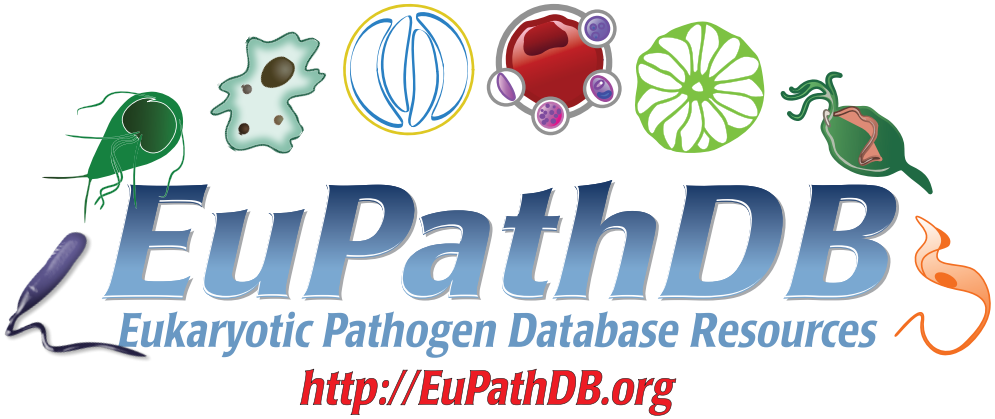
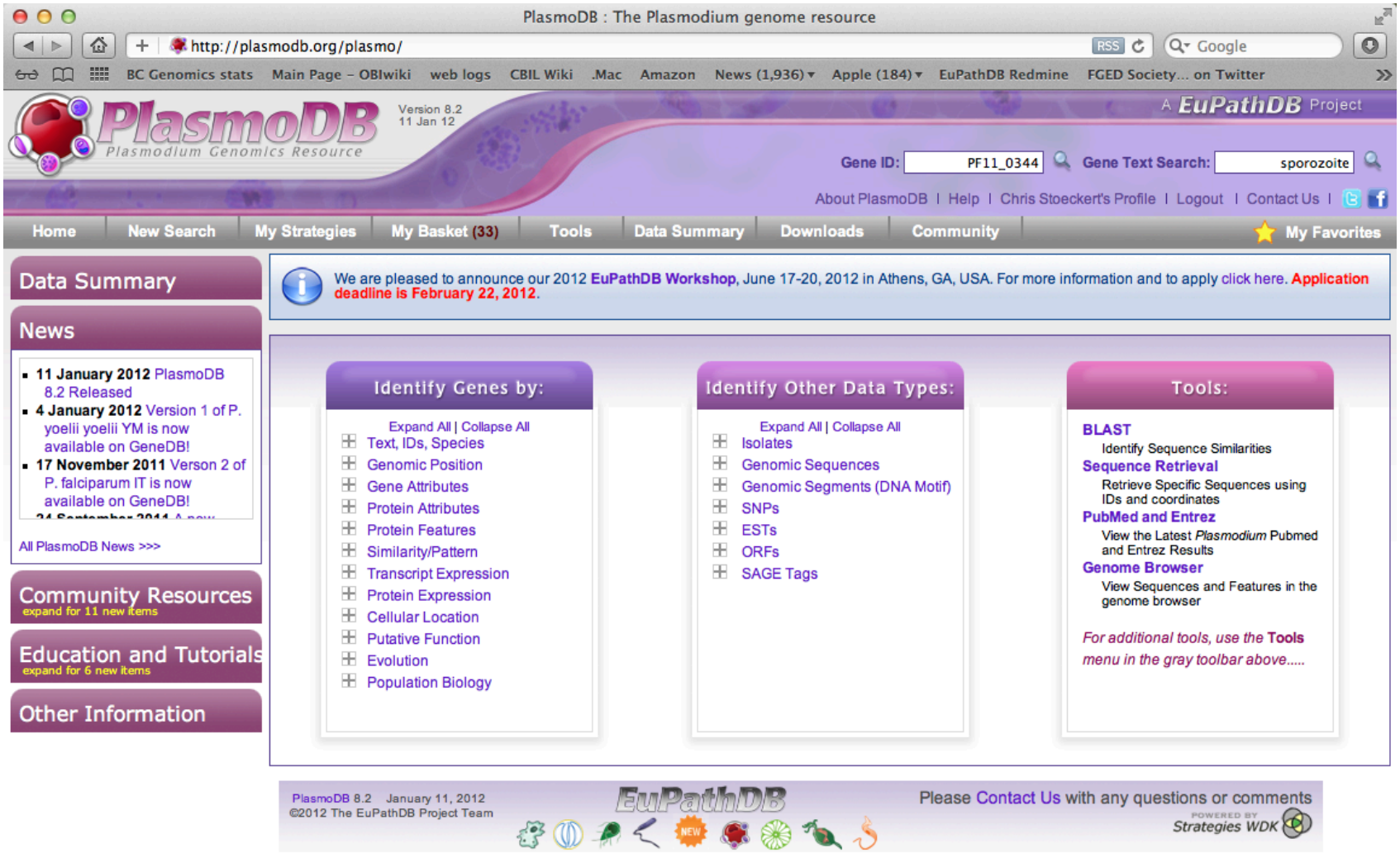


Issues in Infectious Disease Genomics

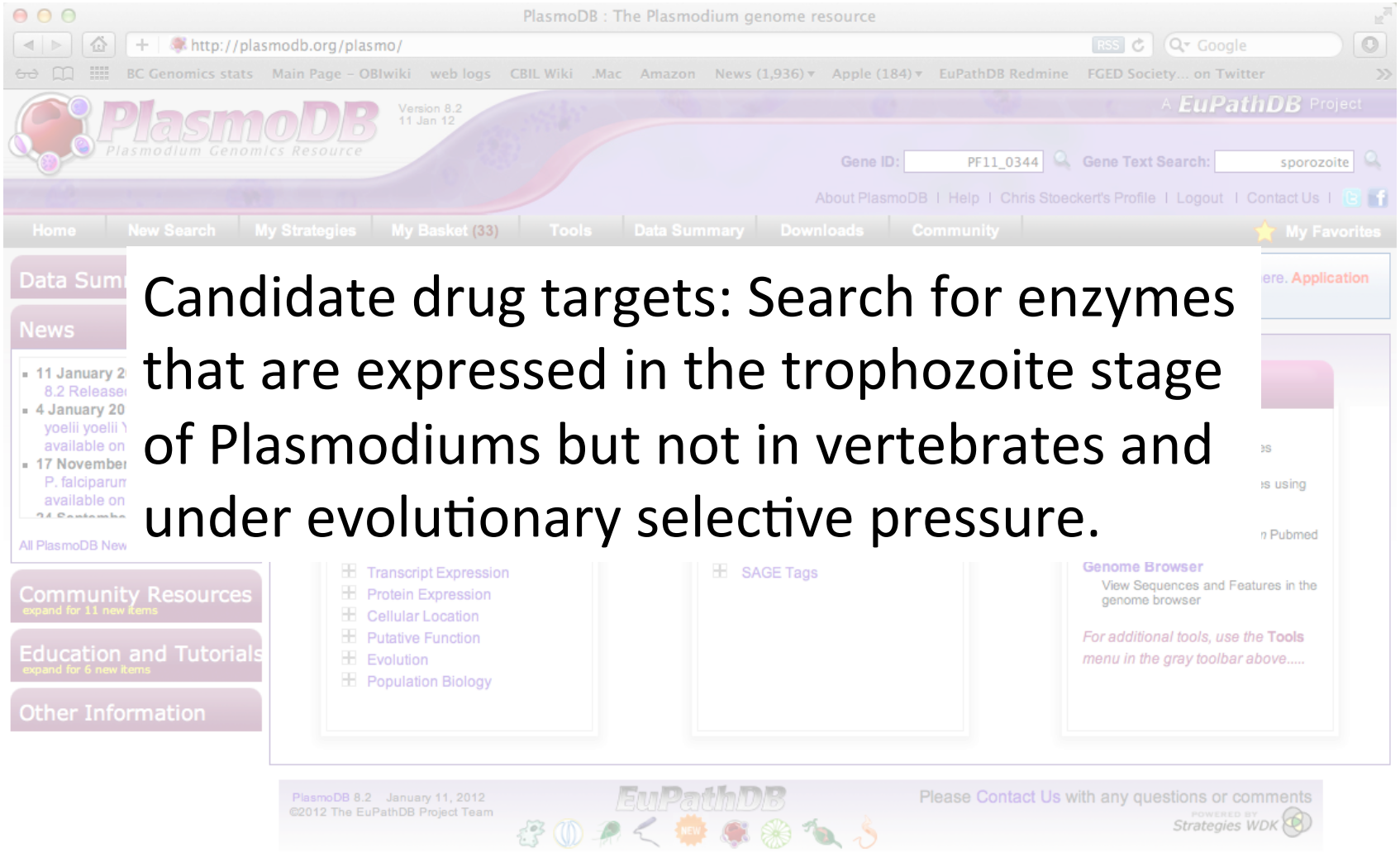
Chris Stoeckert
University of Pennsylvania
NIAID EuPathDB Bioinformatics Resource Center

http://eupathdb.org/eupathdb/

Data mining for malaria treatments

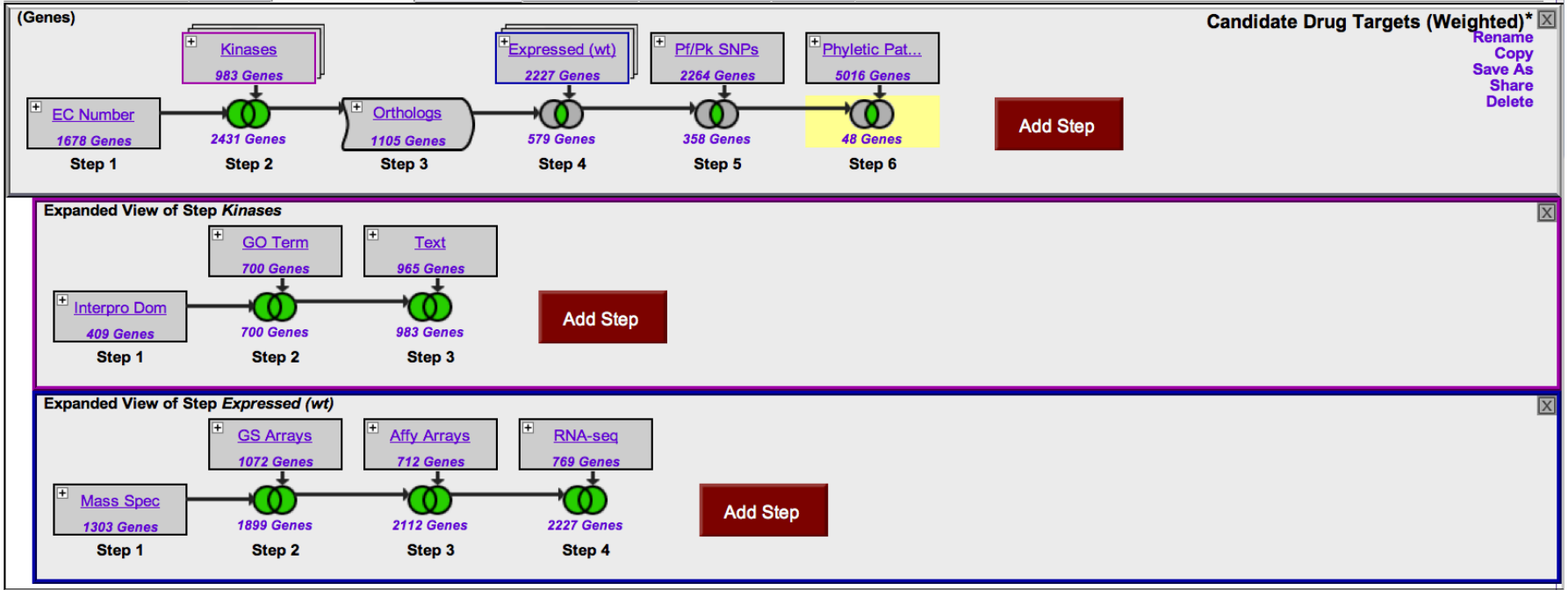


Data mining for malaria treatments



Candidate drug targets: Search for enzymes that are expressed in the trophozoite stage of Plasmodiums but not in vertebrates and under evolutionary selective pressure.

My Strategies: New | Opened (1) | All (136) | Basket | Examples | Help



Filter results by species *(results removed by the filter will not be combined into the next step.)*

Candidate Drug Targets (Weighted) - step 6 - 48 Genes

Add 48 Genes to Basket | Download 48 Genes

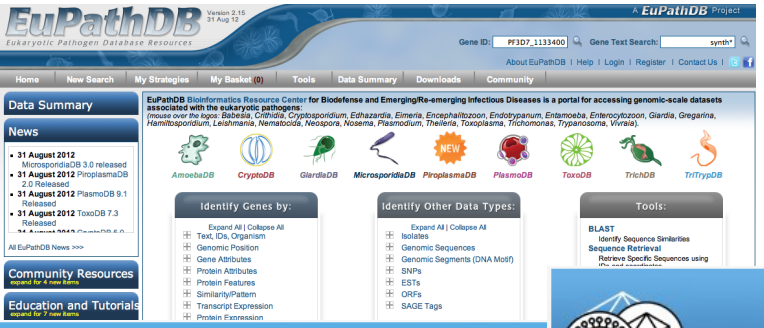
Advanced Paging | Select Columns | Reset Columns

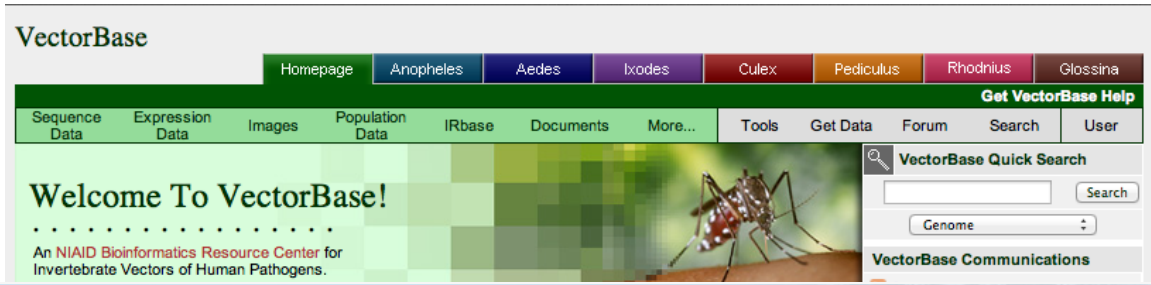
Gene Id	Genomic Location	Product Description	Weight
PFI0755c	Pf3D7_09: 650,576 - 654,832 (-)	6-phosphofructokinase	60
PF08_0132	Pf3D7_08: 147,210 - 151,403 (+)	glutamate dehydrogenase, putative	60
PFD0670c	Pf3D7_04: 626,769 - 627,785 (-)	lysine decarboxylase-like protein, putative	60
PFE0660c	Pf3D7_05: 569,180 - 569,917 (-)	purine nucleoside phosphorylase	60
PF13_0257	Pf3D7_13: 1,968,221 - 1,970,812 (-)	glutamate--tRNA ligase, putative	50
PF14_0541	Pf3D7_14: 2,329,869 - 2,332,022 (-)	V-type H( )-translocating pyrophosphatase, putative	50
MAL7P1.19	Pf3D7_07: 271,404 - 284,452 (-)	ubiquitin transferase, putative	40
PF11_0086	Pf3D7_11: 301,283 - 311,287 (-)	MIF4G domain containing protein	40
PF14_0649	Pf3D7_14: 2,786,244 - 2,794,259 (-)	conserved Plasmodium protein, unknown function	40
PF14_0614	Pf3D7_14: 2,619,614 - 2,624,122 (+)	phosphatase, putative	40
PF14_0063	Pf3D7_14: 239,747 - 243,772 (+)	ATP-dependent Clp protease, putative	40

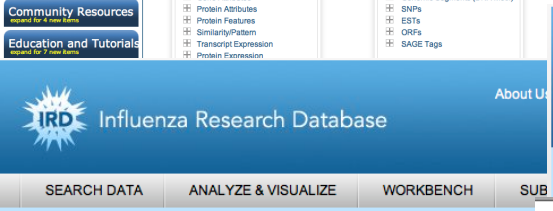
Strategies provides a way to dynamically create a data mining workflow.

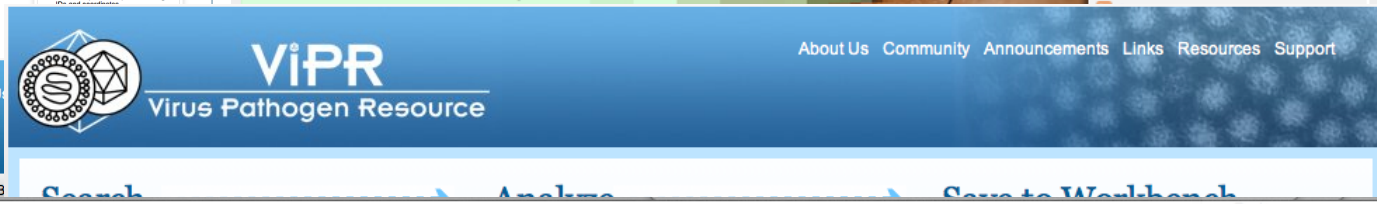
- Can save your strategies and share providing a reproducible record of your data mining
- Also used by TBDB, FungiDB, Schistodb (beta version), and Beta Cell Genomics
- http://code.google.com/p/strategies-wdk/
- Fischer et al. Database (Oxford) 2011:bar027
- **Although powerful, strategies are limited by the attributes that can be queried.**
  - Datasets are focused on quantifiable aspects with minimal information about the context (meta-data).
  - No common semantics for the meta-data in place requiring exploring individual datasets for details

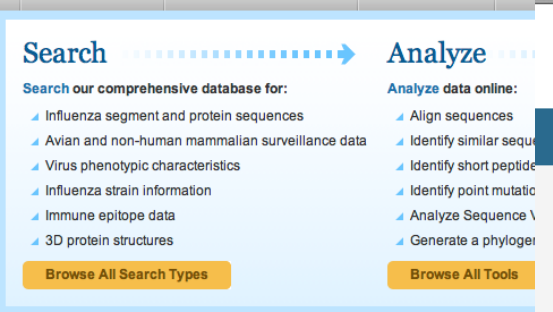
EuPathDB and the NIAID BRCs

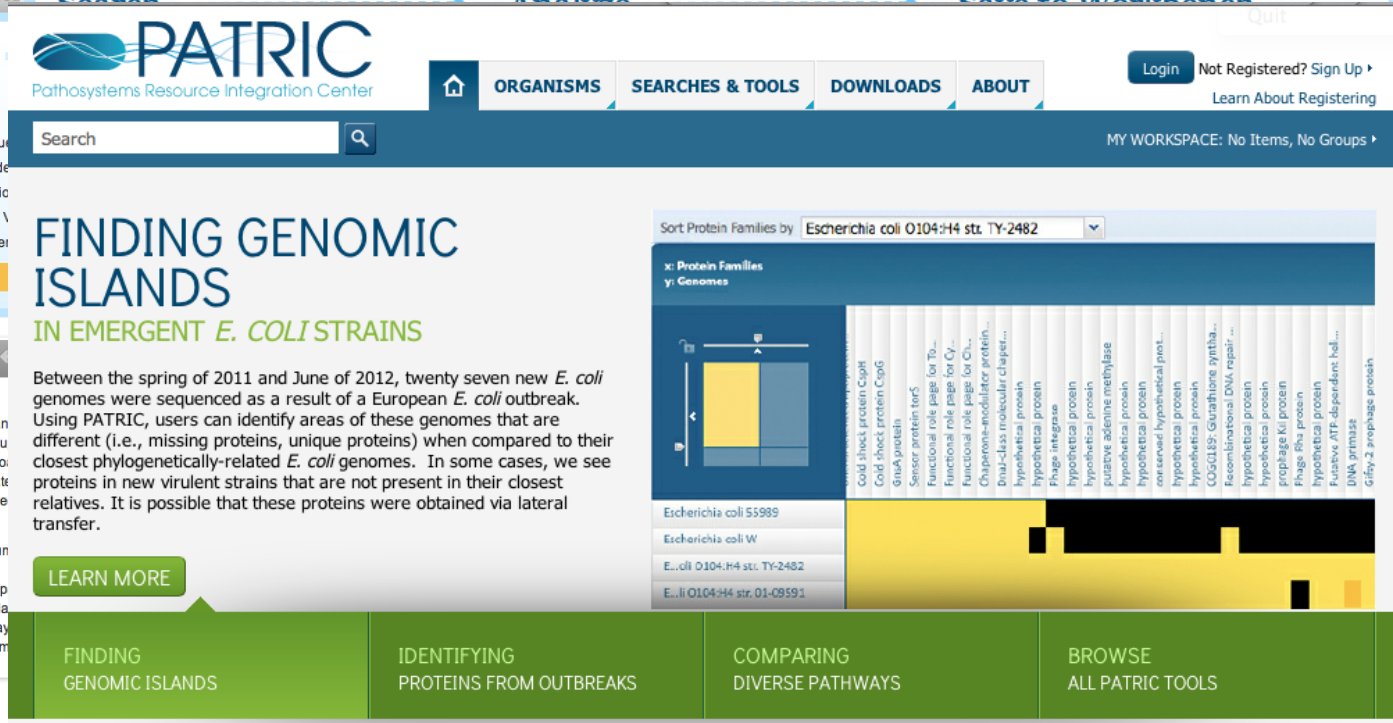


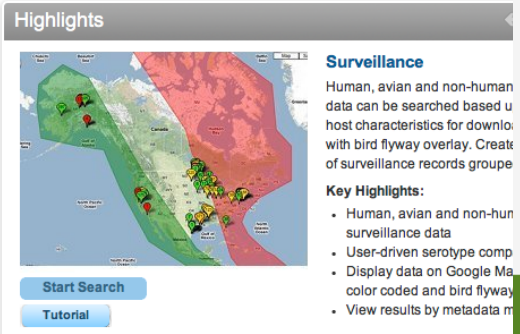




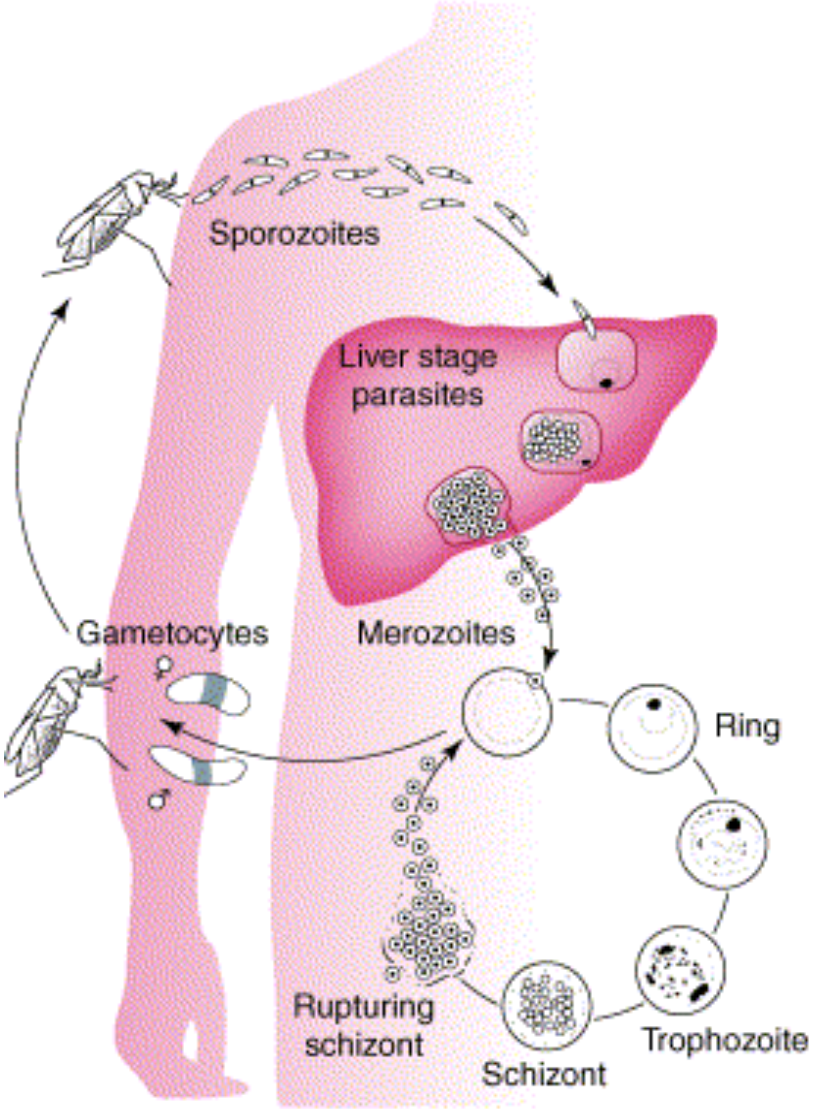






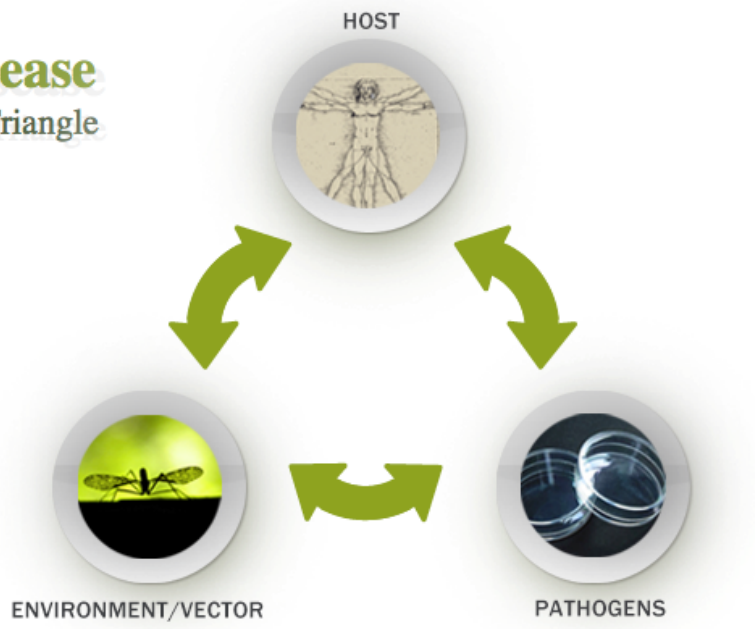


Infectious disease investigations requires integration of information on multiple species



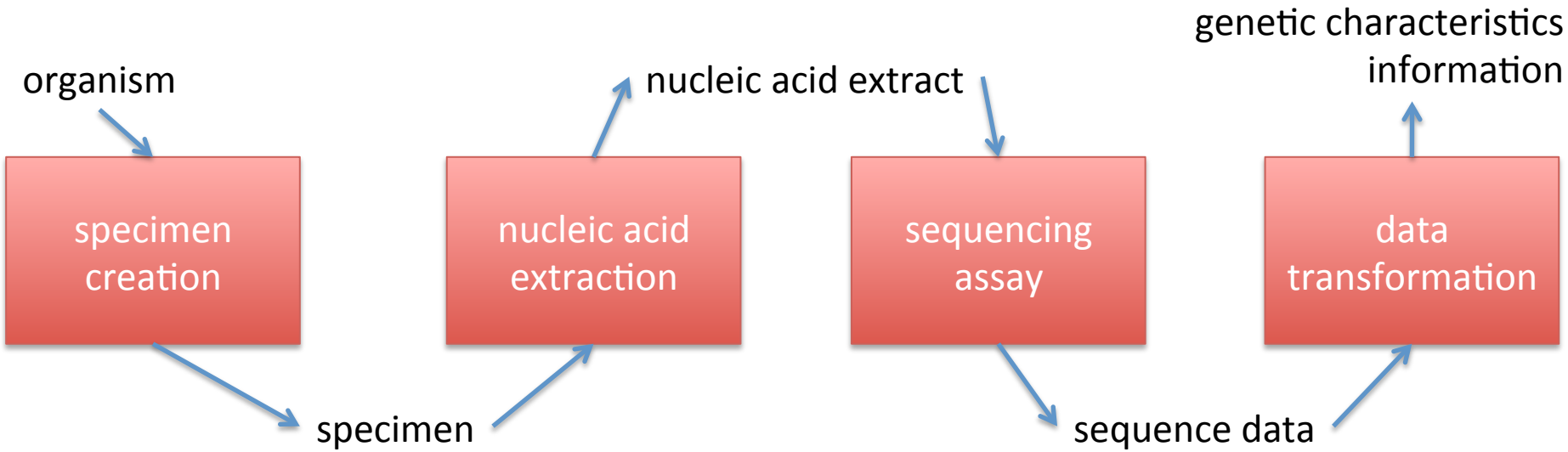
About | News and Announcements

isease
e Triangle



A Genome Sequencing Center (GSC)-Bioinformatics Resource Center (BRC) Investigation (Project)

Types of studies – drugs, vaccines, epidemiology



Specimen collection

Sample management

Sequencing

Data processing

Host info: vital stats, diagnosis, treatment
Pathogen info: which species/strain, level of infection

May be a mixture of host (vector) and pathogen

How to correlate phenotypes, SNPs, expression between host and pathogen?

Host-pathogen datasets in EuPathB

- Focus on parasite with host as experimental factor
  - find isolates based on host, location, source, ..
    - HTS SNPs from field isolates (Broad, Sanger)
  - compare expression profiles of pathogens based on host characteristics
    - NSR-seq transcriptional profiling enables identification of a gene signature of Plasmodium falciparum parasites infecting children. J. Clin. Invest. 2011;121(3):1119-29 Vignali et al.
- Need host-pathogen genomic comparisons
  - correlate expression between host and pathogen
    - *A comprehensive catalog of the T. gondii & N. caninum parasite & infected host cell transcriptome & proteome*. Brian Gregory (Univ. Pennsylvania) and Jonathan Wastling (Univ. Liverpool).

NIAID GSC-BRC Meta-data Working Group

- Standardized Metadata for Human Pathogen/Vector Genomic Sequences
- Bruce Birren[2,b], Laura Brinkac[1,a], Vincent Bruno[3,c], Elizabeth Caler[1,a], Ishwar Chandramouliswaran[1,a], Sinéad Chapman[2,b], Frank Collins[8,h], Christina Cuomo[2,b], Joana Carneiro Da Silva[3,c], Valentina Di Francesco[4], Vivien Dugan[1,a], Scott Emrich[8,h], Mark Eppinger[3,c], Michael Feldgarden[2,b], Claire Fraser[3,c], W. Florian Fricke[3,c], Maria Giovanni[4], Gloria Giraldo-Calderon[8,h], Omar S. Harb[5,g], Matt Henn[2,b], Erin Hine[3,c], Julie Dunning Hotopp[3,c], Jessica C. Kissinger[6,g], Eun Mi Lee[4], Punam Mathur[4], Garry Myers[3,c], Emmanuel Mongodin[3,c], Cheryl Murphy[2,b], Dan Neafsey[2,b], Karen Nelson[1,a], Ruchi Newman[2,b], William Nierman[1,a], Brett E. Pickett[1,d,e], Julia Puzak[4], David Rasko[3,c], David S. Roos[5,g], Lisa Sadzewica[3,c], Richard H. Scheuermann[1,d,e], Lynn M. Schriml[3,c], Bruno Sobral[7,f], Tim Stockwell[1,a], Chris Stoeckert[5,g], Dan Sullivan[7,f], Luke Tallon[3,c], Herve Tettelin[3,c], Doyle V. Ward[2,b], David Wentworth[1,a], Owen White[3,c], Rebecca Will[7,f], Jennifer Wortman[2,b], Alison Yao[4], Jie Zheng[5,g]

[1]J. Craig Venter Institute, Rockville, MD and San Diego, CA
[2]Broad Institute, Cambridge, MA
[3]Insitute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD
[4]National Institute of Allergy and Infectious Diseases, Rockville, MD
[5]University of Pennsylvania, Philadelphia, PA
[6]University of Georgia, Athens, GA
[7]Cyberinfrastructure Division, Virginia Bioinformatics Institute, Blacksburg, VA
[8]University of Notre Dame, South Bend, IN
[a]J. Craig Venter Institute Genome Sequencing Center for Infectious Diseases
[b]Broad Institute Genome Sequencing Center for Infectious Diseases
[c]Institute for Genome Sciences Genome Sequencing Center for Infectious Diseases
[d]Influenza Research Database Bioinformatics Resource Center
[e]Virus Pathogen Resource Bioinformatics Resource Center
[f]PATRIC Bioinformatics Resource Center
[g]EuPathDB Bioinformatics Resource Center
[h]VectorBase Bioinformatics Resource Center

NIAID GSC-BRC Meta-data Working Group

- Goal of group: development of an approach for the capture of standardized human pathogen/vector sequencing metadata designed to support epidemiologic and genotype-phenotype association studies
- Richard Scheuermann presented on this at the FGED Boston 2012 meeting
- Current status: The standard includes data fields about:
  - project leadership and support
  - the organism/environmental source of the pathogen/vector specimen
  - spatial-temporal information about the specimen isolation event
  - phenotypic characteristics of the pathogen/vector isolated
  - information about the sequencing and data processing methodologies used
  - information about the quality and coverage of the resulting sequence

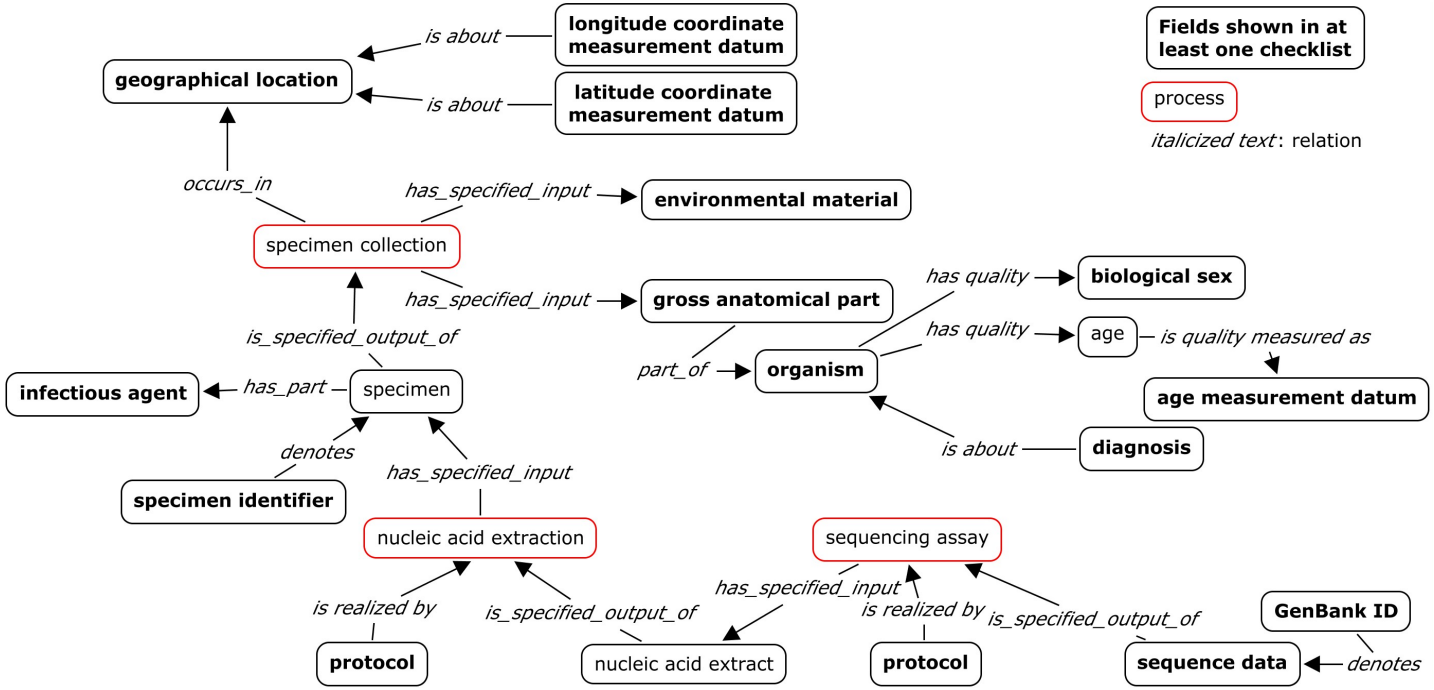
  For each data field, recommendations for preferred vocabularies, ontologies, and syntaxes, and mappings to other related data standards are provided.
- Mapped to terms from other metadata standards initiatives, including:
  - Genomic Standards Consortium's minimal information (MIxS) checklists
  - NCBI's BioSample metadata
  - Ontology of Biomedical Investigations (OBI). Serve as common semantic framework
- Coordination with others?

NIAID GSC-BRC Meta-data Working Group

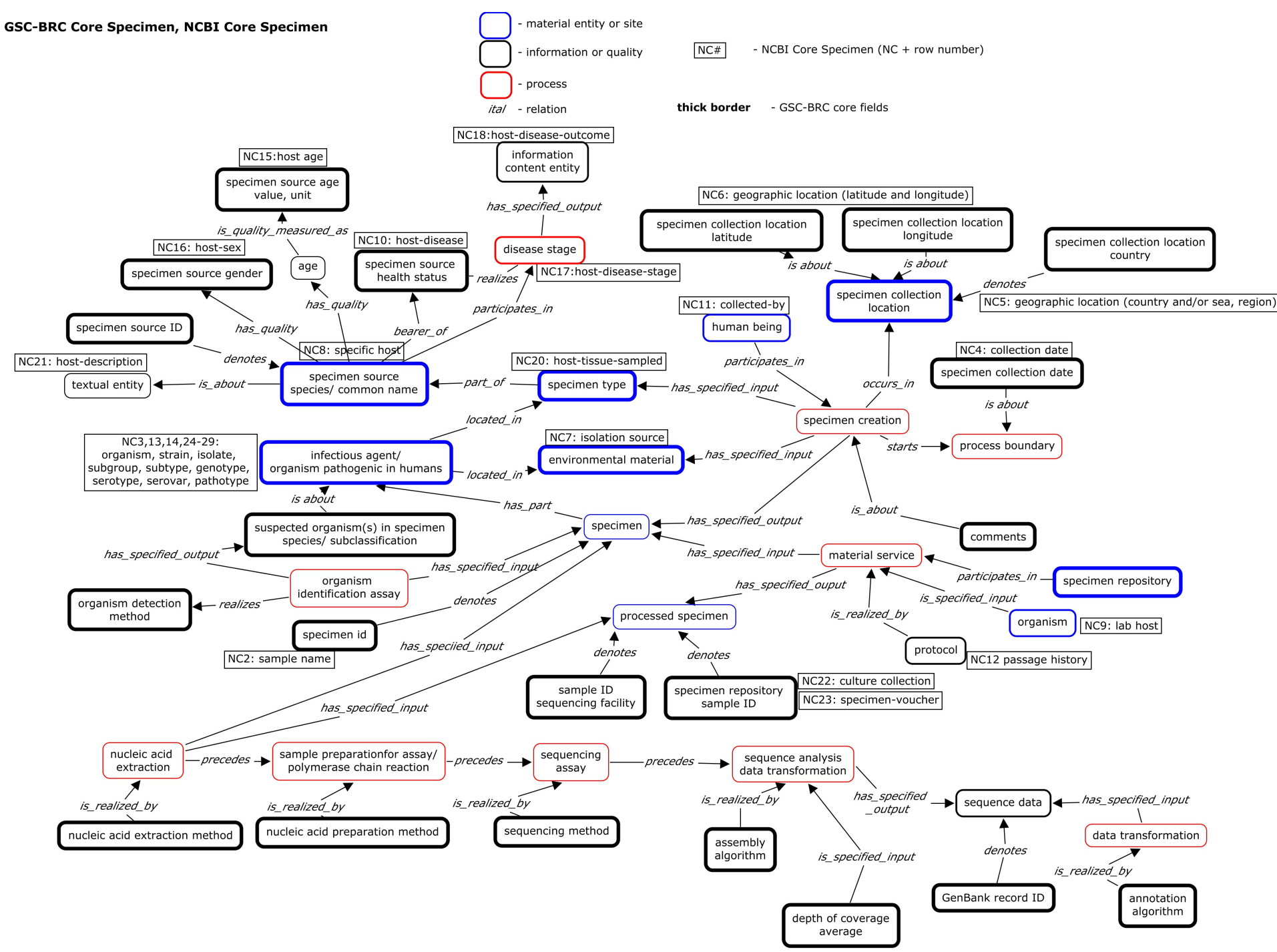
- Now beta testing a form for collection
  - Incentive is getting your sample sequenced
- Have a Clinical meta-data working group
  - Address additional details needed for patient-based data
  - Address privacy issues
  - work in progress (chaired by Jennifer Wortman, Chris Stoeckert)

Ontology of Biomedical Investigations (OBI) as unifying semantic framework

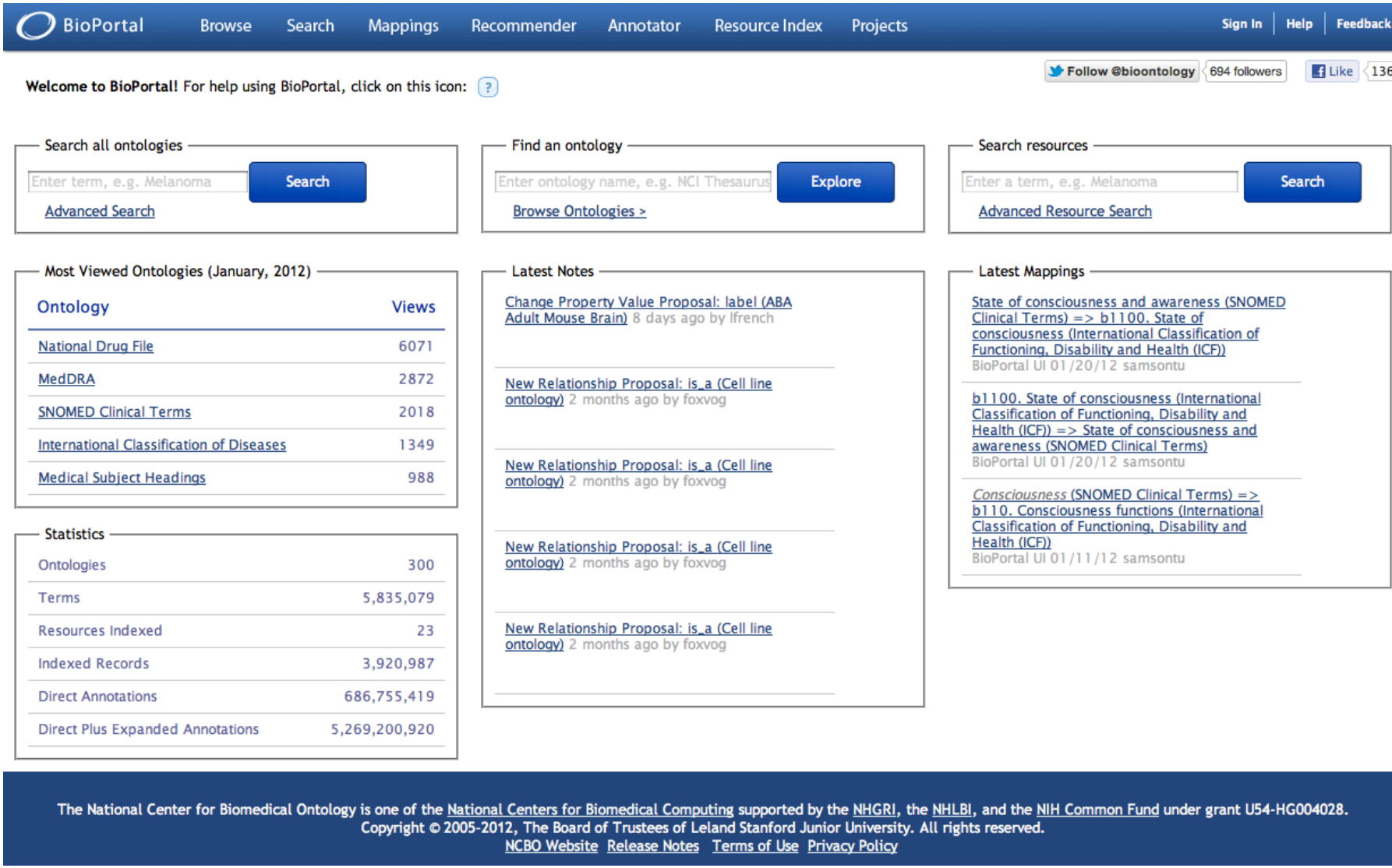
GSC-BRC Core Specimen	NCBI BioSample Specimen	EuPathDB Isolates	OBI/OBO Foundry ontology terms
Specimen ID	sample name	Isolate ID	specimen identifier
Suspected Organism(s) in Specimen	organism	Isolate Genus/Species	infectious agent
Specimen Type	host-tissue-sampled	Host Material Isolated from	gross anatomical part
Specimen Source Species	specific host	Host Species	organism
Specimen Source Gender	host-sex	Sex	biological sex
Specimen Source Age (Value, Unit)	host-age	Age	age measurement datum
Specimen Source Health Status	host-disease	Host Status/Symptoms	diagnosis (OGMS)
Environmental Material	isolation-source	Isolate Environmental Source	environmental material
Specimen Collection Location (Location,Country, Latitude,Longitude)	geographic location (country and/or sea, region)	Geographic Location (Country,Region,County,City/Village/locality, latitude/longitudeCoordinates, Altitude)	geographical location
Specimen Collection Location - Latitude		Latitude	latitude coordinate measurement datum
Specimen Collection Location - Longitude		Longitude	longitude coordinate measurement datum
Nucleic Acid Extraction Method			protocol is realized by nucleic acid extraction
Sequencing Method			protocal is realized by sequencing assay
		Sequence	sequence data
GenBank Record ID		GenBank # Identical to sequence	GenBank ID



Jie Zheng also using OBI as semantic framework for U Penn LIMS.


GSC-BRC Core Specimen, NCBI Core Specimen
- material entity or site
- information or quality
- process
ital - relation
NC# - NCBI Core Specimen (NC + row number)
thick border - GSC-BRC core fields
NC18:host-disease-outcome
information content entity
has_specified_output
NC15:host age
specimen source age value, unit
is_quality_measured_as
age
NC16: host-sex
specimen source gender
NC10: host-disease
specimen source health status
disease stage
realizes
NC17:host-disease-stage
has_quality
participates_in
bearer_of
NC6: geographic location (latitude and longitude)
specimen collection location latitude
specimen collection location longitude
specimen collection location country
is about
denotes
specimen collection location
NC5: geographic location (country and/or sea, region)
NC11: collected-by
human being
specimen source ID
has_quality
NC8: specific host
denotes
NC21: host-description
textual entity
is_about
specimen source species/ common name
part_of
NC20: host-tissue-sampled
specimen type
has_specified_input
participates_in
occurs_in
NC4: collection date
specimen collection date
is about
specimen creation
starts
process boundary
located_in
NC3,13,14,24-29: organism, strain, isolate, subgroup, subtype, genotype, serotype, serovar, pathotype
infectious agent/ organism pathogenic in humans
NC7: isolation source
environmental material
has_specified_input
located_in
is about
has_part
specimen
has_specified_output
is_about
comments
suspected organism(s) in specimen species/ subclassification
has_specified_output
has_specified_input
material service
has_specified_input
organism identification assay
has_specified_ouput
participates_in
specimen repository
organism detection method
realizes
denotes
processed specimen
is_specified_input
is_realized_by
organism
NC9: lab host
specimen id
NC2: sample name
has_speciied_input
denotes
denotes
protocol
NC12 passage history
sample ID sequencing facility
specimen repository sample ID
NC22: culture collection
NC23: specimen-voucher
has_specified_input
nucleic acid extraction
precedes
sample preparationfor assay/ polymerase chain reaction
precedes
sequencing assay
precedes
sequence analysis data transformation
is_realized_by
nucleic acid extraction method
is_realized_by
nucleic acid preparation method
is_realized_by
sequencing method
is_realized_by
assembly algorithm
has_specified_output
sequence data
has_specified_input
data transformation
is_specified_input
depth of coverage average
denotes
GenBank record ID
is_realized_by
annotation algorithm

OBI is available at the NCBO Bioportal http://bioportal.bioontology.org/



Versions

VERSION	RELEASE DATE	UPLOAD DATE	DOWNLOADS
2012-07-01	07/16/2012	07/16/2012	Ontology
2012-03-29	03/30/2012	04/10/2012	Ontology
2011-12-13	01/27/2012	01/27/2012	Ontology
2011-07-20	07/20/2011	08/05/2011	Ontology
2011-04-20	04/20/2011	05/10/2011	Ontology

more...

Views

Create new view

Expand All | Collapse All

FGED View

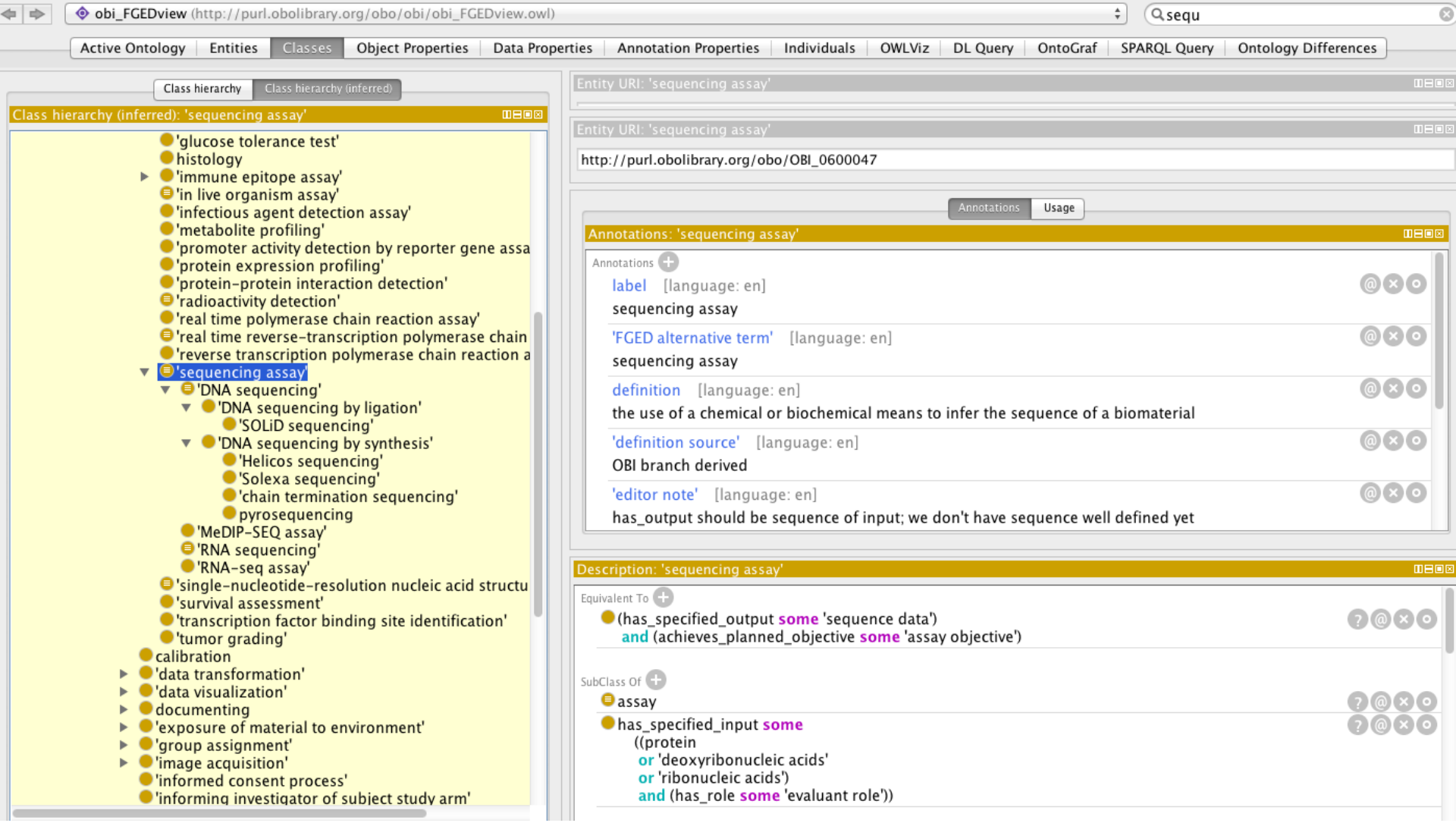
- Description: Subset of OBI terms of interest to the FGED Community.
- Ontology ID: 2070
- Definition Language: Manual

VERSION	BASE VERSION	CREATED	CREATED BY	ONTOLOGY FILE	DIFF FILE	VISIBILITY
2011-12-13	2011-12-13	01/27/2012	Jie Zheng, jiezheng@pcbi.upenn.edu	Download View		Public
2	2011-07-20	10/04/2011	Jie Zheng, jiezheng@pcbi.upenn.edu	Download View		Public
1	2011-04-20	06/20/2011	Jie Zheng, jiezheng@pcbi.upenn.edu	Download View		Public

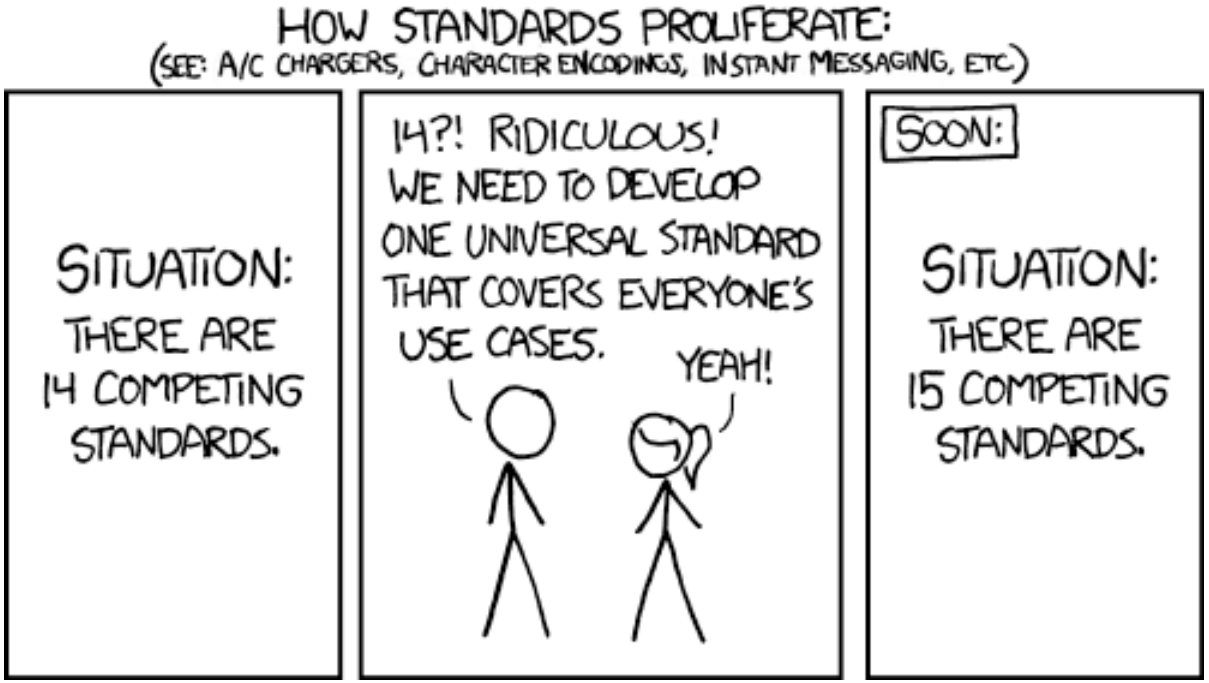
IEDB View

OBI Device branch

Sequencing assay in FGED View of OBI



Reality check



http://imgs.xkcd.com/comics/standards.png

Recommendations

- What recommendations do you (or the community you represent) have for better ways to share genomic data and ensuring reproducible research?
  - Standards for *accessing* annotations and expression data to compare across species (get pile-ups or calls and protocols?)
  - Standards for *accessing* metadata for functional genomics datasets (SPARQL endpoints?)
- What challenges have you met and/or steps forward have you made with creating or using standards?
  - Core meta-data elements for NIAID GSC-BRCs
  - Coordinating with NCBI and potentially other groups
  - Ontology of Biomedical Investigations (OBI) as unifying semantic framework

Team Science

- EuPathDB
  - David Roos Jessie Kissinger
  - Brian Brunk, Eileen Kraemer, Omar Harb, Steve Fischer, Cristina Aurrechio, John Brestelli, Mark Heiges, Debbie Pinney,
  - Ana Barreto, JaShon Cade, RyanDoherty, Bindu Gajria, Xin Gao, Alan Gingle, Greg Grant, Sufen Hu, John Iodice, Wei Li, Brian Pitts, Ganesh Srinivasamoorthy, Haiming Wang, Susanne Warrenfeltz, Mariann Winkelmann
- OBI
  - OBI Consortium: http://obi-ontology.org/page/Main_Page
    - Jie Zheng
  - NIAID GSC-BRC Meta-data Working Group
    - Richard Scheuermann
- NIAID, NIGMS, NHGRI, Bill & Melinda Gates Foundation