

Issues in Infectious Disease Genomics

Chris Stoeckert
University of Pennsylvania
NIAID EuPathDB Bioinformatics
Resource Center

http://eupathdb.org/eupathdb/

Data Summary

News

EuPathDB Bioinformatics associated with the euk (mouse over the logos: Bal, Hamiltosporidium, Leish)

Genome annotation
Gene structure
Gene expression & regulation
Genomic variation

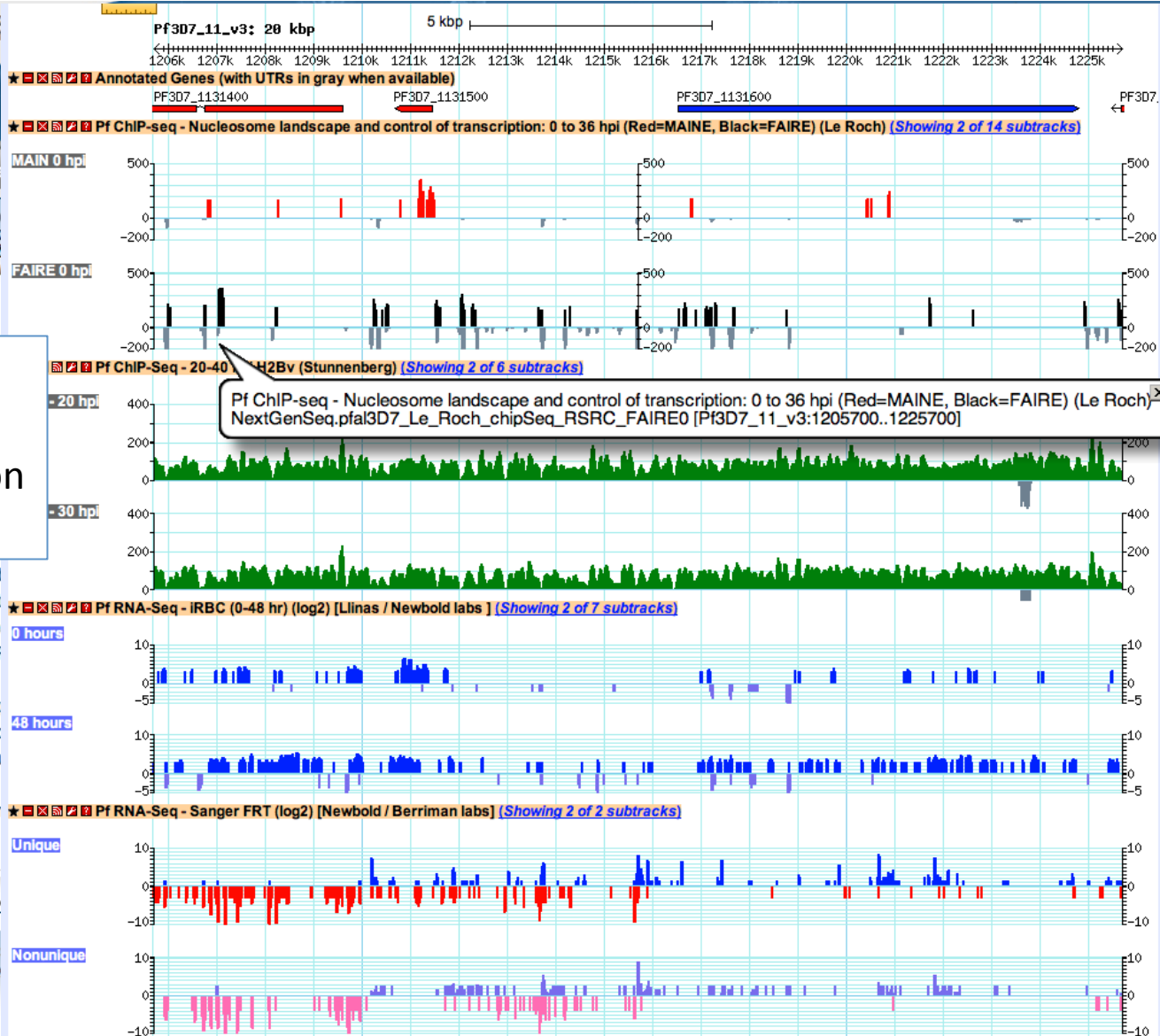
Community Resources
expand for 4 new items

Education and Tutorials
expand for 7 new items

Other Information
expand for 7 new items

- Gene Attril
- Protein Att
- Protein Fe
- Similarity/f
- Transcript
- Protein Ex
- Cellular Lc
- Putative Fi
- Evolution
- Populatio

D



Data mining for malaria treatments

PlasmoDB : The Plasmodium genome resource

http://plasmodb.org/plasmo/

BC Genomics stats Main Page - OBIwiki web logs CBIL Wiki .Mac Amazon News (1,936) Apple (184) EuPathDB Redmine FGED Society... on Twitter

PlasmoDB Plasmodium Genomics Resource Version 8.2 11 Jan 12

A EuPathDB Project

Gene ID: PF11_0344 Gene Text Search: sporozoite

About PlasmoDB | Help | Chris Stoeckert's Profile | Logout | Contact Us

Home New Search My Strategies My Basket (33) Tools Data Summary Downloads Community My Favorites

Data Summary

News

- 11 January 2012 PlasmoDB 8.2 Released
- 4 January 2012 Version 1 of P. yoelii yoelii YM is now available on GeneDB!
- 17 November 2011 Version 2 of P. falciparum IT is now available on GeneDB!

All PlasmoDB News >>>

Community Resources
expand for 11 new items

Education and Tutorials
expand for 6 new items

Other Information

Identify Genes by:

- Expand All | Collapse All
- Text, IDs, Species
- Genomic Position
- Gene Attributes
- Protein Attributes
- Protein Features
- Similarity/Pattern
- Transcript Expression
- Protein Expression
- Cellular Location
- Putative Function
- Evolution
- Population Biology

Identify Other Data Types:

- Expand All | Collapse All
- Isolates
- Genomic Sequences
- Genomic Segments (DNA Motif)
- SNPs
- ESTs
- ORFs
- SAGE Tags

Tools:

- BLAST**
Identify Sequence Similarities
- Sequence Retrieval**
Retrieve Specific Sequences using IDs and coordinates
- PubMed and Entrez**
View the Latest Plasmodium Pubmed and Entrez Results
- Genome Browser**
View Sequences and Features in the genome browser

For additional tools, use the Tools menu in the gray toolbar above.....

PlasmoDB 8.2 January 11, 2012 ©2012 The EuPathDB Project Team

EuPathDB

Please Contact Us with any questions or comments

POWERED BY Strategies WDK

Data mining for malaria treatments

The screenshot shows the PlasmoDB website interface. At the top, the browser address bar displays "http://plasmodb.org/plasmo/". The website header includes the PlasmoDB logo, "Version 8.2 11 Jan 12", and "A EuPathDB Project". A search bar contains "Gene ID: PF11_0344" and "Gene Text Search: sporozoite". The navigation menu includes "Home", "New Search", "My Strategies", "My Basket (33)", "Tools", "Data Summary", "Downloads", "Community", and "My Favorites".

On the left side, there are sections for "Data Summary" and "News". The "News" section lists several updates, including "11 January 2012 8.2 Release" and "4 January 2012 yoellii yoellii \ available on".

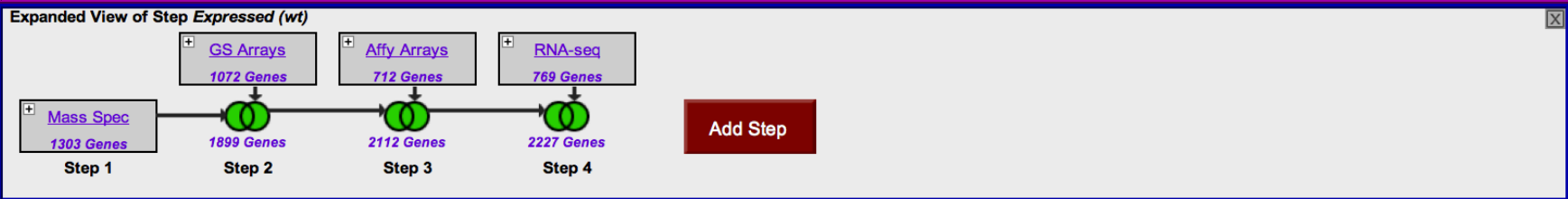
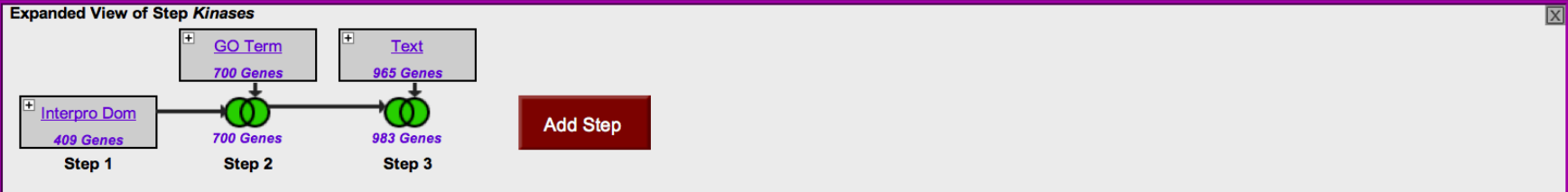
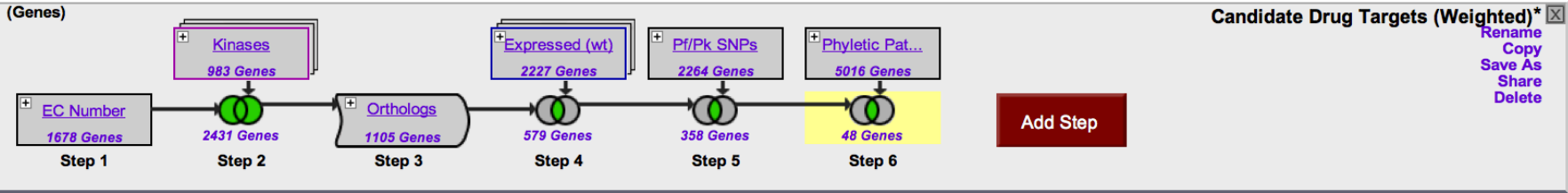
The main content area features a navigation menu with "Home", "New Search", "My Strategies", "My Basket (33)", "Tools", "Data Summary", "Downloads", "Community", and "My Favorites". Below this, there are several panels:

- Transcript Expression**: A list of categories including Protein Expression, Cellular Location, Putative Function, Evolution, and Population Biology.
- SAGE Tags**: A panel for SAGE Tags.
- Genome Browser**: A panel for viewing sequences and features in the genome browser, with a note: "For additional tools, use the Tools menu in the gray toolbar above...."

At the bottom, the footer includes "PlasmoDB 8.2 January 11, 2012 ©2012 The EuPathDB Project Team", the EuPathDB logo, and "Please Contact Us with any questions or comments". It also mentions "POWERED BY Strategies WDK".

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

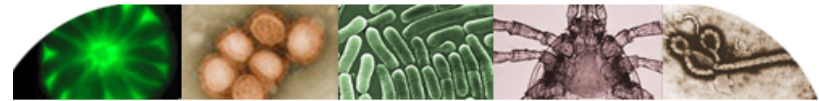
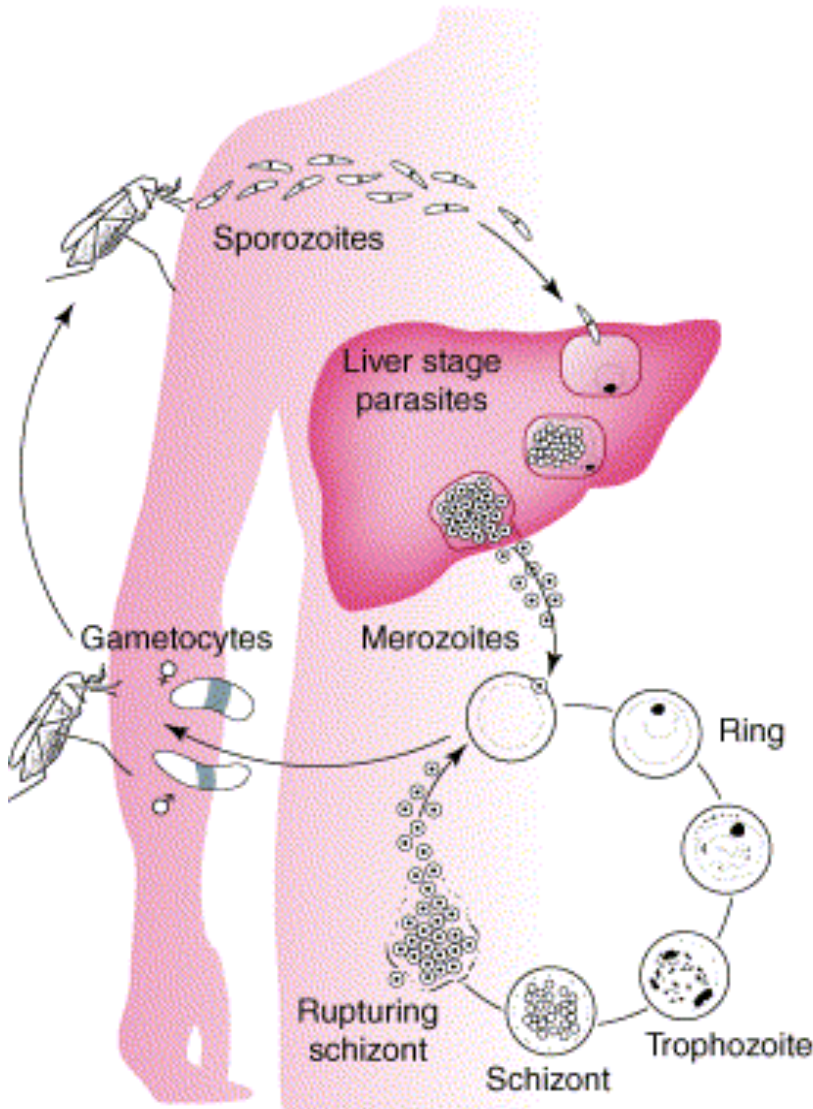
Gene Id	Genomic Location	Product Description	Weight
PF10755c	PF3D7_09: 650,576 - 654,832 (-)	6-phosphofructokinase	60
PF08_0132	PF3D7_08: 147,210 - 151,403 (+)	glutamate dehydrogenase, putative	60
PFD0670c	PF3D7_04: 626,769 - 627,785 (-)	lysine decarboxylase-like protein, putative	60
PFE0660c	PF3D7_05: 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

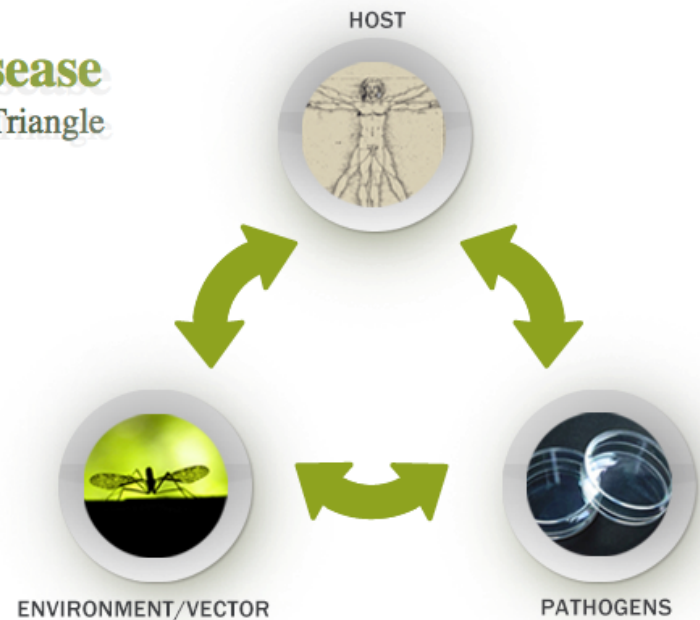
Infectious disease investigations requires integration of information on multiple species



About

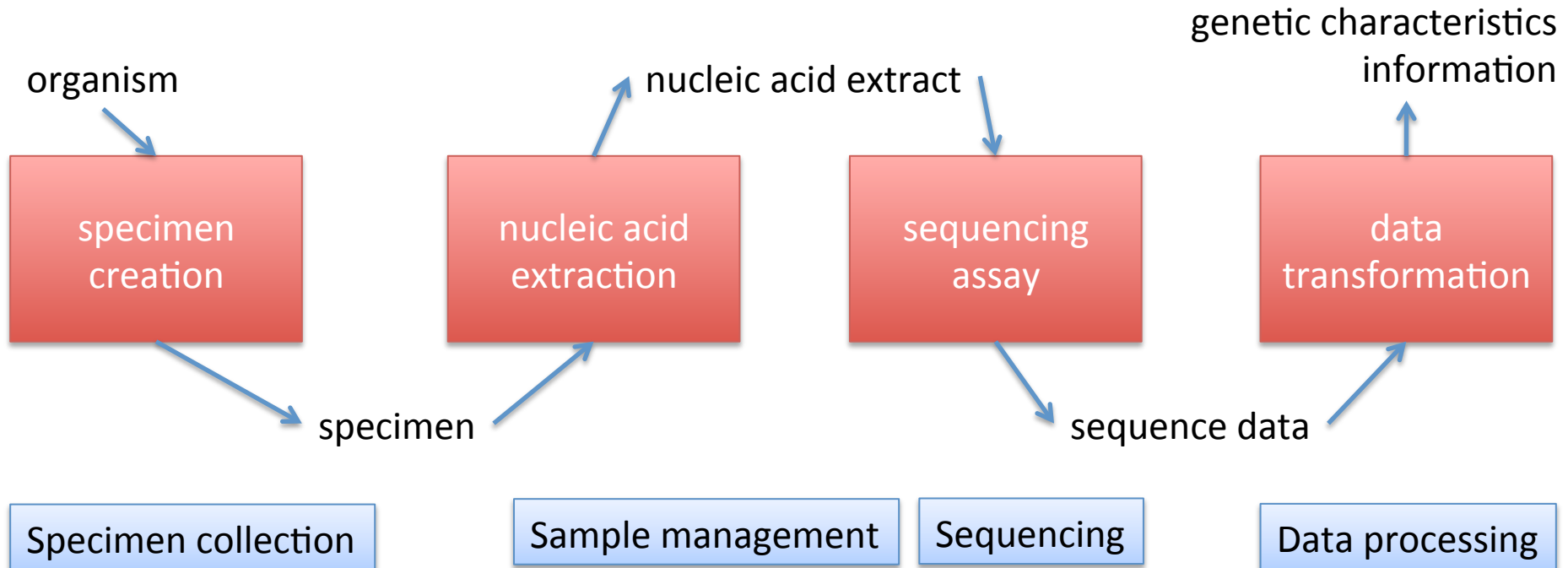
News and Announcements

disease
e Triangle



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

Types of studies – drugs, vaccines, epidemiology



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 Francesco⁴, 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 Giovanni⁴, 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 Lee⁴, Punam Mathur⁴, 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 Puzak⁴, David Rasko^{3,c}, David S. Roos^{5,g}, Lisa Sadzewicz^{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 Yao⁴, Jie Zheng^{5,g}

¹J. Craig Venter Institute, Rockville, MD and San Diego, CA

²Broad Institute, Cambridge, MA

³Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD

⁴National Institute of Allergy and Infectious Diseases, Rockville, MD

⁵University of Pennsylvania, Philadelphia, PA

⁶University of Georgia, Athens, GA

⁷Cyberinfrastructure Division, Virginia Bioinformatics Institute, Blacksburg, VA

⁸University of Notre Dame, South Bend, IN

^aJ. Craig Venter Institute Genome Sequencing Center for Infectious Diseases

^bBroad Institute Genome Sequencing Center for Infectious Diseases

^cInstitute for Genome Sciences Genome Sequencing Center for Infectious Diseases

^dInfluenza Research Database Bioinformatics Resource Center

^eVirus Pathogen Resource Bioinformatics Resource Center

^fPATRIC Bioinformatics Resource Center

^gEuPathDB Bioinformatics Resource Center

^hVectorBase Bioinformatics Resource Center

NIAID GSC-BRC Meta-data Working Group

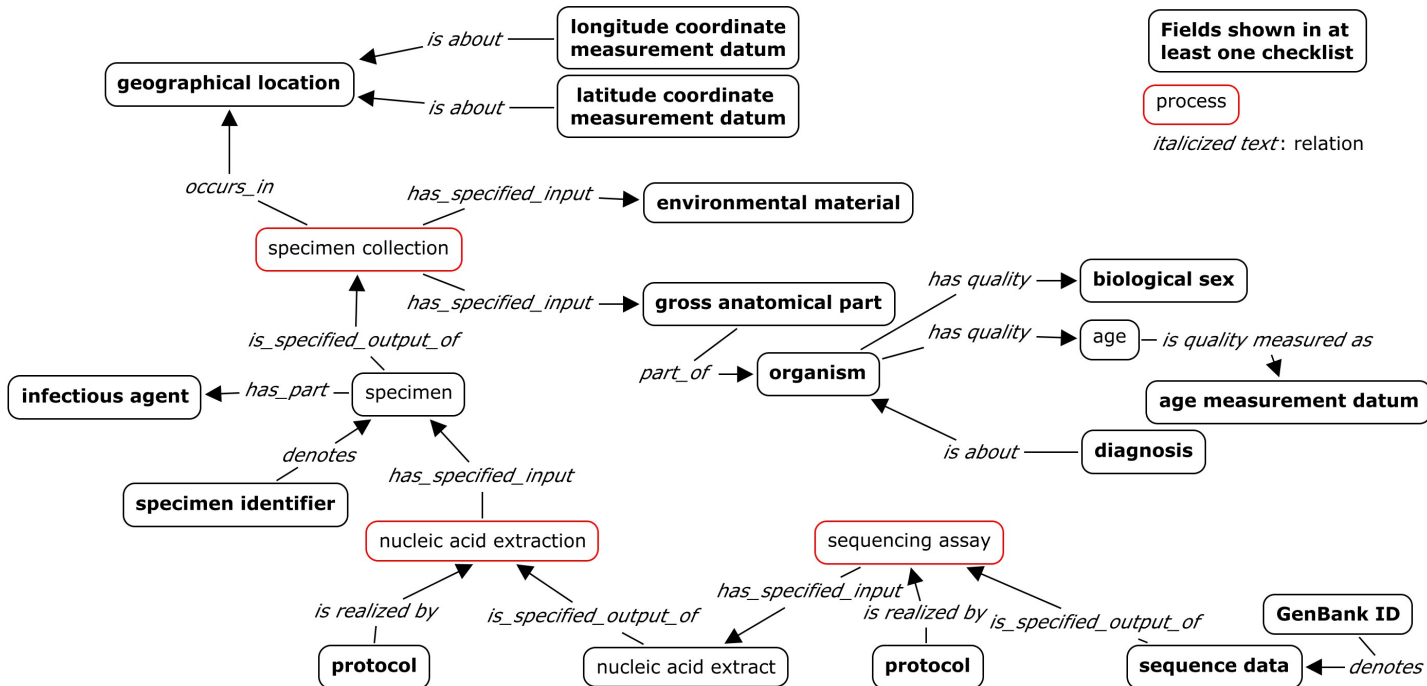
- Goal of group: development of an approach for the capture of standardized human pathogen/vector sequencing metadata designed to support epidemiologic and genotype-phenotype association studies
- Richard Scheuermann presented on this at the FGED Boston 2012 meeting
- Current status: The standard includes 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 sequenceFor 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			protocol 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.

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



Welcome to BioPortal! For help using BioPortal, click on this icon: [?](#)

Search all ontologies

[Advanced Search](#)

Find an ontology

[Browse Ontologies >](#)

Search resources

[Advanced Resource Search](#)

Most Viewed Ontologies (January, 2012)

Ontology	Views
National Drug File	6071
MedDRA	2872
SNOMED Clinical Terms	2018
International Classification of Diseases	1349
Medical Subject Headings	988

- Latest Notes
- [Change Property Value Proposal: label \(ABA Adult Mouse Brain\)](#) 8 days ago by lfrench
 - [New Relationship Proposal: is_a \(Cell line ontology\)](#) 2 months ago by foxvog
 - [New Relationship Proposal: is_a \(Cell line ontology\)](#) 2 months ago by foxvog
 - [New Relationship Proposal: is_a \(Cell line ontology\)](#) 2 months ago by foxvog
 - [New Relationship Proposal: is_a \(Cell line ontology\)](#) 2 months ago by foxvog

- Latest Mappings
- [State of consciousness and awareness \(SNOMED Clinical Terms\) => b1100. State of consciousness \(International Classification of Functioning, Disability and Health \(ICF\)\)](#)
BioPortal UI 01/20/12 samsontu
 - [b1100. State of consciousness \(International Classification of Functioning, Disability and Health \(ICF\)\) => State of consciousness and awareness \(SNOMED Clinical Terms\)](#)
BioPortal UI 01/20/12 samsontu
 - [Consciousness \(SNOMED Clinical Terms\) => b110. Consciousness functions \(International Classification of Functioning, Disability and Health \(ICF\)\)](#)
BioPortal UI 01/11/12 samsontu

Statistics

Ontologies	300
Terms	5,835,079
Resources Indexed	23
Indexed Records	3,920,987
Direct Annotations	686,755,419
Direct Plus Expanded Annotations	5,269,200,920

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

The screenshot displays the OBI FGED View interface. The top navigation bar includes tabs for 'Active Ontology', 'Entities', 'Classes', 'Object Properties', 'Data Properties', 'Annotation Properties', 'Individuals', 'OWL Viz', 'DL Query', 'OntoGraf', 'SPARQL Query', and 'Ontology Differences'. The search bar contains the text 'sequ'.

The left panel shows the 'Class hierarchy (inferred): 'sequencing assay'' with a tree structure of classes. The 'sequencing assay' class is highlighted, and its sub-classes are expanded, including 'DNA sequencing', 'RNA sequencing', and 'RNA-seq assay'.

The right panel shows the 'Entity URI: 'sequencing assay'' with the URI 'http://purl.obolibrary.org/obo/OBI_0600047'. Below this, the 'Annotations: 'sequencing assay'' section lists several annotations:

- label** [language: en]: sequencing assay
- 'FGED alternative term'** [language: en]: sequencing assay
- definition** [language: en]: the use of a chemical or biochemical means to infer the sequence of a biomaterial
- 'definition source'** [language: en]: OBI branch derived
- 'editor note'** [language: en]: has_output should be sequence of input; we don't have sequence well defined yet

The 'Description: 'sequencing assay'' section shows the following description:

Equivalent To: (has_specified_output **some** 'sequence data') and (achieves_planned_objective **some** 'assay objective')

SubClass Of:

- assay
- has_specified_input **some** ((protein or 'deoxyribonucleic acids' or 'ribonucleic acids') and (has_role **some** 'evaluant role'))

Reality check

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)



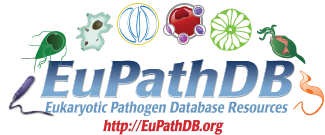
<http://imgs.xkcd.com/comics/standards.png>

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