July 11-13, 2008

National Evolutionary Synthesis Center

(NESCent)

Durham, North Carolina, USA

http://gmod.org/GMOD_Summer_School

Applications due April 15!

July 16-17, 2008

University of Toronto

(before ISMB)

Toronto, Ontario, Canada

http://gmod.org/July_2008_GMOD_Meeting

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

Chado

A Database Schema for Integrating Biological Data

Arthropod Genomics Symposium

April 11, 2008

Scott Cain, PhD.
GMOD Project Coordinator
Cold Spring Harbor Laboratory
cain@cshl.edu

Dave Clements
GMOD Help Desk
National Evolutionary Synthesis Center
clements@nescent.org

Agenda

Overview & Goals

Getting data in/out, interoperability

Examples of what can be represented

Resources & Wrapup

GMOD

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

This Talk

~~Computing folks~~ Biologists

Goal

Give you an understanding

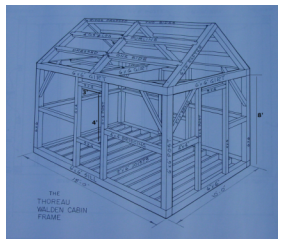
of what Chado can do

GMOD

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

Chado is a database schema for biological data

- A *schema* is a database design
  - Blueprint for a database, a way of organizing data
- Independent of specific data
  - Chado provides structure
  - You provide the hard work and data



+

=

Chado is a core GMOD component

GMOD

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

What is GMOD?

- Set of interoperable tools for managing, annotating, and viewing genomic data:
  - GBrowse - genome viewer
  - Apollo - genome annotation editor
  - CMap - comparative mapping viewer
  - Sybil - comparative genomics viewer
  - plus many other tools.
- Active community of users and developers
  - Mailing lists, semi-annual meetings

http://gmod.org

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

GMOD and you?

- Lots of communities now have genomic data
- Managing, visualizing, annotating and exploiting this data is non-trivial
- GMOD provides
  - tools for this and
  - a community of people dealing with the same challenges.

GMOD

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

GMOD @ Arthropod Genomics Symposium

GMOD Users
wFleaBase
BeetleBase
FlyBase
AphidBase
ButterflyBase
VectorBase
BeeBase
HeliconiusBase
GnpAnnot-Lep
+ hundreds more

GMOD Tools
Chado
Apollo
GBrowse
CMap
Table Editor
GMODWeb
Ergatis
Textpresso
Sybil
BioMart
LuceGene
SynBrowse
Galaxy
InterMine
…

Workshops

Thursday 7-9pm
Community Contributions to Genome Annotation
Christine Elsik, Christopher Childers, Darren Hagen

Friday 7:30-9:30pm
Chado Databases and Integration with GMOD Tools
Scott Cain, Dave Clements

Talks

Friday 11:25-11:45am
Unlocated arthropod genes and ways to find them
Don Gilbert

Posters

Friday 5-6:30pm
A comparative annotation of Drosophilid dicistronic genes

Saturday 10-11:30am
A Complete System For Community Genome Annotation
GMOD: Database Resources for Emerging Model Organisms
Comparative genomics and database construction for Lepidoptera
Unlocated Arthropod genes, and ways to find them
VectorBase: A genome resource for arthropod vectors of human pathogens
The VectorBase Manual and Community Annotation Submission pipeline
ButterflyBase: A framework for comparative genomics in butterflies and moths

http://gmod.org

Got data?

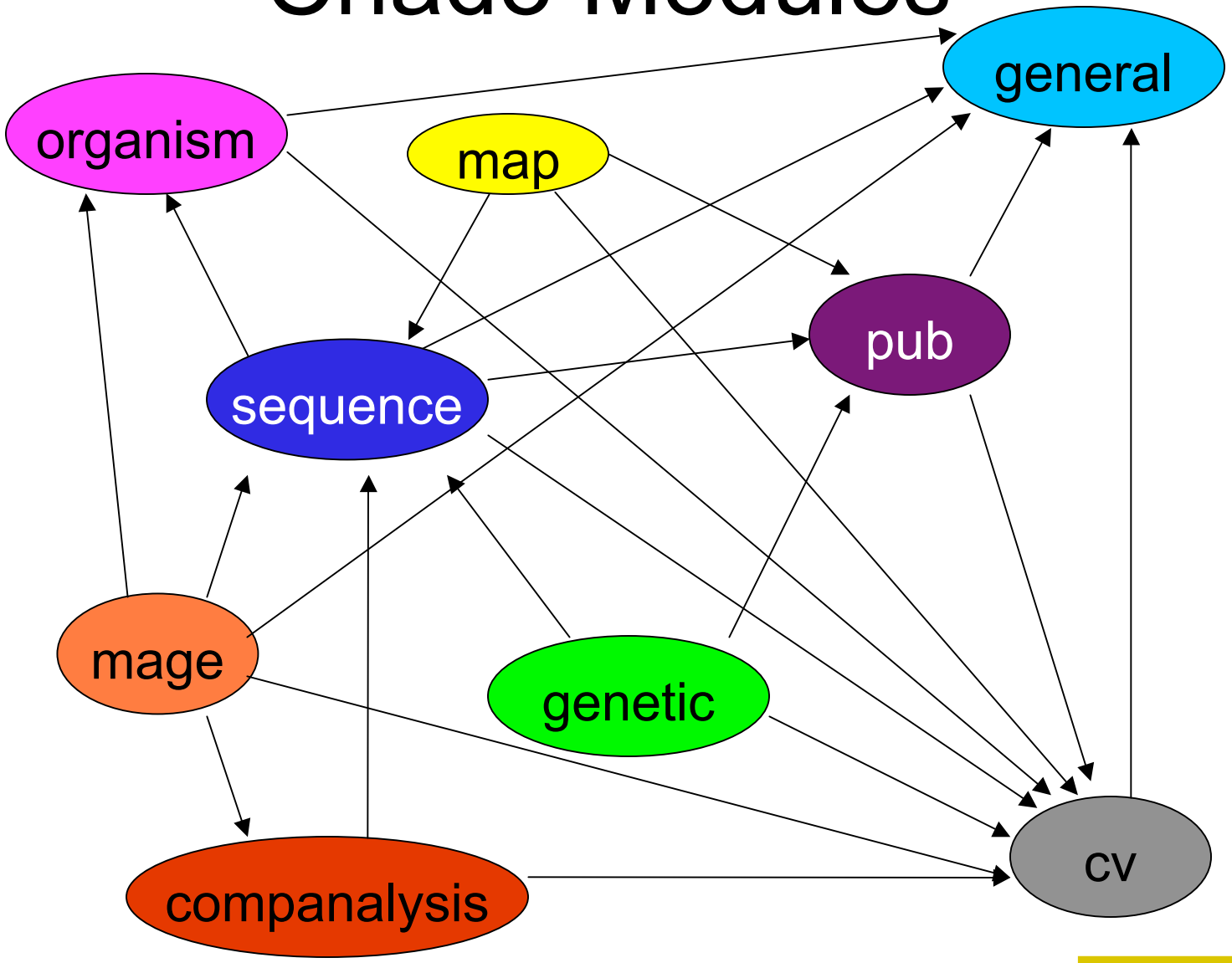
More on Databases

- Chado is a schema, a database design
- Distinct from
  - Database Management System (DBMS)
    - Software system for storing databases
    - e.g.,Oracle, PostgreSQL, MySQL
  - Database, a very loose term
    - Any set of organized data that is readable by a computer
    - A web site with database driven content, e.g., FlyBase
    - Schema + DBMS + Data

Why use Chado?

- Good for genomic data
- Widely used
  - AphidBase, BeeBase, BeetleBase, FlyBase, GnpAnnot-Lep, HeliconiusBase, VectorBase, wFleaBase, many others
- Integrates with other GMOD tools
- Great community of support
  - Mailing List (779 *threads* in 2007)
  - Plus Scott and Dave!
- Modular & Extensible

Chado Modules



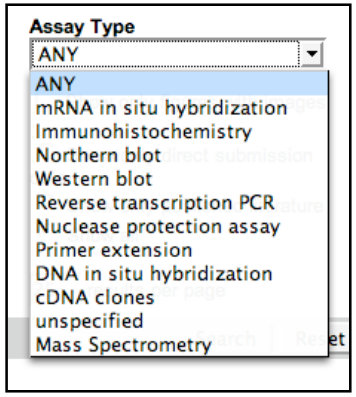
GMOD

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

Controlled Vocabulary (CV)

List of terms from which a value *must* come

Pull down menus are examples of CVs



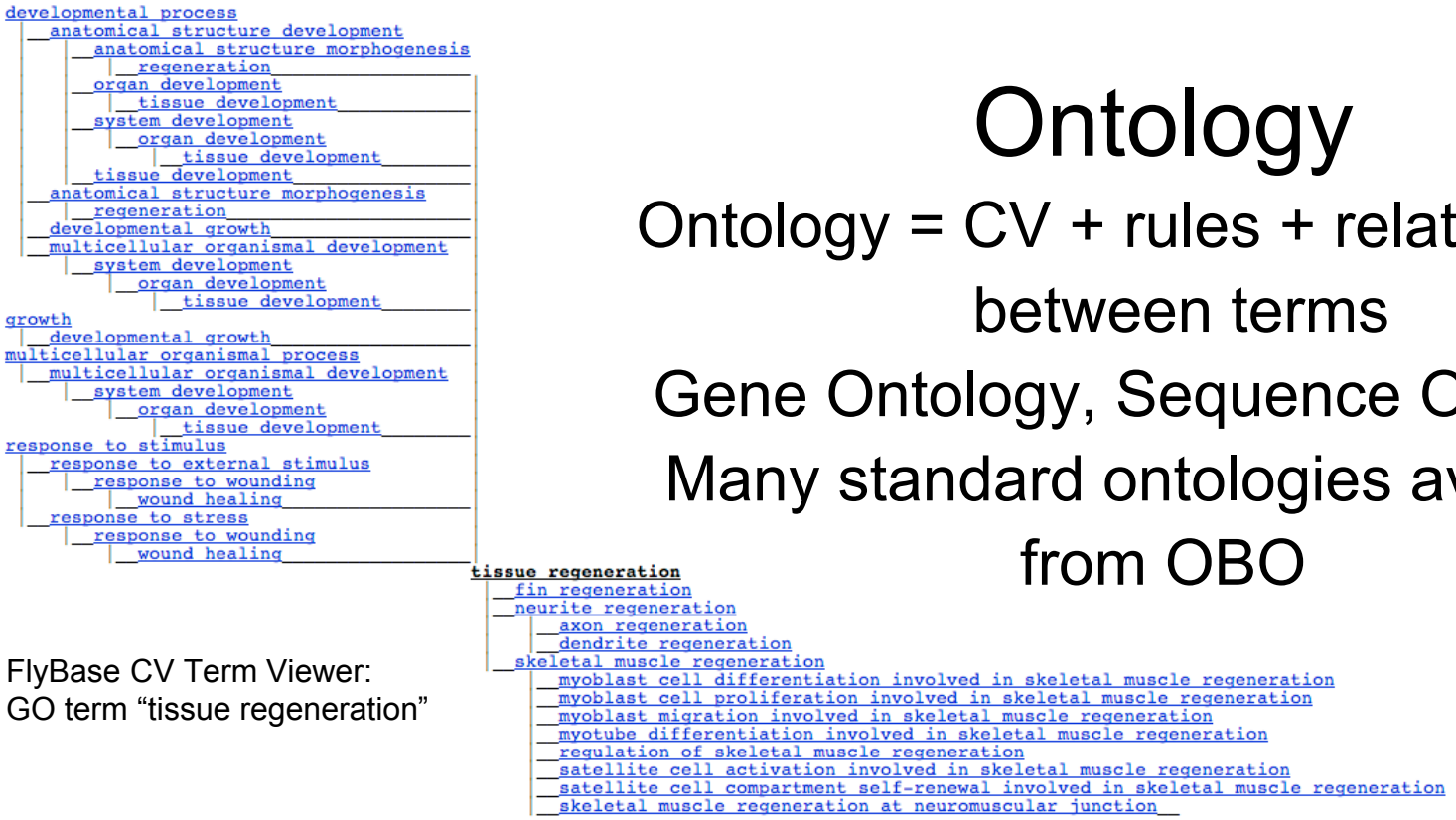
ZFIN assay type CV

Ontology

Ontology = CV + rules + relationships between terms

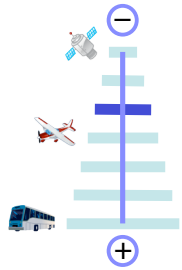
Gene Ontology, Sequence Ontology

Many standard ontologies available from OBO

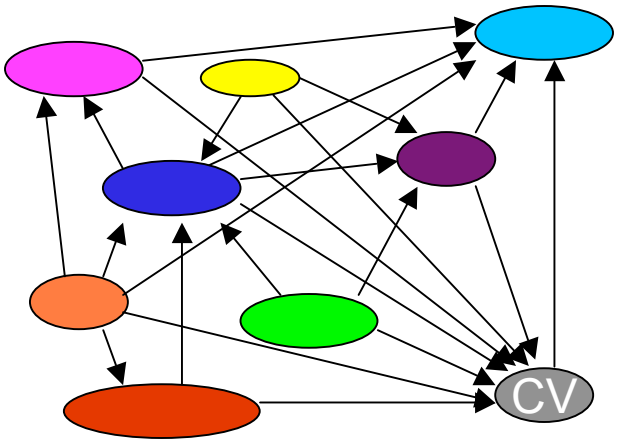


FlyBase CV Term Viewer:
GO term “tissue regeneration”

CVs and Ontologies in Chado



- Controlled vocabularies and ontologies are key in Chado
- Maximally used for
  - Integrity
  - Interoperability
- Can create your own, *but* …
  - Please use standard ontologies when they exist
    - e.g., SIBO: Social Insect Behavior Ontology (Abbas & Smith)
  - See OBO: http://www.obofoundry.org/



GMOD

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

Requirements

- Operating System (OS)
  - Linux preferred
    - Recommend CentOS; most distributions will do
  - Mac OS X, other Unix - workable
  - Windows - doable, but hard

- Database Management System (DBMS)
  - PostgreSQL preferred
  - MySQL, Oracle, Sybase, … - doable, but hard

- GMOD Systems Administrator
  - Understands Linux, relational databases

http://gmod.org/Computing_Requirements

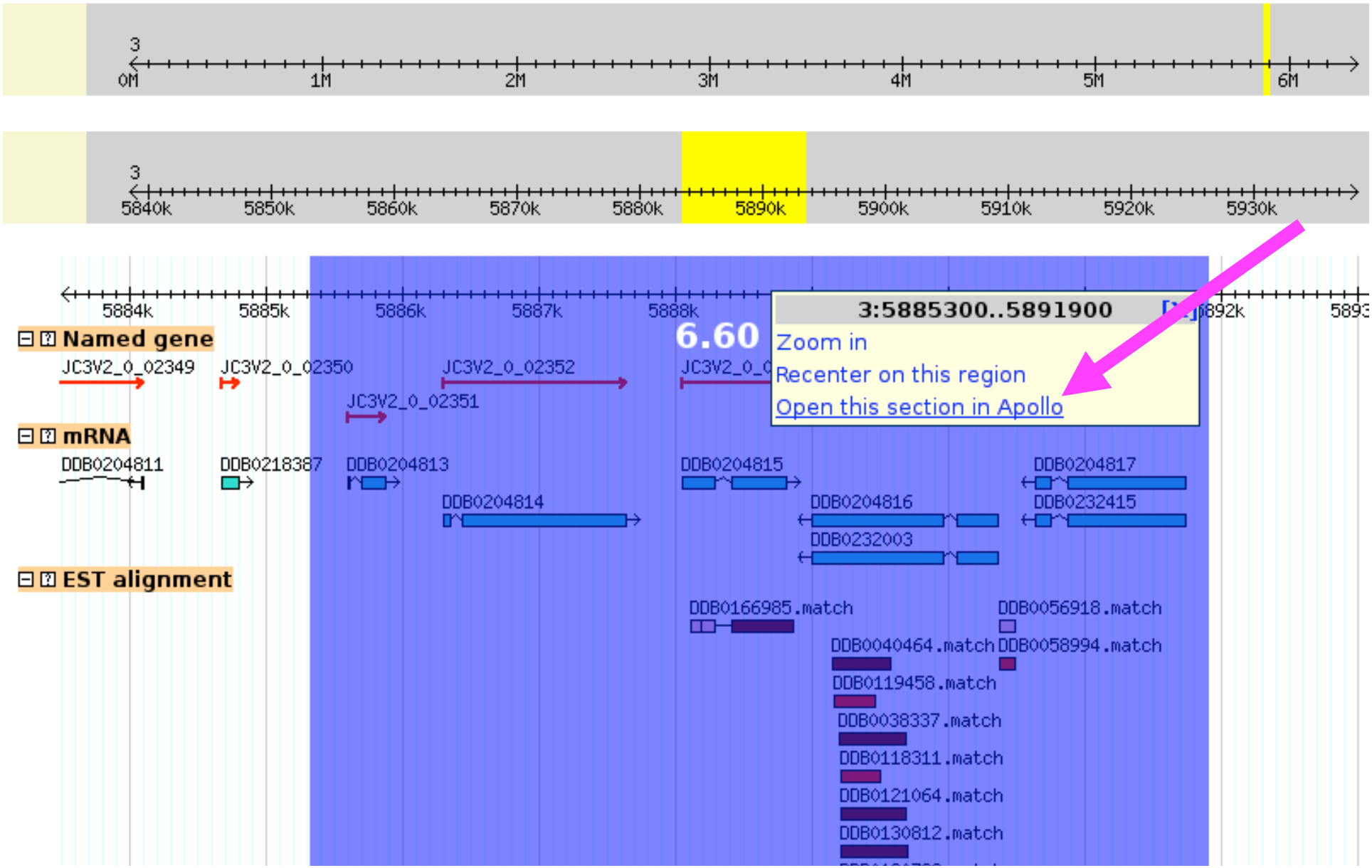
Importing and Exporting Data

- GFF3
  - dedicated loader
  - bulk file dumper
    - configurable GFF3, Fasta dumps
- Chado-XML
  - XORT
- Chaos-XML, oboxml
  - xslt

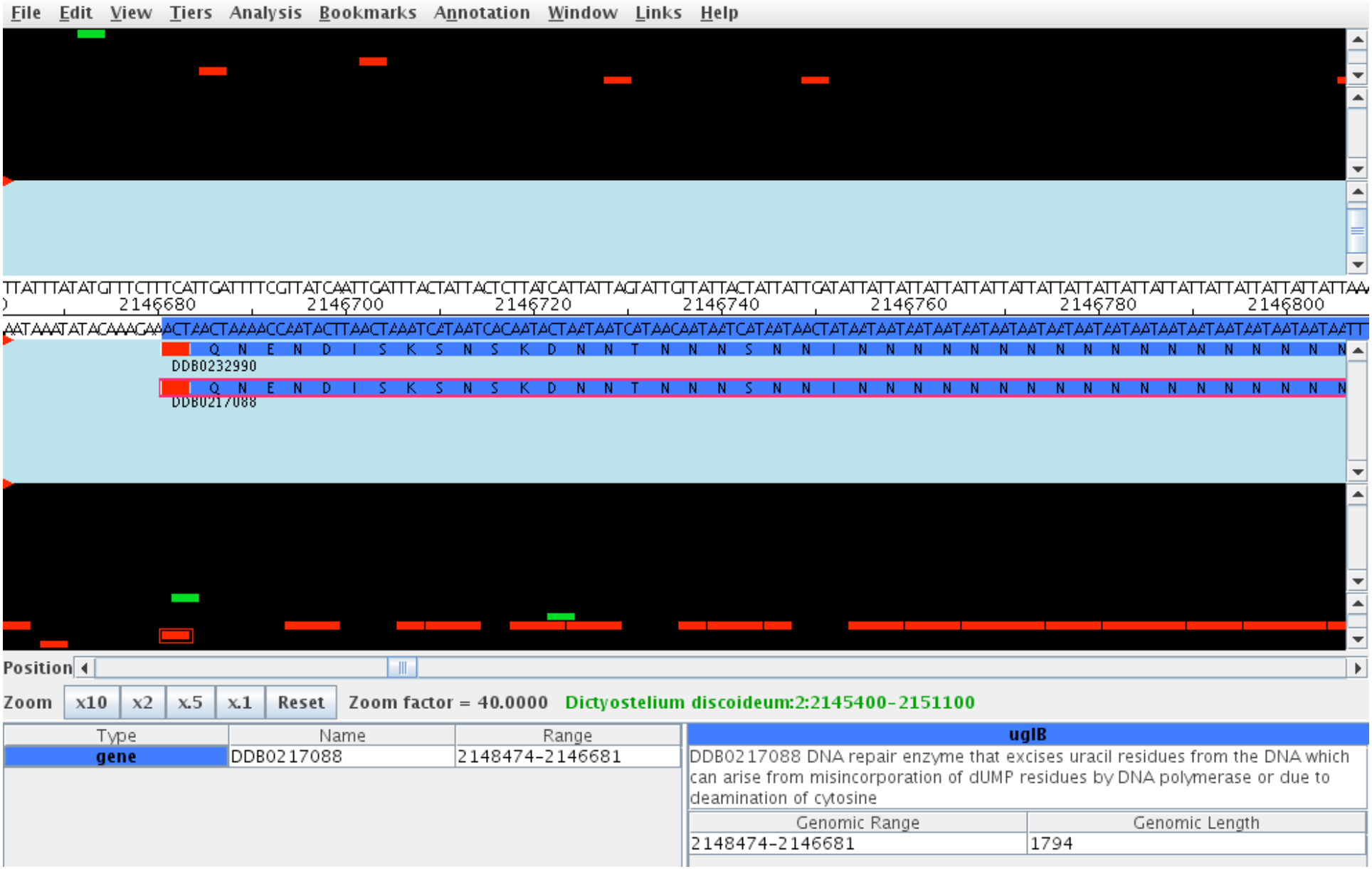
Integration with Other GMOD Tools

- GBrowse
  - Web-based feature browser
- Apollo
  - Desktop-based feature editor
- GMODWeb
  - Web front end for Chado
- CMap
  - Comparative mapping

GBrowse


3
0M
1M
2M
3M
4M
5M
6M
3
5840k
5850k
5860k
5870k
5880k
5890k
5900k
5910k
5920k
5930k
5884k
5885k
5886k
5887k
5888k
6.60
3:5885300..5891900
Zoom in
Recenter on this region
Open this section in Apollo
Named gene
JC3V2_0_02349
JC3V2_0_02350
JC3V2_0_02352
JC3V2_0_02351
mRNA
DDB0204811
DDB0218387
DDB0204813
DDB0204814
DDB0204815
DDB0204816
DDB0204817
DDB0232415
DDB0232003
EST alignment
DDB0166985.match
DDB0056918.match
DDB0040464.match
DDB0058994.match
DDB0119458.match
DDB0038337.match
DDB0118311.match
DDB0121064.match
DDB0130812.match

Apollo


File Edit View Tiers Analysis Bookmarks Annotation Window Links Help
2146680
2146700
2146720
2146740
2146760
2146780
2146800
DDB0232990
DDB0217088
Position
Zoom x10 x2 x.5 x.1 Reset Zoom factor = 40.0000 Dictyostelium discoideum:2:2145400-2151100
Type Name Range
gene DDB0217088 2148474-2146681
uglB
DDB0217088 DNA repair enzyme that excises uracil residues from the DNA which can arise from misincorporation of dUMP residues by DNA polymerase or due to deamination of cytosine
Genomic Range Genomic Length
2148474-2146681 1794

GMODWeb

ParameciumDB

Version 1.17

Search : Sequence ok

Home | Blast | Browse | Tools | Submit | Download | About

Properties

- description
  Non-discharge of trichocysts.

Phenotypes

- non-discharge of trichocysts
  - environment : standard
  - recessive

Stocks

- nd242 F2 1a
- 7d4-NO 131 nd242 Cur
- d4-NE127 nd242 Cur
- nd242 F2 19a
- nd242 F2 19c
- nd242 F2 1b

Basic Information :

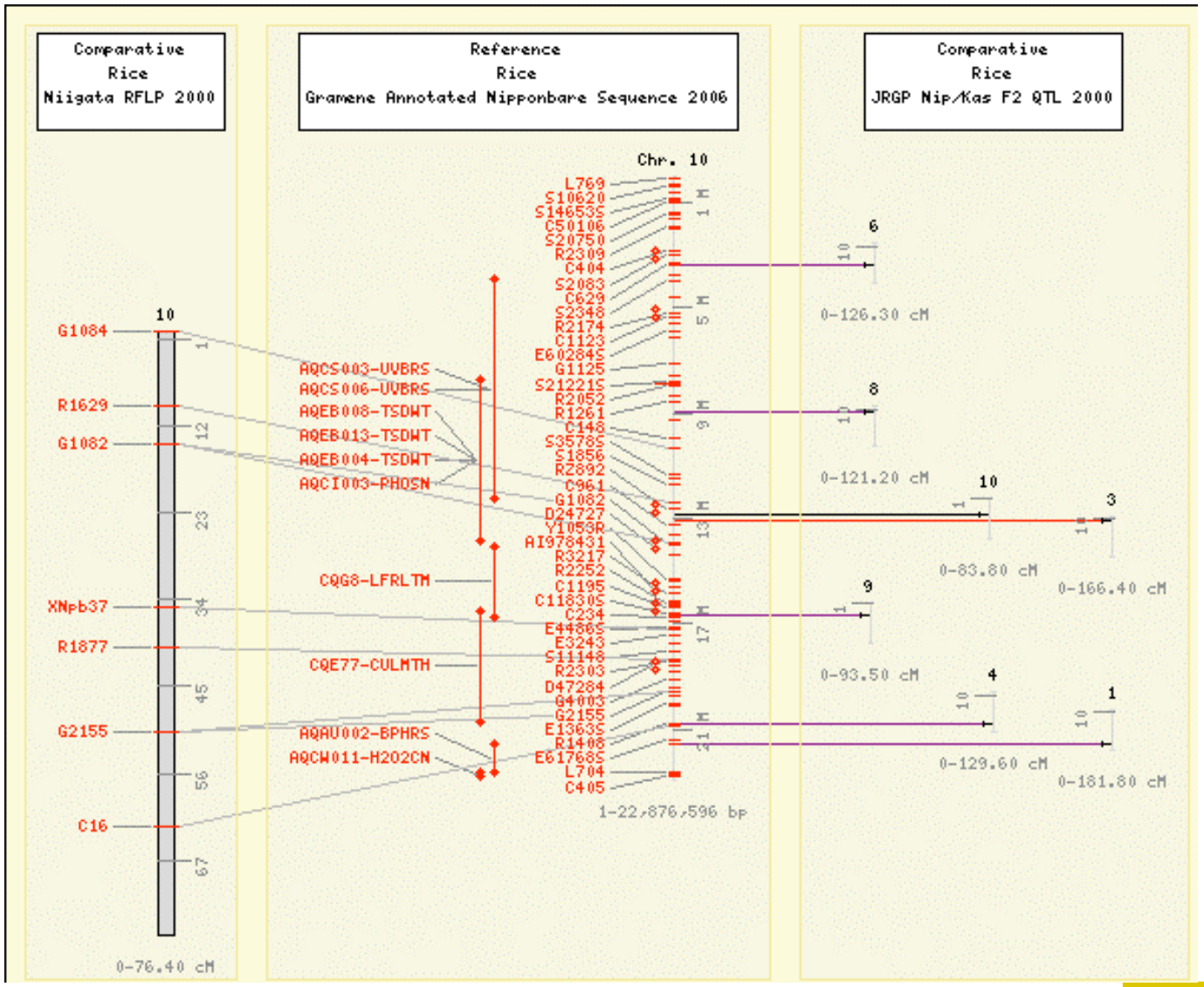
Name :	nd126-2
Type :	allele
Synonyms :	nd126b, nd242,
Publications :	NCBI link
Heredity :	micronuclear heredity

No sequence available

Contact: Olivier Arnaiz , Linda Sperling

GMOD CNRS CGM

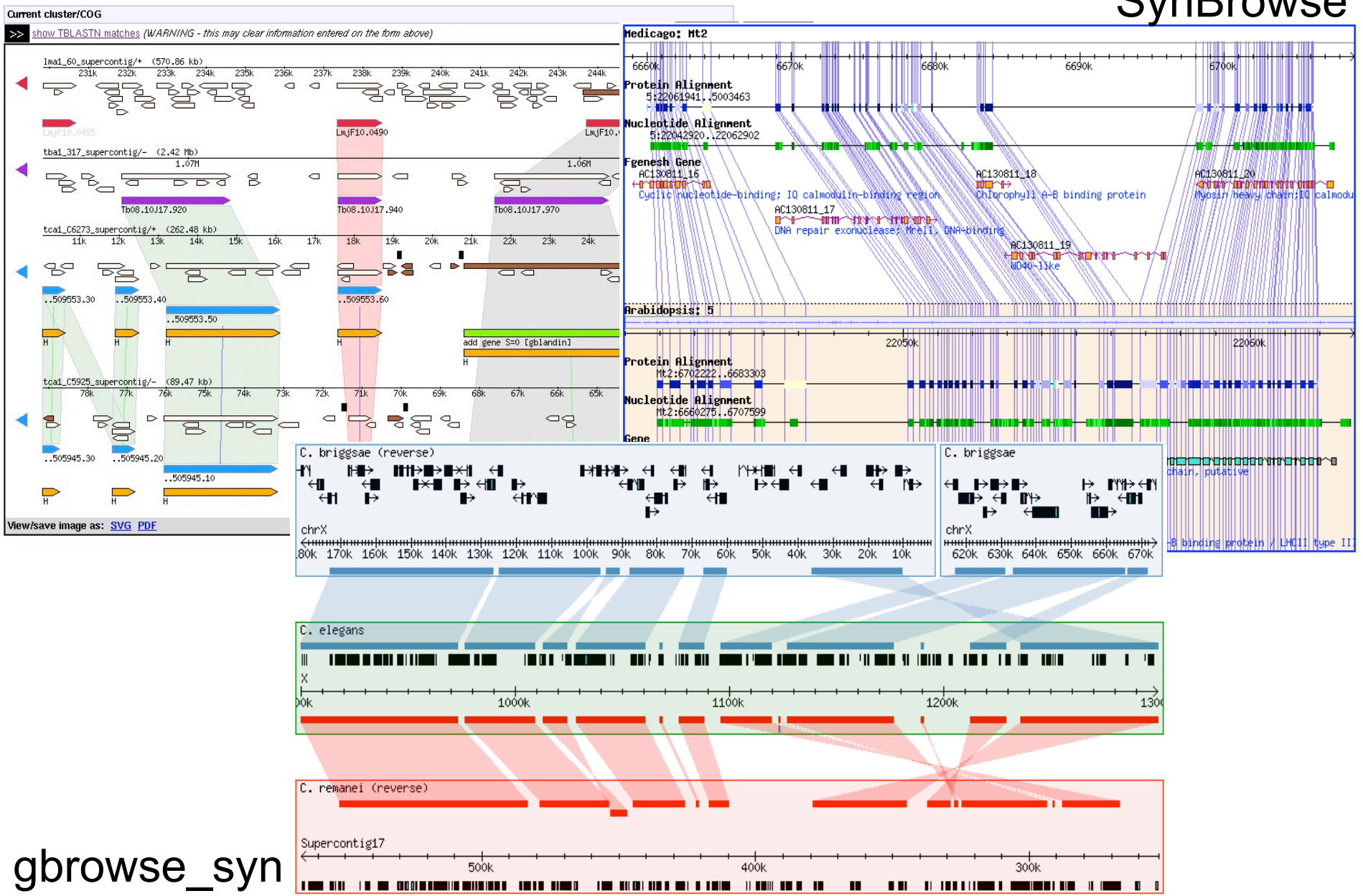
CMap



Synteny

Sybil

SynBrowse



gbrowse_syn

Sequence Module

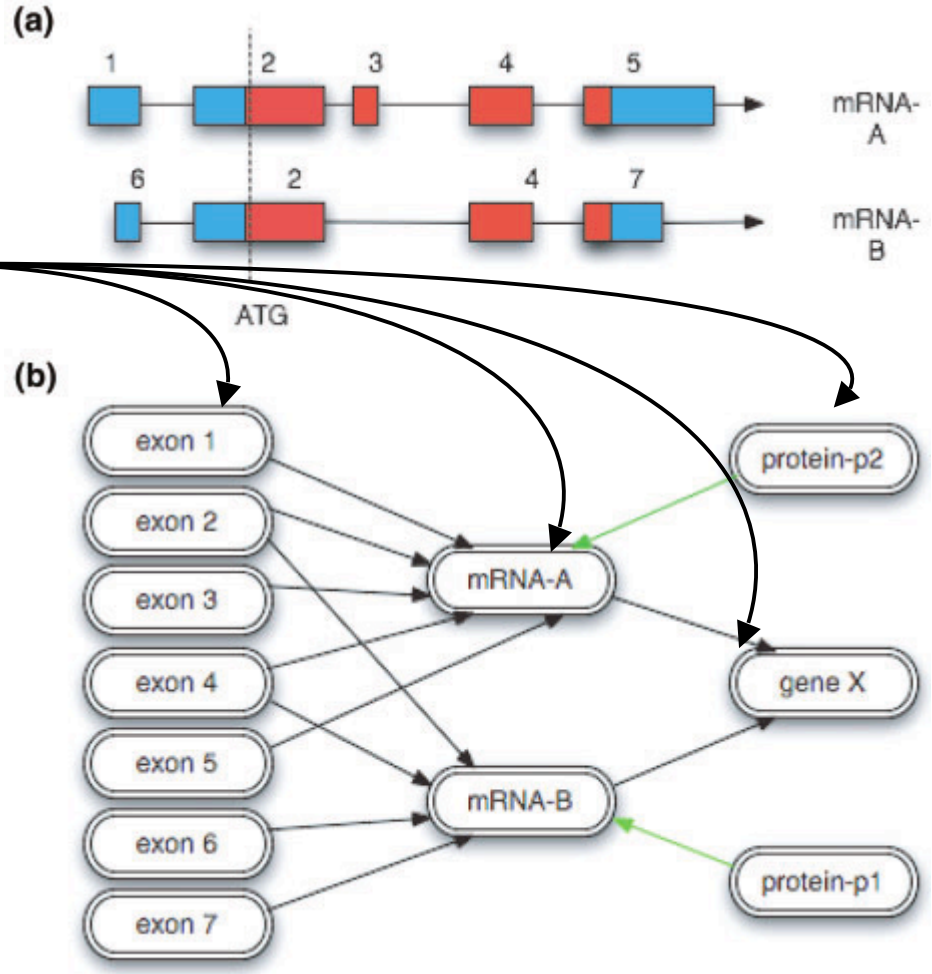
- All features (chromosome, gene, exon, etc) go into one table
- Typed using the Sequence Ontology
- Linking tables for annotating db xrefs, ontology terms, publications, properties, relationships between features
- Separate tables for synonyms (which are also typed)

GMOD

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

Storing Sequence Features

features

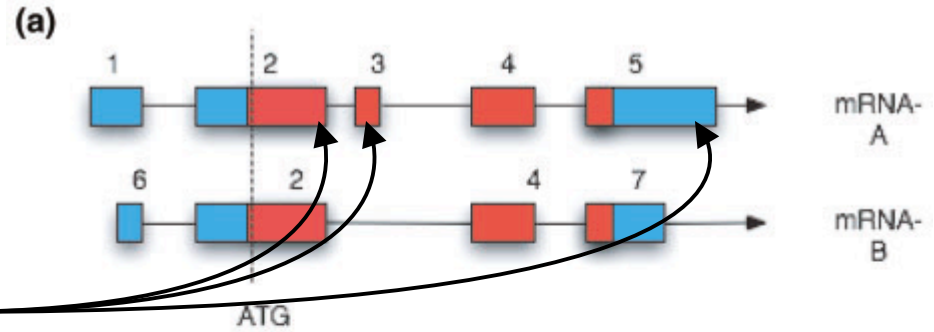


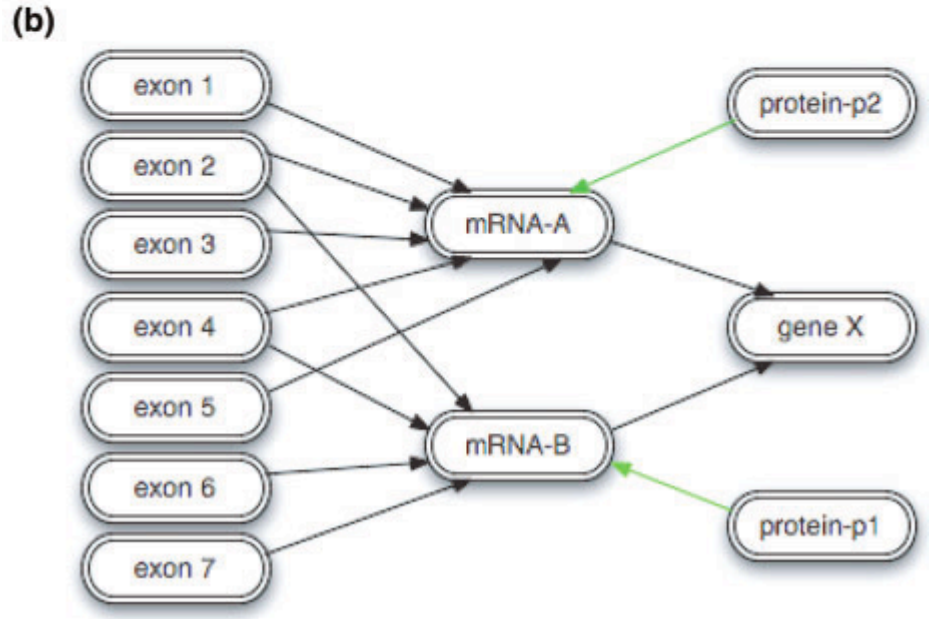
Figures based on *A Chado case study: an ontology-based modular schema for representing genome-associated biological information,* by Christopher J. Mungall, David B. Emmert, and the FlyBase Consortium (2007)

SECOND ANNUAL Arthropod Genomics SYMPOSIUM

Feature coordinates

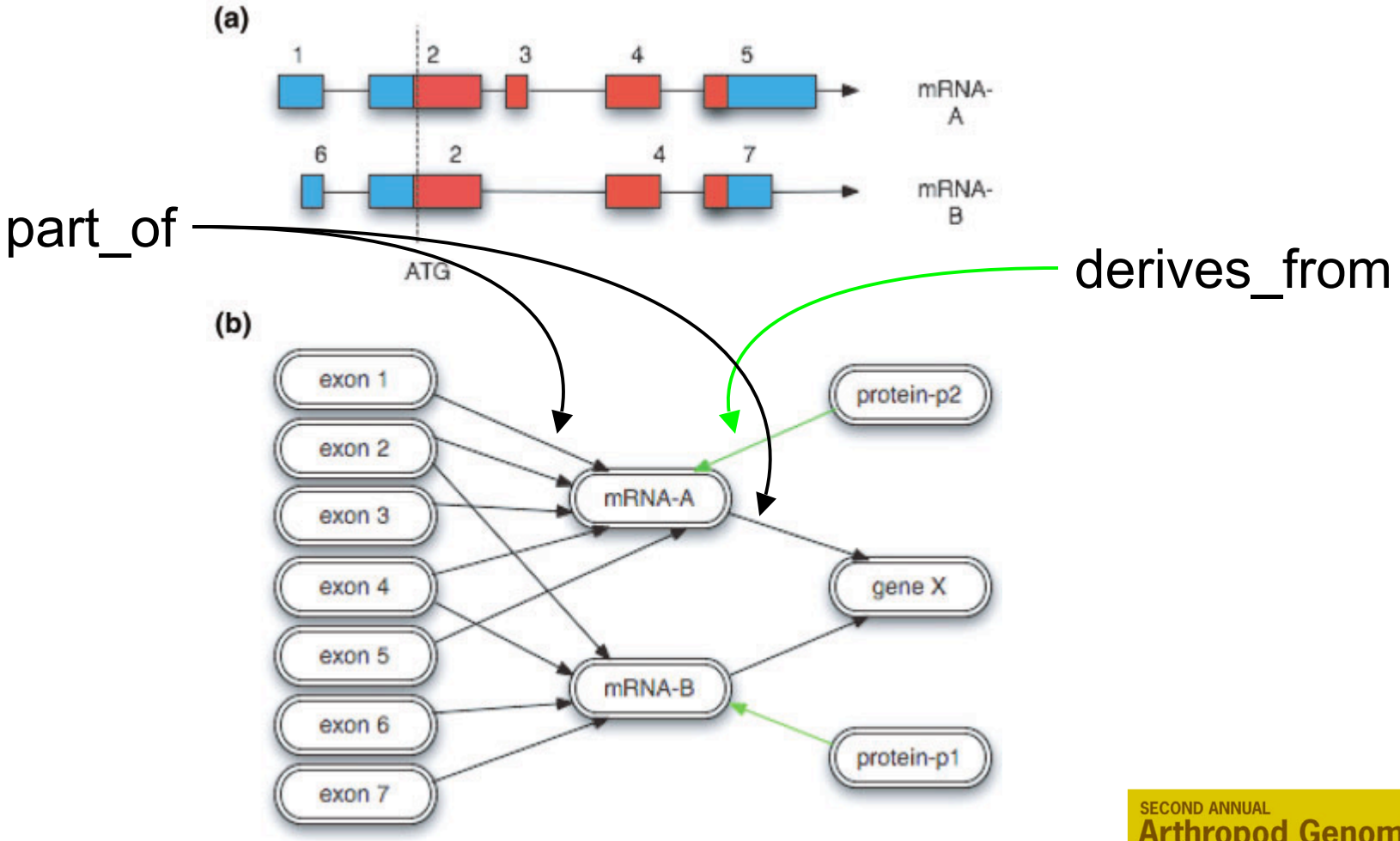
featurelocs





using "interbase coordinates" relative to any other feature (contig, chromosome, transcript)

feature_relationship



GMOD

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

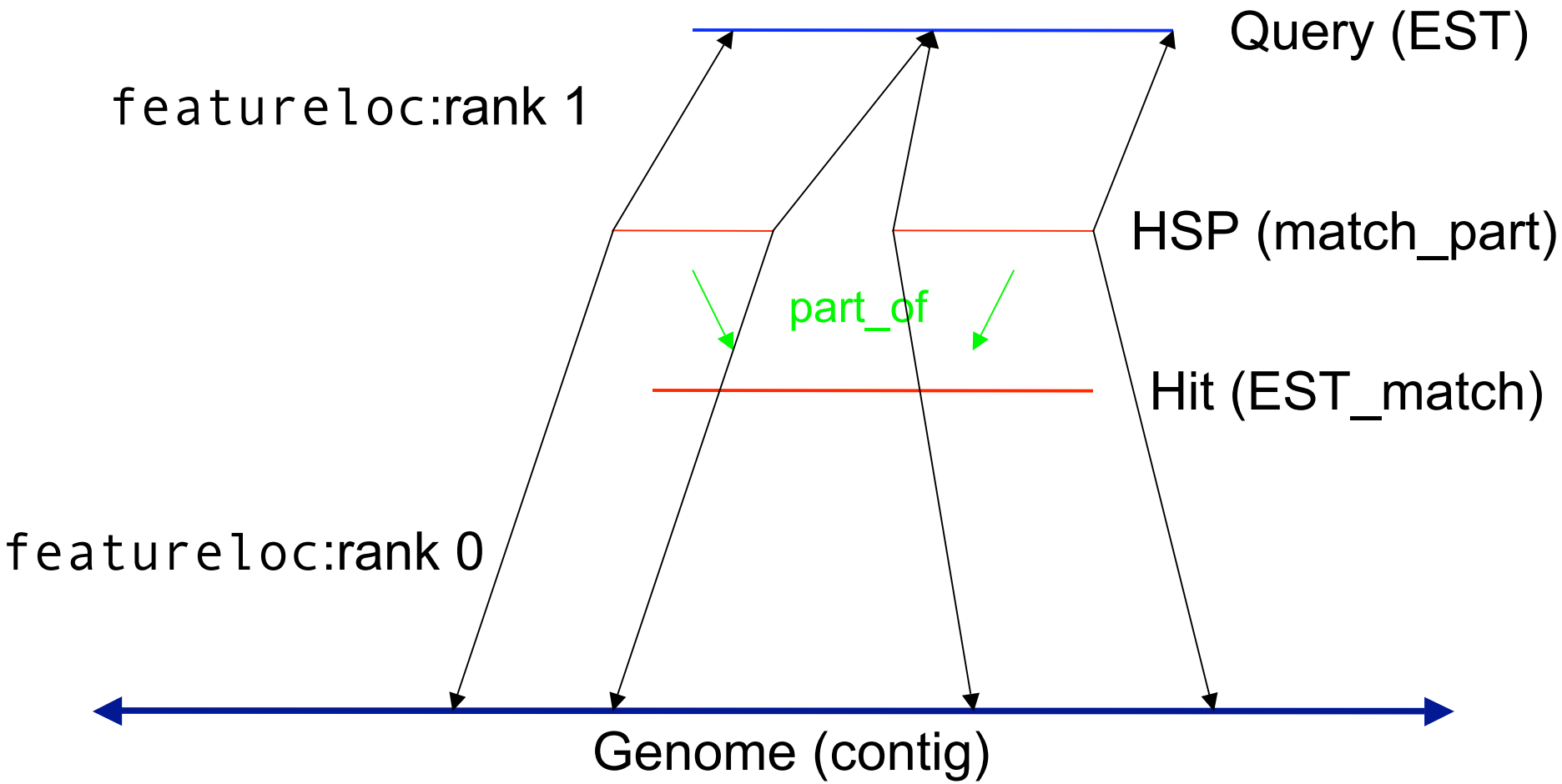
Other annotations

- Link to any feature via `feature_id`:
  - GO terms in `feature_cvterm`
  - DB links in `feature_dbxref`
  - Miscellaneous features in `featureprop`
  - References in `feature_pub`

Storing BLAST/BLAT results

- A little more complicated, since there are two reference sequences (which means two entries in `featureloc`).
- `analysis` and `analysisfeature` tables used to store information about how the analysis was done and what scores resulted.

Feature locations: Rank



Resources: GMOD.org

GMOD for the Biologist

Download Software

FAQs and HOWTOs

Project Events

Contribute Code! GMOD Is open source

Project News

Contribute Doc! GMOD.org is a Wiki

Support: Help Desk, mailing lists, meetings

Popular GMOD Components

SECOND ANNUAL Arthropod Genomics SYMPOSIUM

Mailing List: GMOD-Schema

Log in / create account

article | discussion | edit | history

navigation
- GMOD Home
- Categories / Tags
- Getting Started
- Downloads
- View all pages

documentation
- Chado
- FAQs
- HOWTOs
- For the Biologist
- Glossary

community
- GMOD News
- Mailing Lists
- Calendar
- About this site

developers
- CVS
- SourceForge Site

search

Go | Search

GMOD Mailing Lists

This list contains most of the mailing lists relevant to GMOD. This table does not contain the more technical lists that automatically deliver every change or *commit* to the GMOD CVS repositories, see SourceForge for these mailing lists. Nabble has a GMOD mail mirror of these lists, which helps as the SourceForge mail services, especially searches, are (often) unavailable.

Overview Lists

Topic	List Link	Comment
Announcements	gmod-announce (SourceForge)	Low volume GMOD announcements. Moderated.
Architecture	gmod-architecture (SourceForge)	GMOD architecture working group list. Very low traffic.
GMOD Developers List	gmod-devel (SourceForge)	General GMOD developer list.

Component Lists

Mailing lists about specific GMOD Components.

Component	List Link	Comment
Apollo	apollo (FruitFly.org)	Apollo mailing list
BioMart	mart-dev BioMart.org	BioMart mailing list
Chado	gmod-schema (SourceForge)	All Chado issues
CMap	gmod-cmap (SourceForge)	CMap mailing list

http://gmod.org/GMOD_Mailing_Lists

SECOND ANNUAL Arthropod Genomics SYMPOSIUM

GMOD Help Desk

- User help
- Web site and documentation
- Education and outreach
- Developer support
- Promote GMOD for evolutionary biology
- Sponsored by NESCent
  - Funded by NIH grants to Ian Holmes (UC Berkeley) and James Hu (Texas A&M)
  - Dave in Oregon, Scott in Ohio, Lincoln in New York, Hilmar and Todd in North Carolina

GMOD

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

GMOD Meetings

GMOD 2008 Summer School

July 11-13, 2008

National Evolutionary Synthesis Center
(NESCent)
Durham, North Carolina, USA
http://gmod.org/GMOD_Summer_School
Applications due April 15!

GMOD July 2008 Community Meeting

July 16-17, 2008

University of Toronto
(before ISMB)
Toronto, Ontario, Canada
http://gmod.org/July_2008_GMOD_Meeting

GMOD

SECOND ANNUAL Arthropod Genomics SYMPOSIUM

Acknowledgements

FlyBase
Dave Emmert
FlyBase Team

BBOP
Chris Mungall
Ed Lee
Mark Gibson

CSHL
Lincoln Stein
Sheldon McKay
Ben Faga

ParameciumDB
Linda Sperling
Oliver Arnaiz

UCLA
Brian O'Connor
Allen Day

KSU
Sue Brown
Doris Merrill

U Oregon
Patrick Phillips

U Wisconsin
Xiaokang Pan

JCVI
Jonathan Crabtree

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

Thank You!

Scott Cain, PhD.
GMOD Project Coordinator
Cold Spring Harbor Laboratory
cain@cshl.edu

Dave Clements
GMOD Help Desk
National Evolutionary Synthesis Center
clements@nescent.org

SECOND ANNUAL
Arthropod Genomics
SYMPOSIUM

GMOD @ Arthropod Genomics Symposium

GMOD Users
wFleaBase
BeetleBase
FlyBase
AphidBase
ButterflyBase
VectorBase
BeeBase
HeliconiusBase
GnpAnnot-Lep
+ hundreds more

GMOD Tools
Chado
Apollo
GBrowse
CMap
Table Editor
GMODWeb
Ergatis
Textpresso
Sybil
BioMart
LuceGene
SynBrowse
Galaxy
InterMine
…

Workshops

Thursday 7-9pm
Community Contributions to Genome Annotation
Christine Elsik, Christopher Childers, Darren Hagen

Friday 7:30-9:30pm
Chado Databases and Integration with GMOD Tools
Scott Cain, Dave Clements

Talks

Friday 11:25-11:45am
Unlocated arthropod genes and ways to find them
Don Gilbert

Posters

Friday 5-6:30pm
A comparative annotation of Drosophilid dicistronic genes

Saturday 10-11:30am
A Complete System For Community Genome Annotation
GMOD: Database Resources for Emerging Model Organisms
Comparative genomics and database construction for Lepidoptera
Unlocated Arthropod genes, and ways to find them
VectorBase: A genome resource for arthropod vectors of human pathogens
The VectorBase Manual and Community Annotation Submission pipeline
ButterflyBase: A framework for comparative genomics in butterflies and moths

http://gmod.org

Got data?

External IDs

Database + Accession = Identifier

GO + 0043565 = GO:0043565

InterPro + IPR001356 = InterPro:IPR001356

YourDB + whatever = YourDB:whatever

Shown on web pages, published

Each defined DB has a prefix & URL

IDs can have associated names

FlyBase:

ID (Ontology)	SO:0000010 (Sequence C

wFleaBase:

Dbxref:	JGI:JGI_CBNO3000.rev

ZFIN:

Protein Families, Domains and Sites:

- InterPro:IPR001356 (1)
- InterPro:IPR012287 (1)
- InterPro:IPR013847 (1)
- PROSITE:PS00027 (1)
- PROSITE:PS50071 (1)

Internal IDs

Long integers

Used inside database

Rarely shown or published