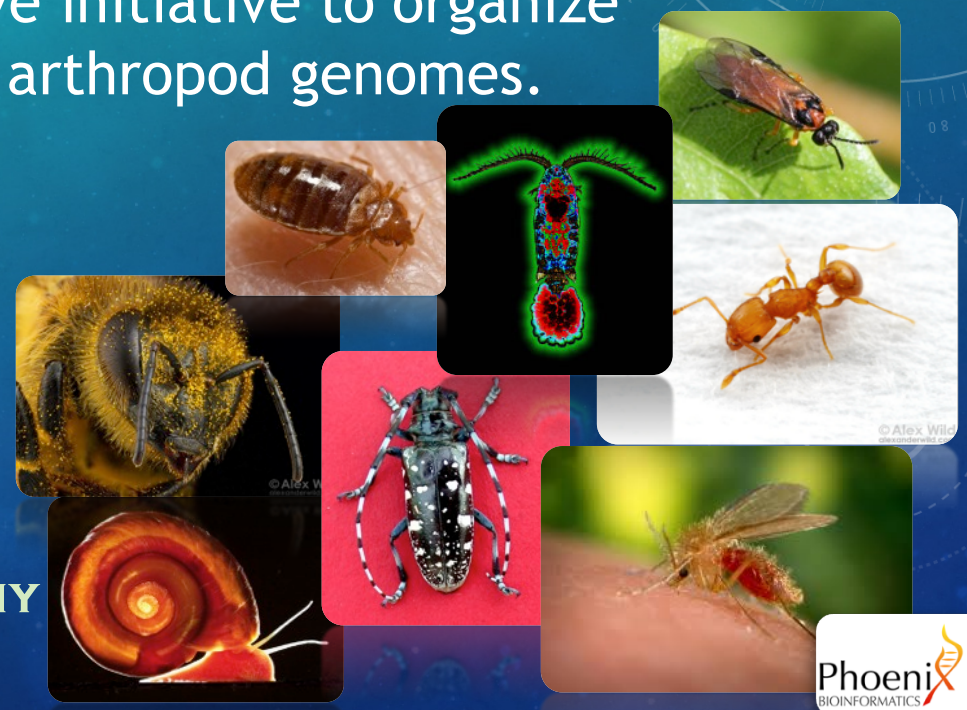


i5k - five thousand arthropod genomes

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



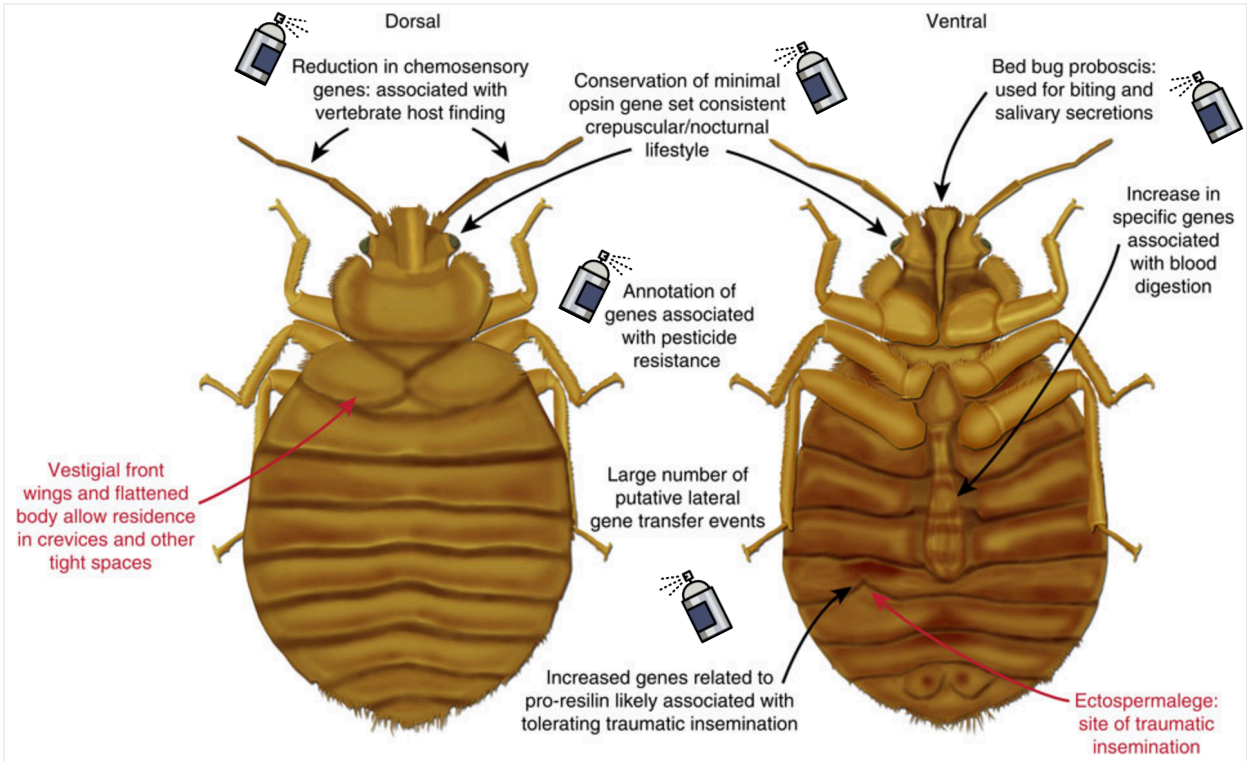
- International Travel and Commerce
- Increased Insecticide Resistance

~80 Curators!

i5k.github.io

The bed bugs, they're back!

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



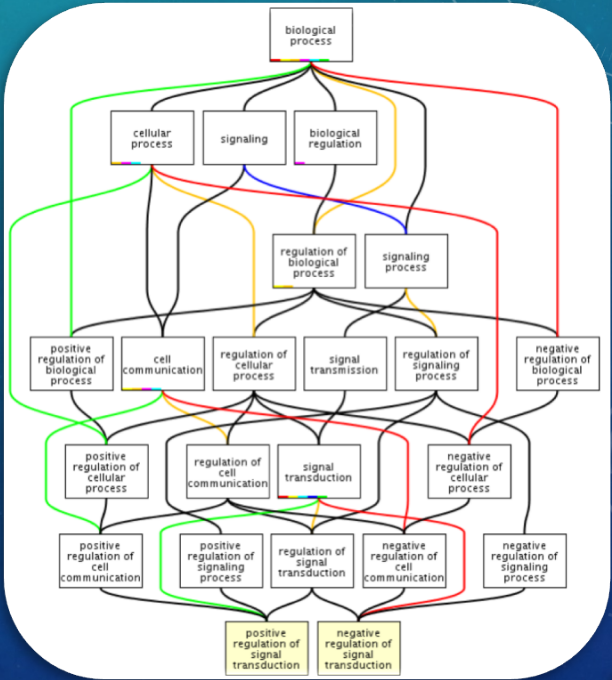
Red, general characteristics of bed bugs; black, key aspects identified and expanded by genome sequencing and manual curation.

- 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



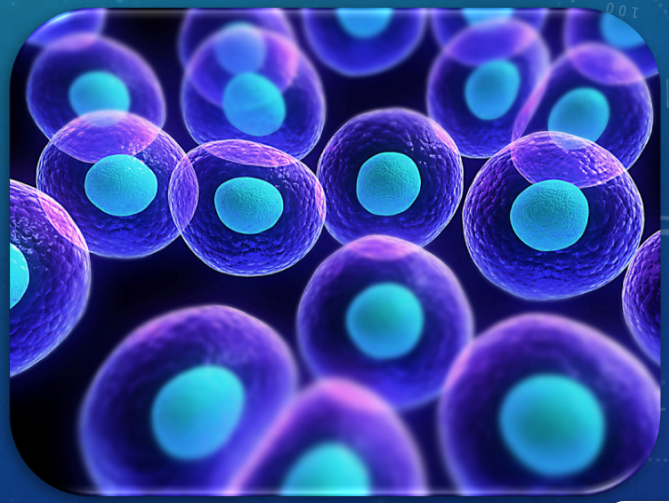
Language inconsistencies in biology

The same name for two different concepts

Cell

www.robertpaterson.com

or



www.biology.usf.edu

BERKELEY LAB

Phoenix BIOINFORMATICS

Language inconsistencies in biology

Two or more different names for the same concept

Eggplant

Brinjal

www.care2.com

Aubergine

Melongene

The same is true for biological concepts:

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

BERKELEY LAB

Phoenix BIOINFORMATICS

Growing number of available biological data

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | dna repair

Create RSS Create alert Advanced

Article types
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

PubMed

Summary 20 per page Sort by Most Recent

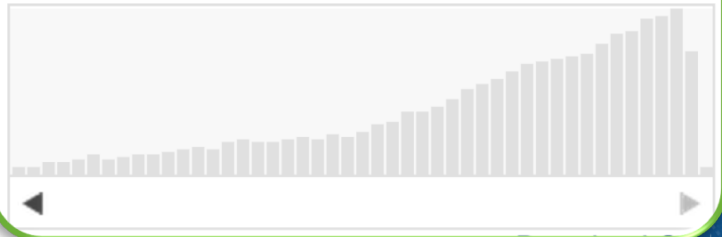
Search results

Items: 1 to 20 of 81894

<< First < Prev Page 1 of 4095 Next > Last >>

1. CITED2 silencing sensitizes cancer cells to cisplatin by inhibiting p53 trans-activation and chromatin relaxation on the ERCC1 **DNA repair** gene.
Liu YC, Chang PY, Chao CC.
leic Acids Res. 2015 Sep 17. pii: gkv934. [Epub ahead of print]
: 26384430
r articles

Results by year



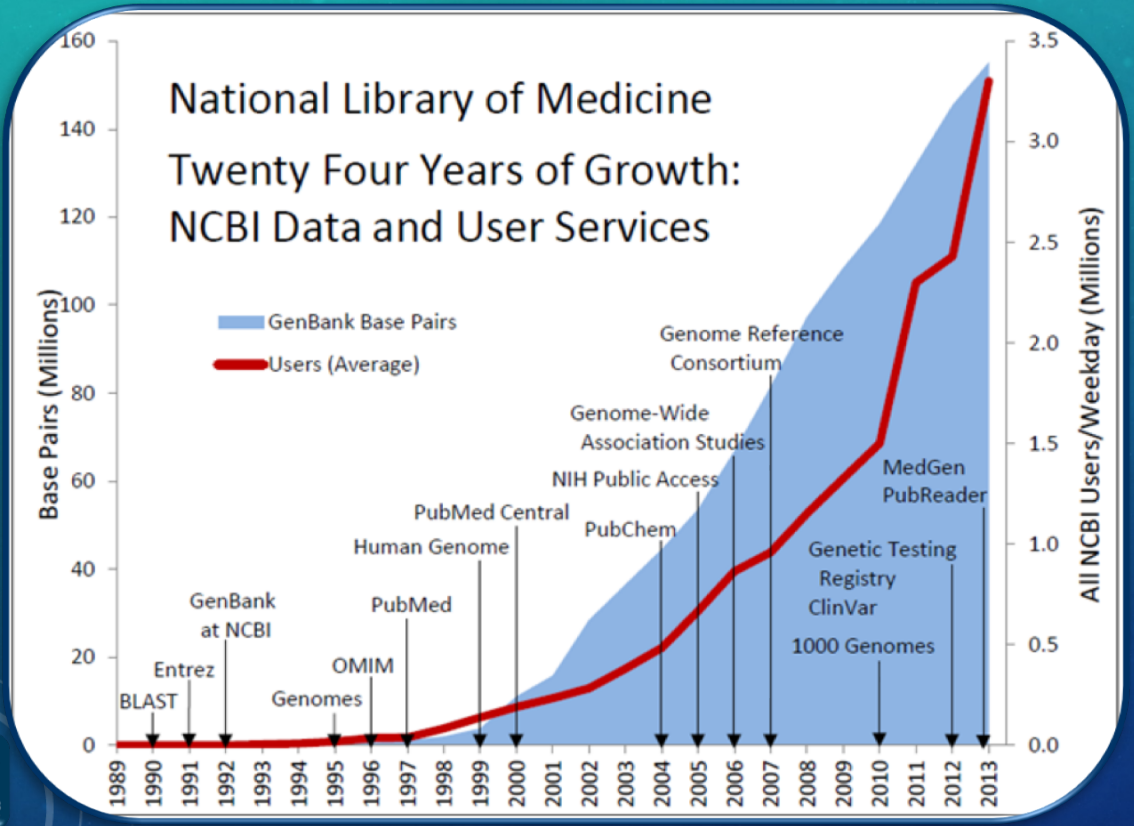
Related searches
dna repair review
cancer **dna repair**
dna repair genes
dna repair mechanisms
mitochondrial **dna repair**

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

BERKELEY LAB

Phoenix BIOINFORMATICS

Growing number of available biological data still to come



Information expansion in base pairs

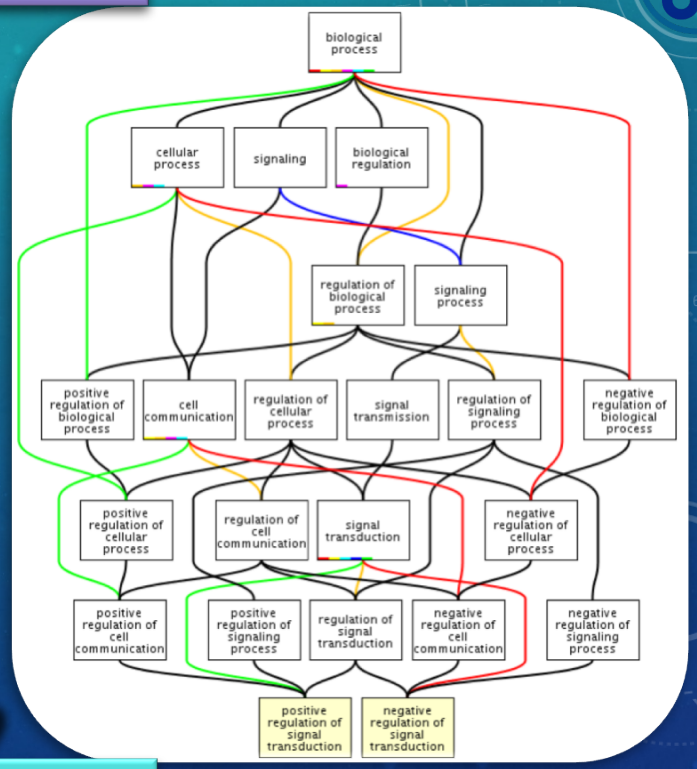
Gene Ontology

- 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

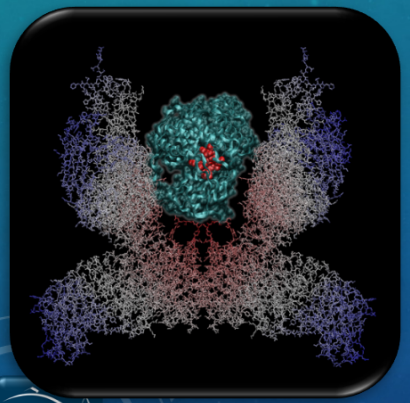
Terms (classes) arranged in a graph: molecular functions, biological processes, cellular locations, and the relationships connecting them all, in a species-independent manner.

~150 Contributors!

1. Molecular Function

An elemental activity or task or job

- protein kinase activity
- insulin receptor activity

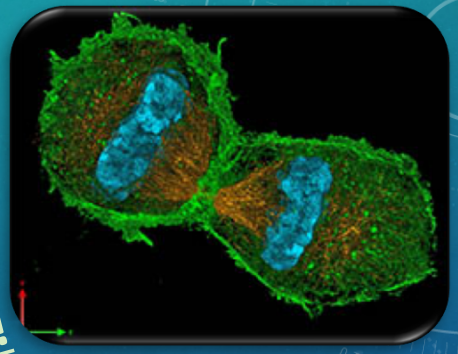


Insulin Receptor
Petrus et al, 2009, *ChemMedChem*

2. Biological Process

A commonly recognized series of events

- cell division

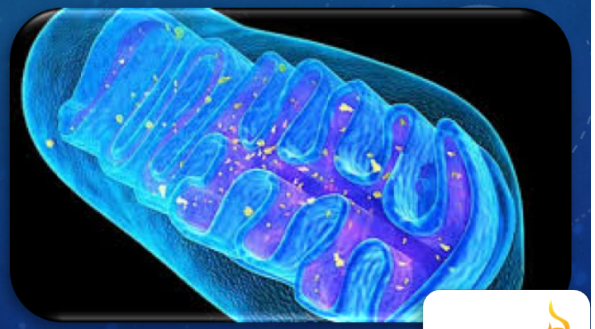


End of Telophase.
Lothar Schermelleh

3. Cellular Component

Where a gene product is located

- mitochondria
- mitochondrial matrix
- mitochondrial inner membrane



Mitochondrion.
PaisekaScience Photo Library

BERKELEY LAB

Phoenix BIOINFORMATICS

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
P02879	Ricin		protein binding		ParkinsonsUK-UCL	Ricinus communis	IPI	UniProtKB:Q8BJT9			PMID:24200403	20150609
P02879	Ricin		protein bindin					UniProtKB:Q925U4			PMID:24200403	20150609

GO:0005515 (go to the term details page for protein binding)

AmiGO 2

Anatomy of a GO term:

Unique Identifier

Term name

Ontology

Synonyms

Definition

Cross-references

Term Information

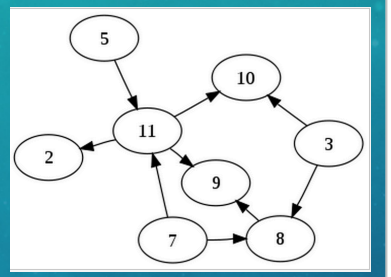
Accession GO:0000725
Name recombinational repair
Ontology biological_process
Synonyms None
Definition A DNA repair process that involves the exchange, reciprocal or nonreciprocal, of genetic material between the broken DNA molecule and a homologous region of DNA. *Source:* GOC:elh
Comment None
History See term history for GO:0000725 at QuickGO
Subset gosubset_prok
Community GN Add usage comments for this term on the GONUKB wiki.
Related Link to all **genes and gene products** as related 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.

BERKELEY LAB

Phoenix BIOINFORMATICS

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**

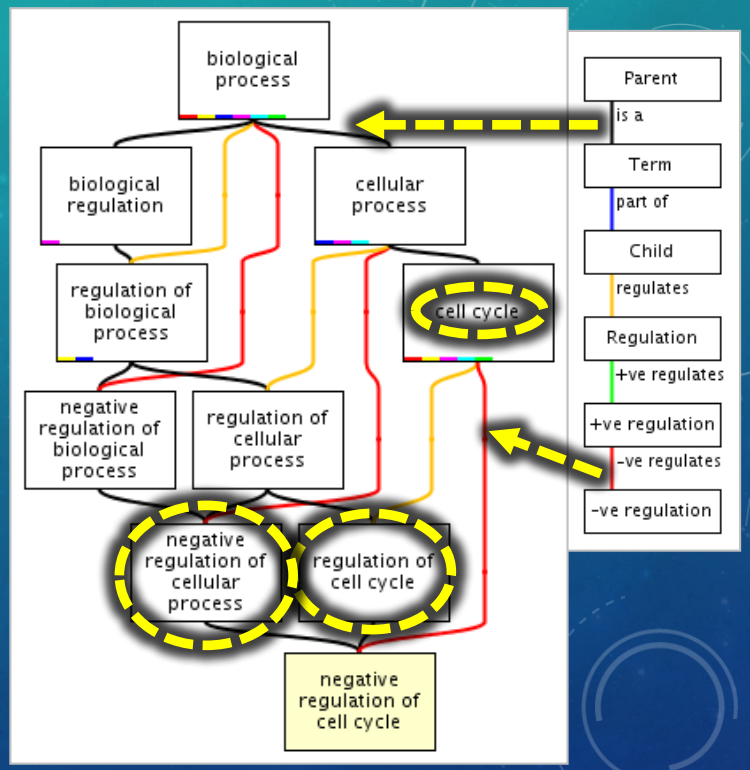
  `is_a`

  `part_of`

  `regulates` (& +/-)

  `has_part`

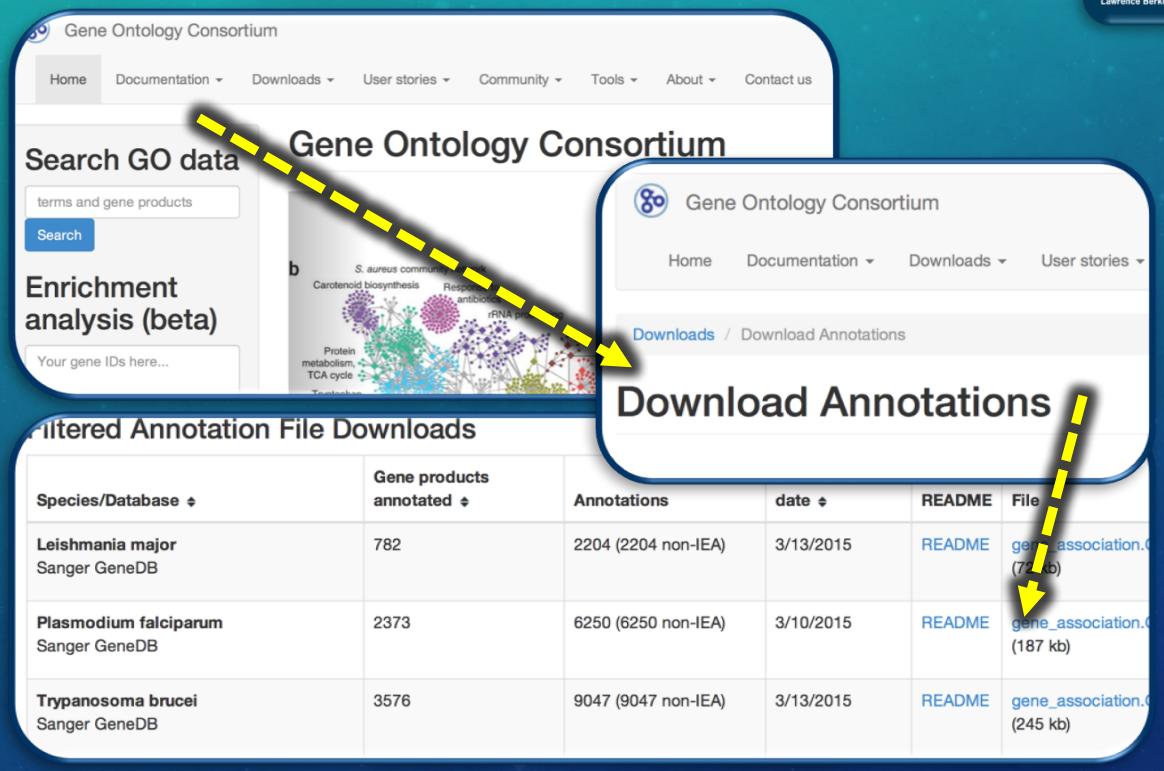
  `occurs_in`



These relationships allow for complex analysis of large datasets

Access to GO

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

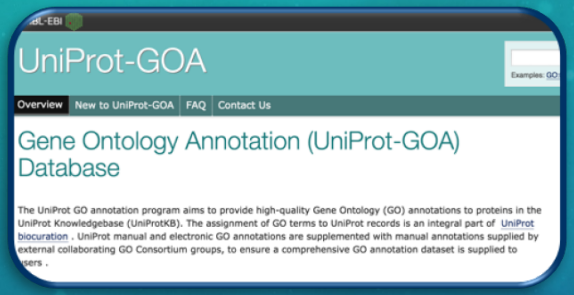


Annotation files store information about associations between a gene product and an ontology term.

There is one annotation file per species

Access to GO

2. UniProt-GOA

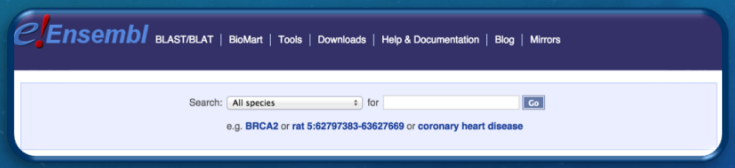


http://www.ebi.ac.uk/GOA

3. UniProtKB

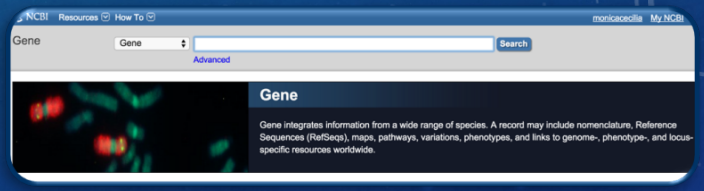
http://www.uniprot.org/uniprot/

4. Ensembl



http://www.ensembl.org/

5. NCBI Gene



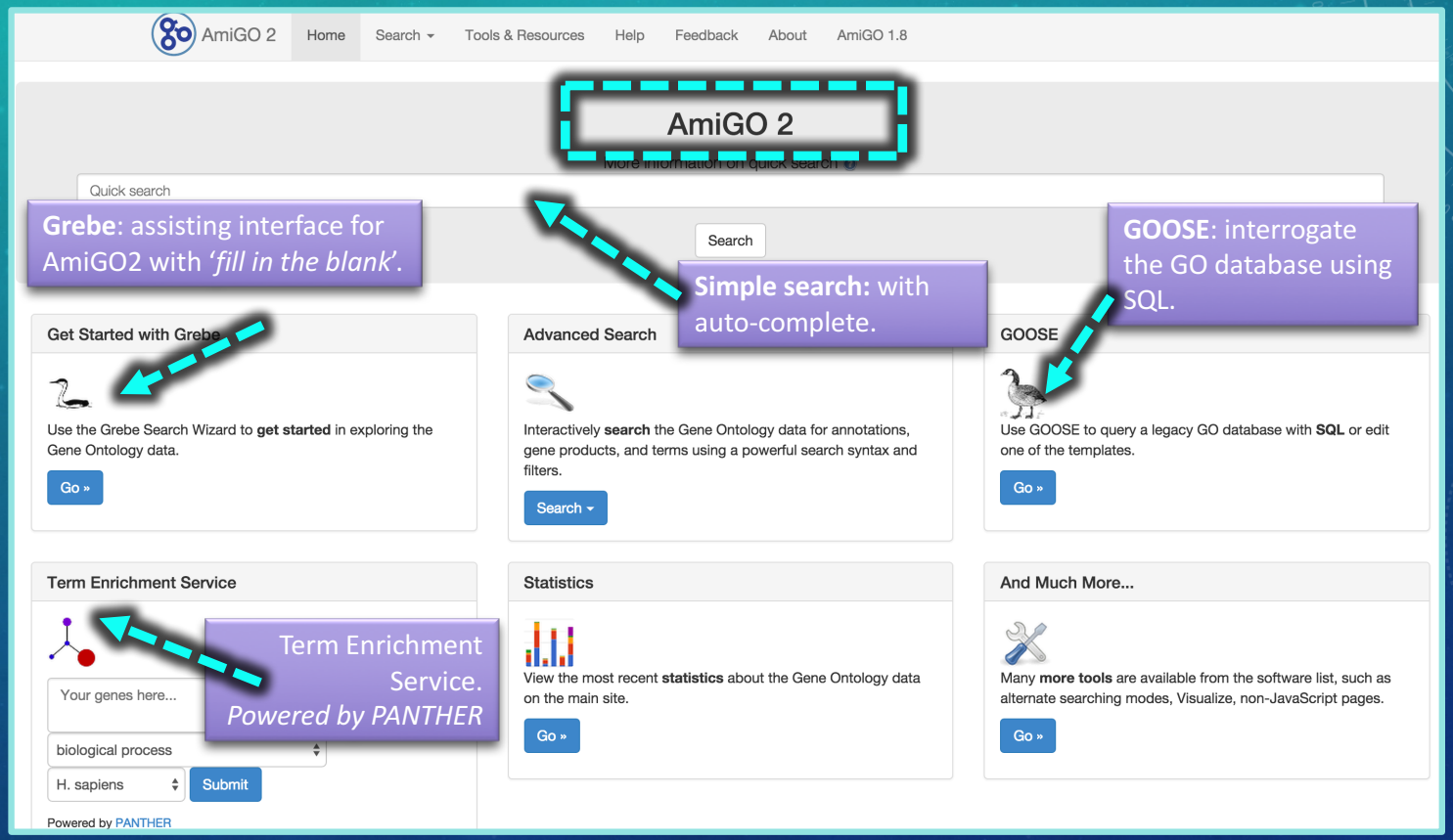
http://www.ncbi.nlm.nih.gov/gene/

BERKELEY LAB

Phoenix BIOINFORMATICS

GO Browsers

http://amigo.geneontology.org/



GO Browsers

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

EMBL-EBI | Services | Research | Training | About us

QuickGO — fast browser for Gene Ontology terms and annotations.

- QuickGO
  - Help
  - Reference
  - FAQs
  - Video tutorials
  - Downloads
  - geneontology.org
  - UniProt-GOA project
  - Web Services

EBI > Databases > QuickGO

QuickGO

Click for example search | Search! | Web Services | Dataset | Term Basket: 0

Search and Filter GO annotation sets

Extensive filters are available from this page to allow the
subsets of GO annotations, mapped to sequence identifie

Investigate GO slims

GO slims are lists of GO terms that have been selected from the full set of terms available from the Gene Ontology project.

GO slims can be used to generate a focused view of part of the GO, or with annotation data they can be used to see how a set of proteins/genes can be broadly categorized (using annot data and the relationships that exist between terms in the ontologies).

Further information on GO slims can be found at the GO Consortium web site.

View the history of changes to GO

This page allows you to view the changes to GO,
term identifier, or type of change.

Other resources for GO analysis

anges to the Term

term history displays

20 April 2011 - Display improvements

QuickGO News Archive

QuickGO Tips

Tutorial

Please send comments, suggestions or bug reports to goa@ebi.ac.uk. Click here for details of how to cite UniProt-GOA and QuickGO. 2ms

Simple search: Search GO terms or proteins; with auto-complete.

Find GO annotation sets: With filters to generate annotation subsets.

Investigate GO Slims: Focused visualization for certain GO subsets.

EMBL-EBI

BERKELEY LAB

Phoenix BIOINFORMATICS

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.

BERKELEY LAB

Phoenix BIOINFORMATICS

Functional Annotation

Assembly:
- Genome
- Transcriptome

Predicted DNA / Protein Sequences

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.

Examples:
- InterProScan
- PANTHER Term Enrichment Tool
- JAMp

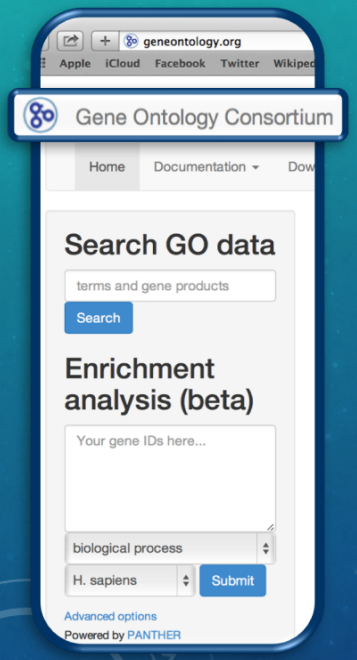
Term Enrichment

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

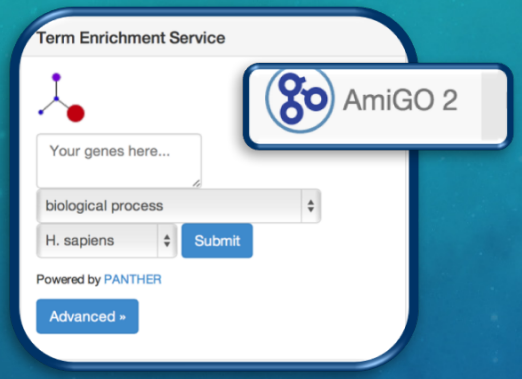
For MODs

Powered by

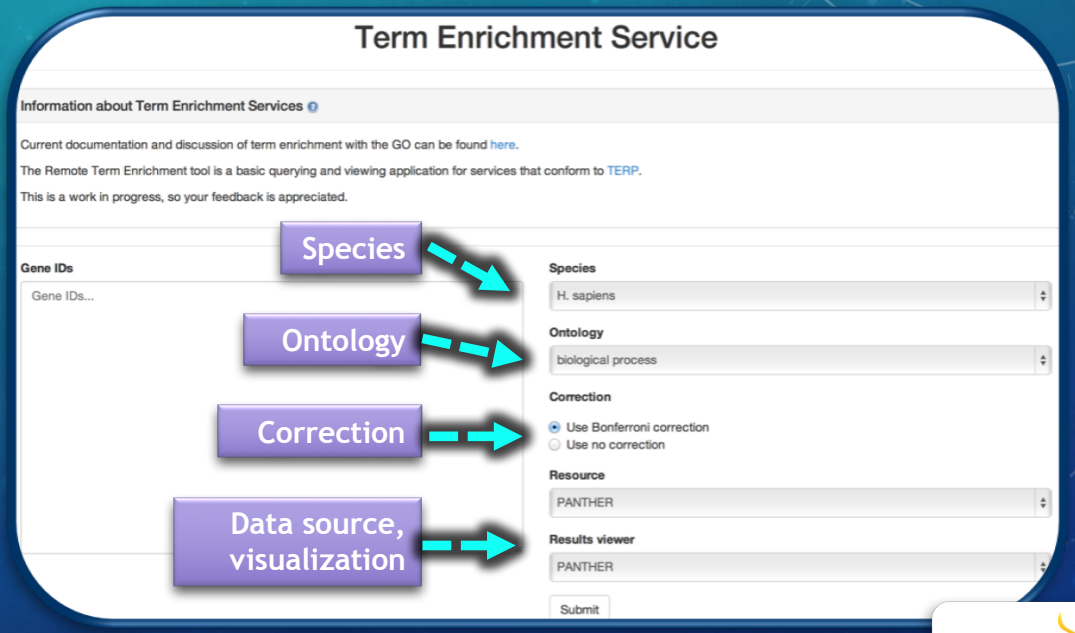
http://www.pantherdb.org/



http://GeneOntology.org/

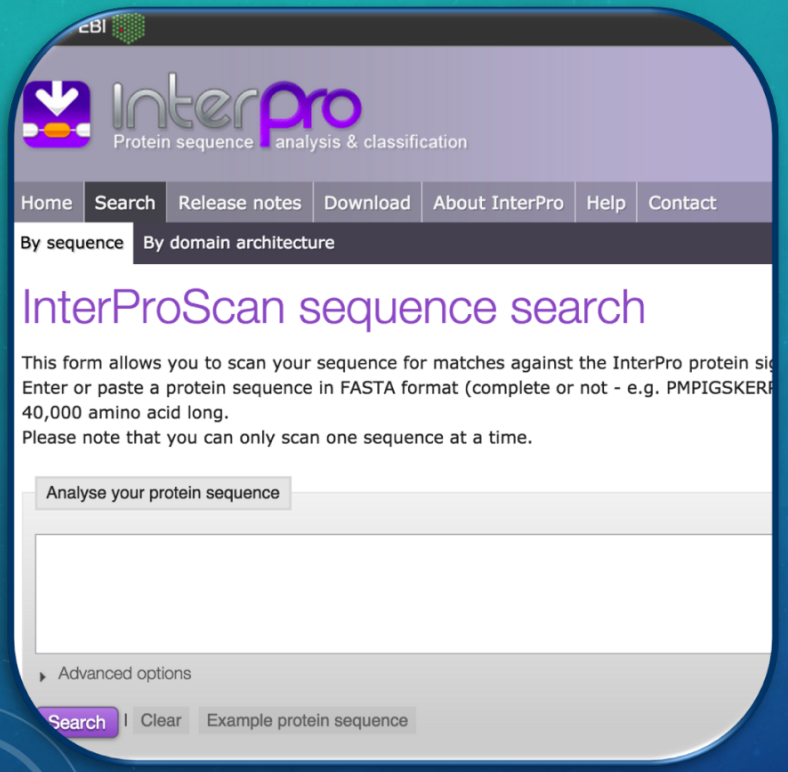


http://amigo.geneontology.org/



“Term Enrichment”

If not working with MODs



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

- **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

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

https://arabidopsis.org

tair

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 | Al

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

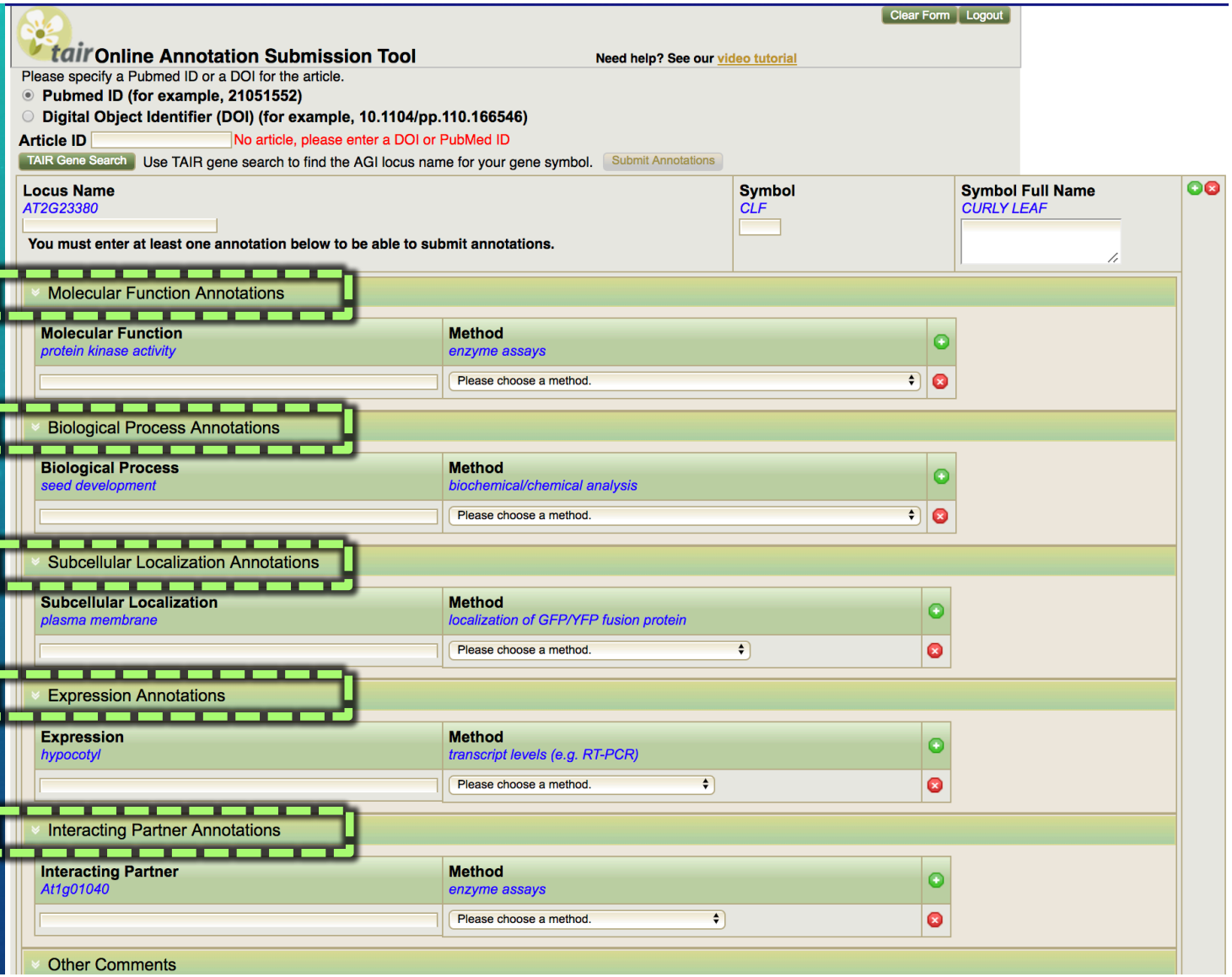
Choose Organism
- ✓ Arabidopsis thaliana
- False Brome (Brachypodium distachyon)
- Chlamydomonas reinhardtii
- Soybean (Glycine max)
- Rice (Oryza sativa)
- Moss (Physcomitrella patens)
- Poplar (Populus trichocarpa)
- Tomato (Solanum lycopersicum)
- Sorghum (Sorghum bicolor)
- Grape (Vitis vinifera)

BERKELEY LAB

Phoenix BIOINFORMATICS

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

https://arabidopsis.org



Contributing to GO

Ad hoc curation efforts

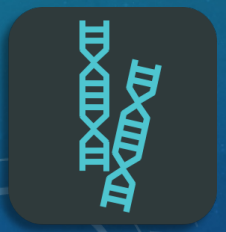
A handful of annotations from manual efforts.

Annotations

Large sets of annotations

Ontology

terms, relationships, etc.



genes, proteins

articles

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


BERKELEY LAB
Lawrence Berkeley National Laboratory
SIB
GO
Annotating cardiovascular genes
PomBase
PANTHER
Classification System
EMBL-EBI
UniProt
dictyBase
EnsemblGenomes
MGI
tair
sanger
GeneDB
WormBase
SGD
GENEONTOLOGY
Unifying Biology
FlyBase
InterPro
Protein sequence analysis & classification
Ensembl
GRAMENE
PORT Eco
MTBBASE
Reactome
PIR
J. Craig Venter
INSTITUTE
AspGD
AgBase
ZFIN
PAMGO
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE
INSTITUTE FOR GENOME SCIENCES
RGD

Noctua

Collaboratively curating gene functions

GO Annotation

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

Total annotations: 33; showing: 1-25

Results count 25

«First <Prev Next> Last» Download (up to 100000)

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference
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
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
HOXA5	Homeobox protein Hox-A5		respiratory system process		Ensembl	Homo sapiens	IEA	UniProtKB:P09021 ensembl:ENSMUSP00000039012	family not named pthr24326		GO_REF:0000107
HOXA5	Homeobox protein Hox-A5		DNA binding		UniProt	Homo sapiens	IDA		family not named pthr24326		PMID:8657138
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
HOXA5	Homeobox protein Hox-A5		transcription factor activity, sequence-specific DNA binding		UniProt	Homo sapiens	IDA		family not named pthr24326		PMID:16756717
HOXA5	Homeobox protein Hox-A5		protein binding		UniProt	Homo sapiens	IPI	UniProtKB:Q15672	family not named pthr24326		PMID:15545268
HOXA5	Homeobox protein Hox-A5		nucleus		UniProt	Homo sapiens	IDA		family not named pthr24326		PMID:15545268
HOXA5	Homeobox protein Hox-A5		transcription from RNA polymerase II promoter		GOC	Homo sapiens	IEA	GO:0001077	family not named pthr24326		GO_REF:0000108
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 10

«First <Prev Next> Last» Download (up to 100000)

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference
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
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
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
HOXD13	Homeobox protein Hox-D13		skeletal system development		Ensembl	Homo sapiens	IEA	UniProtKB:P70217 ensembl:ENSMUSP00000001872	family not named pthr24326		GO_REF:0000107
HOXD13	Homeobox protein Hox-D13		DNA binding		UniProt	Homo sapiens	IDA		family not named pthr24326		PMID:26581570
HOXD13	Homeobox protein Hox-D13		chromatin binding		Ensembl	Homo sapiens	IEA	UniProtKB:P70217 ensembl:ENSMUSP00000001872	family not named pthr24326		GO_REF:0000107
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
HOXD13	Homeobox protein Hox-D13		nucleus		HPA	Homo sapiens	IDA		family not named pthr24326		GO_REF:0000052
HOXD13	Homeobox protein Hox-D13		regulation of transcription, DNA-templated		PINC	Homo sapiens	TAS		family not named pthr24326		PMID:9207113
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

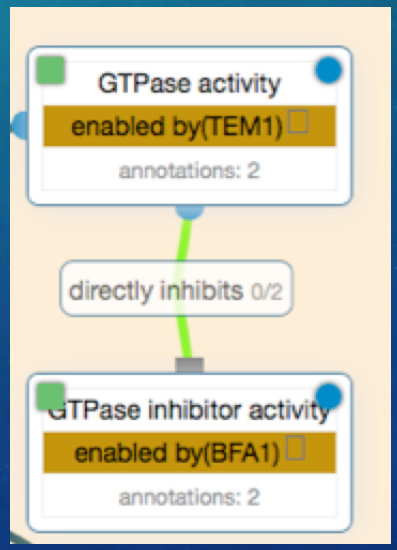
- Classic GO annotations don’t describe how genes work together
- Every annotation is independent
- Limited ability to use other ontologies:
  - E.g. PO for plant anatomy and cell types
  - TO for traits

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>**

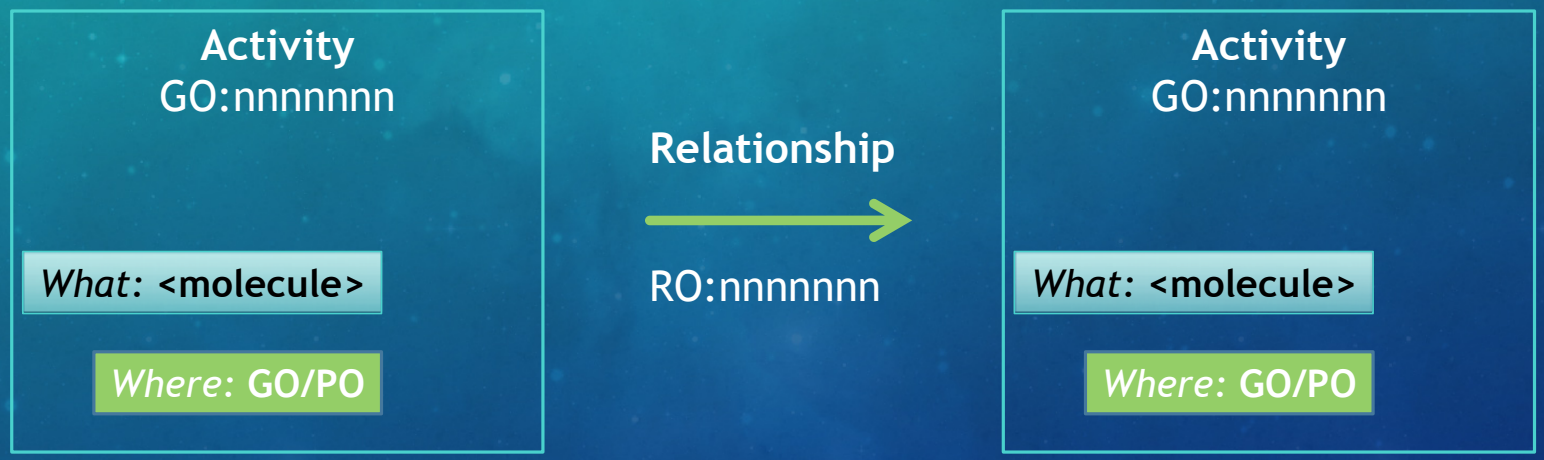
A data model for causal ontology annotations: "LEGO"

Activity
GO:nnnnnnn

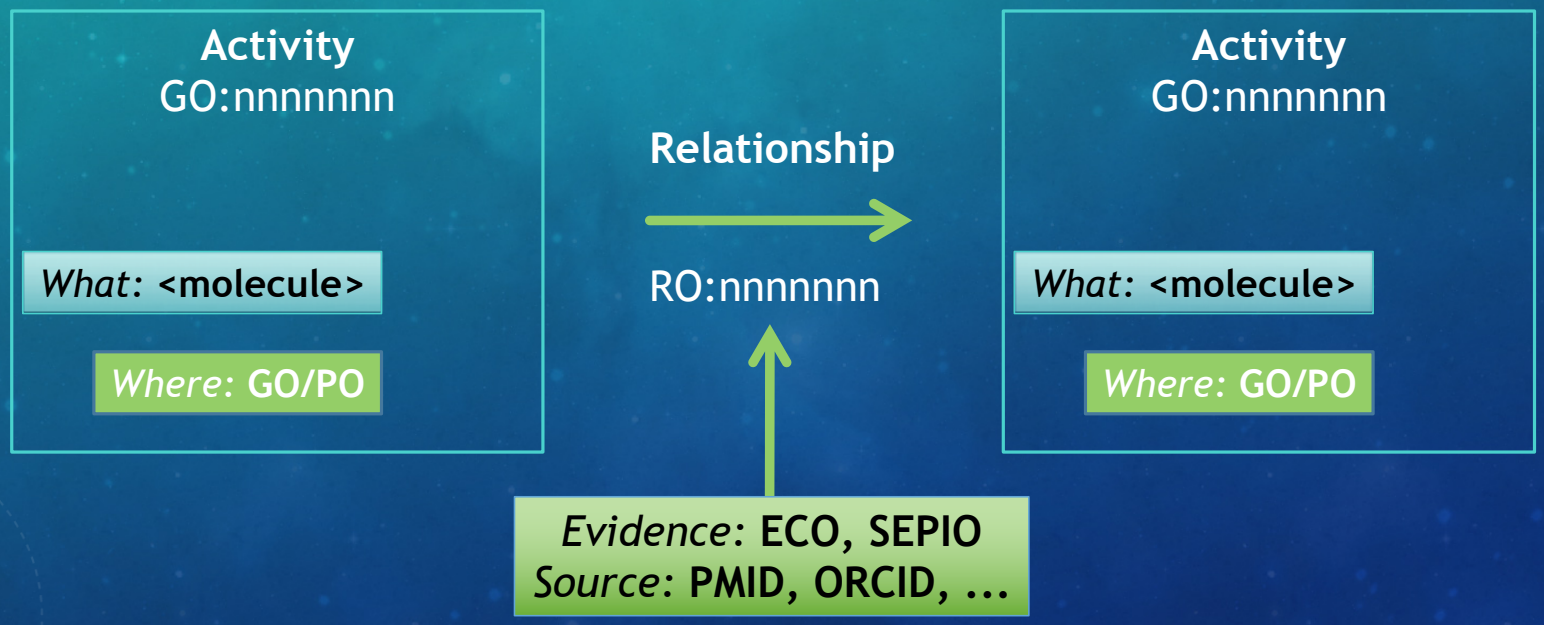
What: **<molecule>**

Where: **GO/PO**

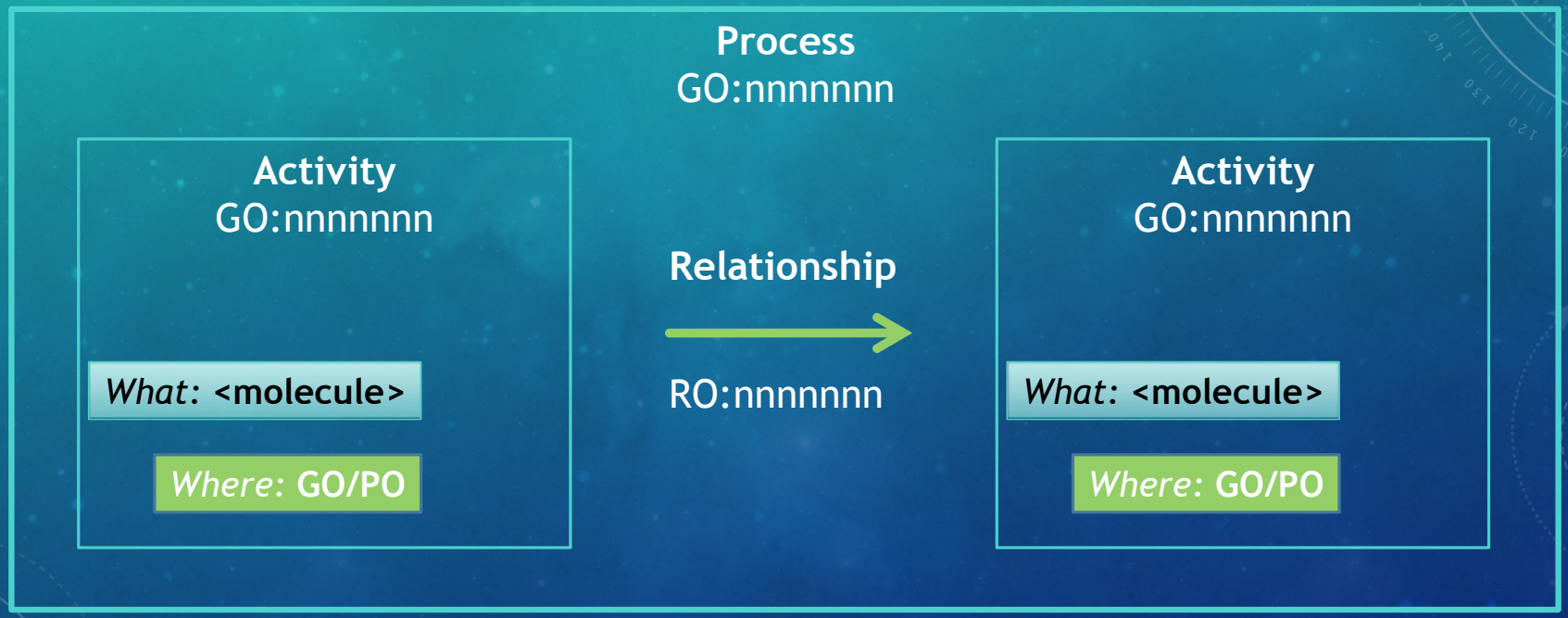
A data model for causal ontology annotations: "LEGO"



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

⊣

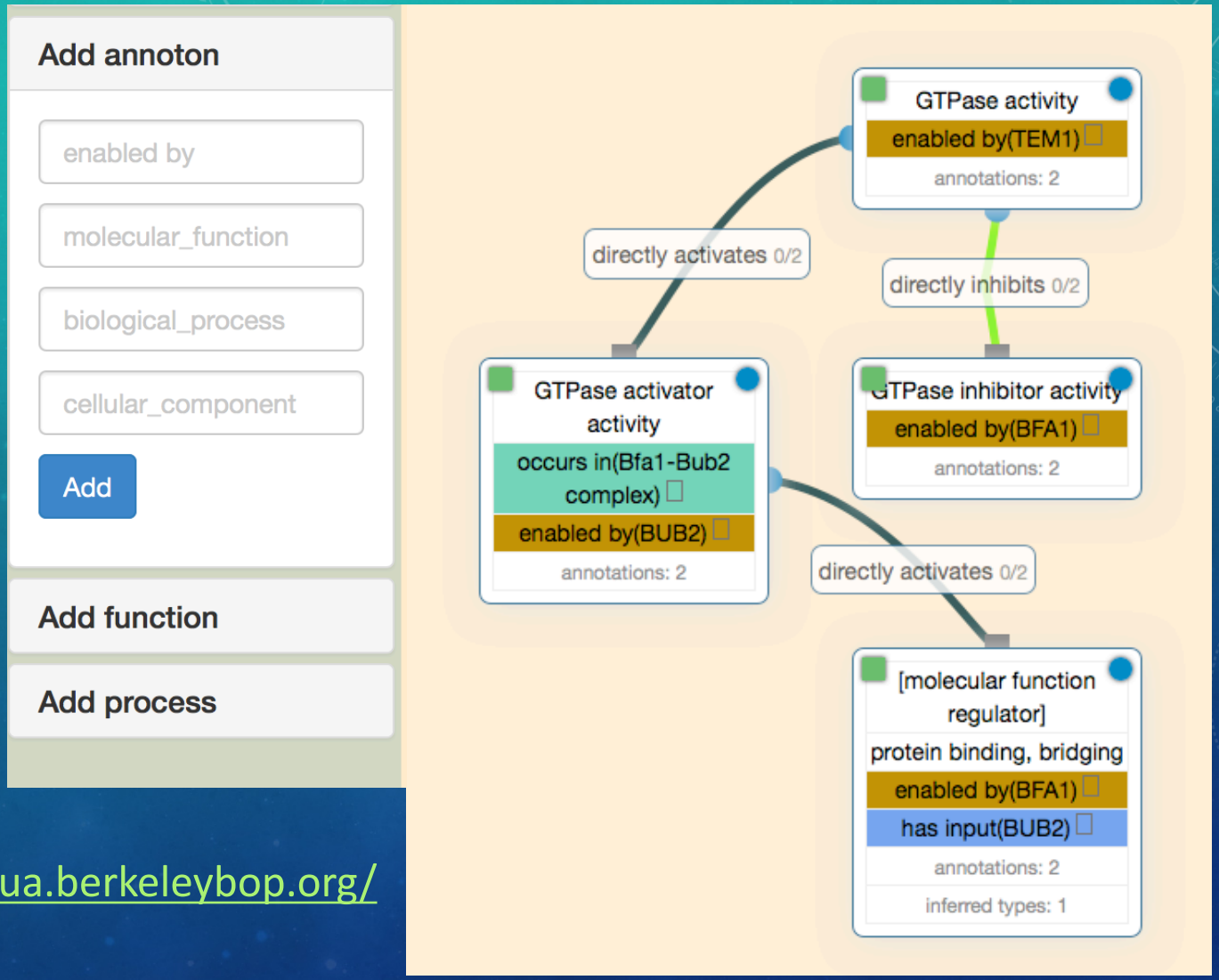
GTPase activity
GO:0003924

What: **TEM1** S000004529

Where: **spindle pole**
GO:0000922

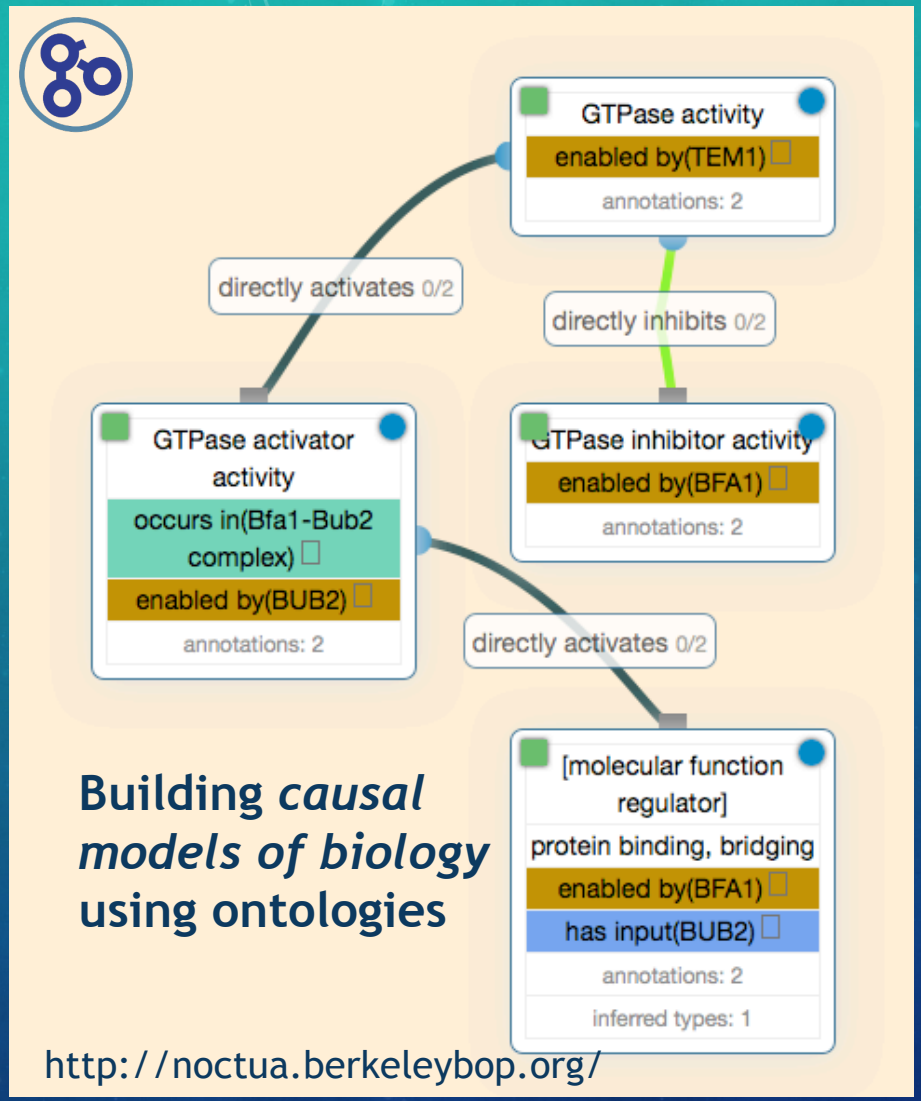
BERKELEY LAB

Phoenix BIOINFORMATICS



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

Pathway data

REACTOME

Gene sets

Phoenix BIOINFORMATICS

BERKELEY LAB

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) <u>now</u>:**
  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 Elsik**, **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

Once upon a time…