

Integrating Data and Models Using the Disease Ontology

Prof. Warren Kibbe & Prof. Lynn Schriml

(Northwestern University)
wakibbe@northwestern.edu

(U. Maryland School of Medicine)
lschriml@som.umaryland.edu



Jan 31-Feb 4, 2009

Acknowledgements

Steering Committee	Contributors to this talk
Rex Chisholm, DO Founder, NU	John Osborne
Michael Ashburner, EBI	Simon Lin
Barry Smith,	Jared Flatow
Richard Scheuerman, UT Southwestern	Julie Zhu
Suzi Lewis, UC Berkeley	Gilbert Feng
Maryann Martone, UC San Diego	Pan Du
Chris Mungall, Lawrence Berkeley Lab	Wendy Wolf
	Abel Kho
	Steve Roessingh
	Dong Fu
	Eric Neumann



DO Mission Statement

The Disease Ontology provides a unifying structure to map human disease knowledge between datasets such as patient records and large scale genome, sequencing and microbiome projects.



Overview

- What is DO
- Current status of DO
 - Source/external vocabularies
- DO Use Cases
- DO Applications
- Join us in improving DO!

Disease Ontology Wiki

Log in / create account

page discussion view source history

Main Page

Contents [hide]

- 1 Disease Ontology Wiki
- 2 Mission
- 3 Scope
- 4 Meetings
- 5 Ongoing Development
- 6 Pls
- 7 Collaborators
- 8 Project Links
- 9 Related Links

Disease Ontology Wiki

Welcome to the Disease Ontology Community Wiki. The Disease Ontology was initially developed as part of the [NUGene](#) project starting in 2003 at Northwestern. Work continues on the [NUGene](#) project at Northwestern.

- The Disease Ontology is organized into five high level bins.
- The Disease Ontology in August 2006 was submitted for inclusion and review to the OBO Foundry.
- This project is open to collaborative development. Please contact the project Pls, Lynn Schriml at lynn.schriml_at_gmail.com or Warren Kibbe at wakibbe_at_gmail.com

Mission

The mission the Disease Ontology (DO) is to provide an open source ontology for the integration of biomedical data that is associated with human disease. DO will have a formally correct (in the ontology sense), semantically computable structure. Terms in DO will be well defined, using standard references. These terms will be linked to well-established, well-adopted terminologies that contain disease and disease-related concepts such as SNOMED (we are working with SNOMED to see if we can release SNOMED codes linked via UMLS to the community), ICD-9 and ICD-10, MeSH, and UMLS. The combination of a semantically computable structure and the external references to these terminologies will enable useful inference between disparate datasets using one or more of these standard terminologies to code disease. The Disease Ontology will be, at the end of this project, a community-driven, community-accepted ontology of diseases for clinical research and medicine inclusive of genetic, environmental and infectious diseases. The Disease Ontology will encapsulate, therefore, a comprehensive theory of disease. The design of the disease ontology will enable greater understanding of disease states by placing heritable disorders in the context of other infectious diseases and related diseases. The structure of Disease Ontology and the external references to other terminologies will enable the integration of disparate datasets through the concept of disease.

DO Subversion Repository

SOURCEFORGE.NET

Communit

[SF.net](#) / [Projects](#) / [SCM Repositories](#) / [diseaseontology](#) / [trunk](#) / [HumanDO.obo](#)

SCM Repositories - [diseaseontology](#)

[Parent Directory](#)

Links to HEAD:	(view) (download) (annotate)
Sticky Revision:	<input type="text"/> <input type="button" value="Set"/>

Revision [31](#) - [\(view\)](#) [\(download\)](#) [\(annotate\)](#) - [\[select for diffs\]](#)
Modified *Fri Jan 16 20:26:32 2009 UTC* (12 days ago) by *lynn_schriml*
File length: 15562247 byte(s)
Diff to [previous 30](#)

updating term names

Revision [30](#) - [\(view\)](#) [\(download\)](#) [\(annotate\)](#) - [\[select for diffs\]](#)
Modified *Wed Dec 10 17:56:30 2008 UTC* (7 weeks ago) by *lynn_schriml*
File length: 15562260 byte(s)
Diff to [previous 29](#)

housekeeping

Revision [29](#) - [\(view\)](#) [\(download\)](#) [\(annotate\)](#) - [\[select for diffs\]](#)
Modified *Mon Dec 8 21:25:50 2008 UTC* (7 weeks, 1 day ago) by *lynn_schriml*
File length: 15562606 byte(s)
Diff to [previous 28](#)

adding definitions to syndromes

<http://svn.sourceforge.net/viewvc/diseaseontology/trunk/HumanDO.obo?view=log>

Disease Ontology Tracker

SF.net » Projects » Phenotype and Disease Ontologies » Tracker » Feature Requests » Browse Tracker Items

Phenotype and Disease Ontologies

Trackers Search Advanced

Summary Tracker Mailing Lists Forums Code Services Download Documentation Tasks

Add new artifact Browse

Stats RSS

Assignee: (?) Status: (?) Category: (?) Group: (?)
Any Open Any Any
Show only: Submitter username: Summary keyword:
Sort By: (?) ID Descending Browse

Request ID	Summary	Open Date	Priority	Assigned To	Submitted By
2529743	yCTUzFuTWjtlCVH	2009-01-22 23:49	5	nobody	nobody
2498333	umANTwNE	2009-01-10 21:35	5	nobody	nobody
2351612	DO to HP alignment	* 2008-11-26 17:39	5	nobody	cmungall
2351598	DO to FMA links	* 2008-11-26 17:32	5	nobody	cmungall
2351572	potential missing is_a links	* 2008-11-26 17:26	5	nobody	cmungall
2314492	name and exact synonym clashes	* 2008-11-19 10:03	5	nobody	cmungall
2215117	NclpwLBzThdjXqom	* 2008-11-02 01:29	5	nobody	nobody
2185374	deal with "Recruitment"	* 2008-10-21 20:40	5	nobody	nobody
2156653	rtWONdcHeh	* 2008-10-10 06:27	5	nobody	nobody
2144571	Obsolete - disease of body substance - DOID:9	* 2008-10-03 15:50	5	nobody	wakibbe
2144568	Obsolete - secretion disease- DOID:41	* 2008-10-03 15:49	5	nobody	wakibbe
2144553	Obsolete: blood disease-DOID:40	* 2008-10-03 15:42	5	nobody	wakibbe
2074404	DO Slim	* 2008-08-25 21:24	5	nobody	lynn_schriml
2074400	Liver Cancer SubTypes	* 2008-08-25 21:23	5	nobody	lynn_schriml
2074397	Colorectal Cancer	* 2008-08-25 21:22	5	nobody	lynn_schriml

* Denotes Requests 45 Days Old

http://sourceforge.net/tracker/?group_id=79168&atid=555739

Ontology Viewer at EBI

OLS - Ontology Lookup Service

Enter Ontology Term

Search Ontology:

Term Name: (Include obsolete terms)

Term ID:

Additional Information:

preferred name	Diabetes Insipidus
exact synonym	Diabetes insipidus
exact synonym	diabetes insipidus
exact synonym	Diabetes insipidus (disorder)
exact synonym	Diabetes insipidus
xref_related_synonym	SNOMEDCT_2005_07_31:15771004
xref_related_synonym	CSP2005:1849-2602
xref_related_synonym	SNOMEDCT_2005_07_31:190484000
xref_related_synonym	ICD9CM_2006:253.5
xref_analog	ICD9CM_1987:253.5
xref_analog	ICD9CM_2005:253.5
xref_analog	UMLS_SNOMEDCT_2005_01_31_AUI:A29286
xref_analog	UMLS_ICD9CM_2005_AUI:A0406458
xref_analog	ICD9CM_2004:253.5

Statistics

Version: 1.15
Ontologies Loaded: 61
Terms Loaded: 771943
Last updated: Sun Nov 30 07:05:12 GMT 2008

See the full breakdown of loaded ontologies [here](#) and load statistics [here](#).

OLS - Ontology Lookup Service

DOID Ontology Browser

- temp holding
- disease
 - disease of infectious agent
 - disease of behavior
 - syndrome
 - disease of environmental origin
 - disease of biological process
 - disease of anatomical entity
 - disease of physical anatomical entity
 - disease of material anatomical entity
 - disease of anatomical structure
 - acellular anatomical structure disease
 - internal elastic lamina
 - disease of body
 - disease of organ system
 - organ disease
 - tissue disease
 - biological macromolecule
 - disease of anatomical set
 - disease of set of heterogenous clusters
 - endocrine system disease
 - immune system disease
 - disease of subdivision of hemolymphoid system

Help [\(hide\)](#)

Double-click a term to see its children. The ontology browser is populated dynamically. If there are many children for a given term, there may be a small delay while the browser fetches. **Click** to highlight a term to see any information associated with it. **Hover** over a term to see its relation with its immediate parent. Root terms will not display any relational information.

Relations

disease of environmental origin is_a disease

Term Information

ID: [DOID:7](#)

[Zoom](#)

Name: disease of anatomical entity

Associated information

definition	A disease that manifests in a defined anatomical structure.
xref_definition	URL: http://www2.merriam-webster.com/cgi-bin/mwmednlm?book=Medical&va=anatomic
xref_definition	DO:wk,ls



Disease Ontology Version 3

- Is an OBO Foundry ontology for the Integration of Biomedical Data
- Is inclusive of genetic, environmental and infectious diseases
- Is semantically organized and computable
- Path to the top is mostly true
- Has 12,564 terms and 21,024 branches
- Has a maximum depth of 13 and is 'node heavy in the middle', meaning many nodes are in the 6-10 node deep range



Medical Vocabularies in DO

- UMLS Metathesaurus and Semantic Network
 - 5 million concepts and a million terms
- MeSH (Medical Subject Headings)
 - Shallow graph with no direct disease mapping
- NCI Thesaurus (National Cancer Institute)
 - Broad coverage, not deep outside cancer domain
 - No direct mapping to ICD9
- SNOMED (Systemized Nomenclature of Medicine)
 - Large, broad but duplicate concepts in different contexts
 - Restrictive, only free for research in the US
- ICD9/ICD10 (International Classification of Disease)
 - Poor coverage, few high level terms, confused terms

DO Mappings

External Reference	Unique xref:DOID Mappings	Unique xrefs
ICD-9	186278	10109
UMLS_SNOM DCT_2005_01_31_AUI	38912	38912
UMLS_NCI2004_11_17_AUI	24049	24049
UMLS_MSH2005_2005_01_17_AUI	21377	21377
UMLS_CUI	17023	17023
UMLS_ST	14674	14674
SNOMEDCT_2005_01_31	13116	13116
UMLS_ICD-9	10048	10048
NCI2004_11_17	6991	6991
UMLS_MTHICD-9_2005_AUI	3611	3611
MSH2005_2005_01_17	3502	3502
UMLS_CSP2004_AUI	2269	2269

Unique mappings in DO to ICD-9, SNOMED CT, NCI metathesaurus (EVS), MESH, UMLS and MESH terms. The compound reference names, such as UMLS_SNOMEDCT_2005_01_31_AUI show the release of UMLS used to perform the SNOMED CT mappings. All existing mappings are to exact or closest concept match between a DO term and the external reference source.

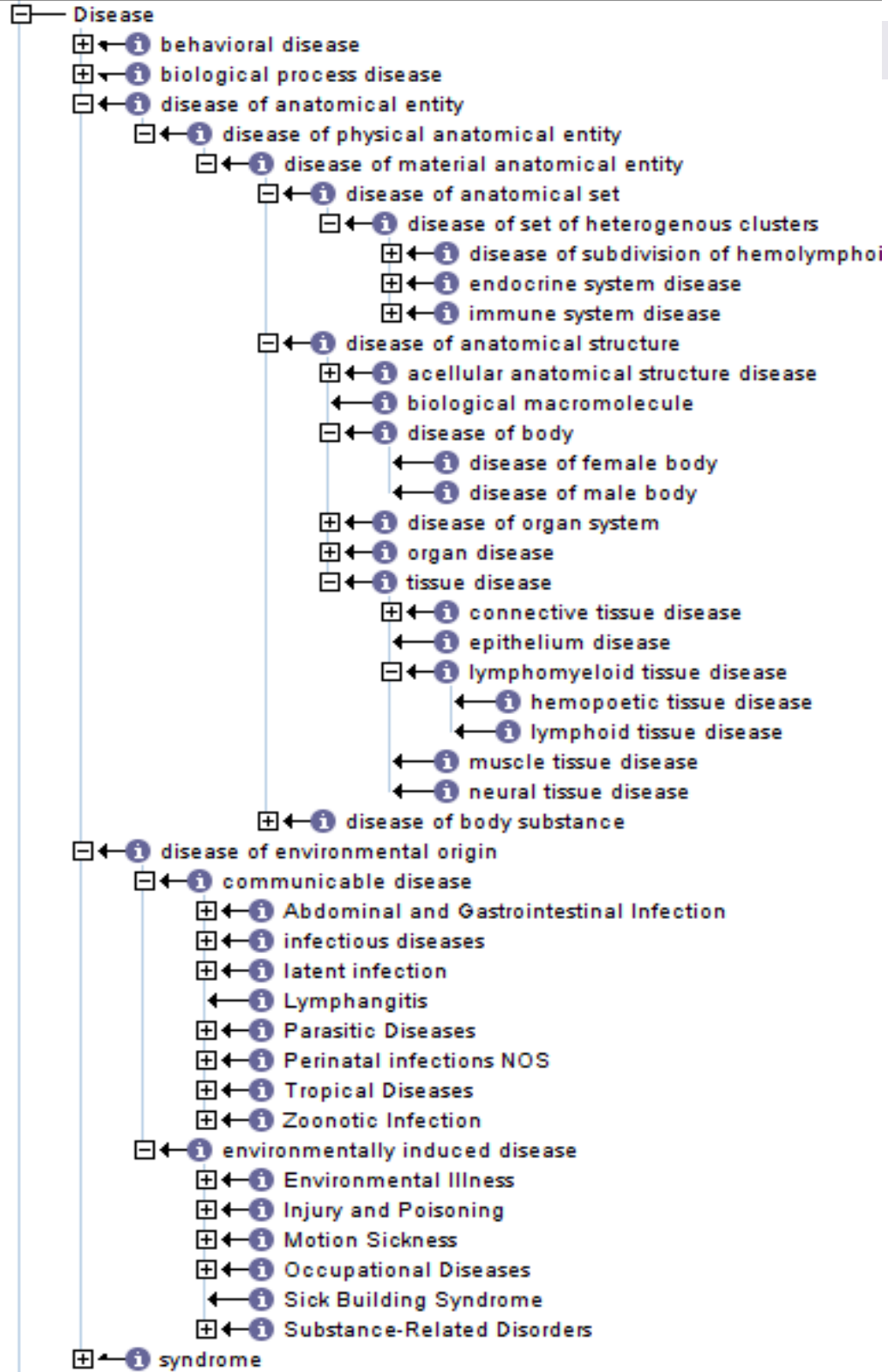
DO V3 To Dos

DO Improvements:

- Ontological structure:
 - Still needs a lot of help
 - Views vs internal structure needs to be better defined
different communities need to view DO from very different perspectives and DO needs to easily support each viewpoint

Three main hurdles which face DO development:

- 1) Proper ontological structuring of DO
- 2) Mapping between source vocabulary terms for all DO terms
- 3) Providing tools and resources for community input of DO and timely updates of DO



Cross references to Anatomical Ontology (FMA)

Original driving Use Case



NUgene collects and stores genetic (DNA) samples along with associated healthcare information from patients of Northwestern-affiliated hospitals and clinics. It is currently the only study of its kind in Chicago and one of a few in the nation. This resource is available to scientists to conduct groundbreaking genetic research.



- The NUgene Project is a genetic banking study which collects and stores DNA samples and associated medical information from its participants
 - Medical information consists of Epic data (mostly free text) and billing/procedure data
- Problem:
 - What types of diseases do our participants have?

Past and Current Applications

- The primary driver for the creation of the Disease Ontology is the ability to integrate disparate datasets that contain disease concepts or concepts that can be mapped to disease.
- The Disease Ontology provides a unifying structure to map disease knowledge between datasets such as patient records and large scale genome, sequencing and microbiome projects.

GeneRIF and NUGene studies: DO fulfills that role in an unbiased and granular fashion by providing the key component in the arsenal of tools by providing computable relationships between disease and concepts that can be mapped to disease such as genetic associations to disease, symptoms and biological process.

Gemina project: DO has been utilized to annotate incidents of infectious pathogens and to provide a query and retrieval vocabulary linking disease to hosts and transmissions and outbreaks of disease. (<http://gemina.igs.umaryland.edu>)

eMERGE Consortium's electronic medical records (EMR) will be validated by the Disease Ontology so that we can more easily map participants to specific disease cohorts and map data coming from each EMR system to common standards.

FunDO, CASIMIR, VPH?

Example Application - DO Browser

Disease Ontology

- Communicable Diseases [1728 patients, 2619 terms]
- Disorders of Environmental Origin [1398 patients, 1419 terms]
- Stomatognathic Diseases [339 patients, 283 terms]
- Syndrome [508 patients, 156 terms]
- Mental and behavioral problems [989 patients, 871 terms]
- Neoplasms [1468 patients, 681 terms]
- Hyperplasia [130 patients, 27 terms]
- Hemic and Lymphatic Diseases [1388 patients, 437 terms]
- Otorhinolaryngologic Diseases [2039 patients, 2043 terms]
- Skin and Connective Tissue Diseases [2213 patients, 3069 terms]
- Degenerative Disease [949 patients, 1192 terms]
- Disorder by Site [2480 patients, 8716 terms]
- Hereditary Diseases [458 patients, 113 terms]
- Digestive System Disorders [1632 patients, 1159 terms]
- Immunodeficiency and Immunosuppression Disorders [1284 patients, 422 terms]
- Deformity [758 patients, 629 terms]
- Lifestyle-related condition [627 patients, 375 terms]
- Organic brain syndrome [31 patients, 41 terms]
- Socialized Conduct Disorder [519 patients, 259 terms]
 - Socialized conduct disorder, mild degree [0 patients, 1 terms]
 - Socialized conduct disorder, severe degree [0 patients, 1 terms]
 - Undersocialized Conduct Disorder, Aggressive Type [0 patients, 5 terms]
 - Phobic anxiety disorder [5 patients, 6 terms]
 - Socialized conduct disorder, moderate degree [0 patients, 1 terms]
- Impulse Control Disorders [0 patients, 5 terms]
- Panic Disorder [17 patients, 2 terms]
- Communication impairment [506 patients, 237 terms]
 - Hearing problem [158 patients, 38 terms]
 - Vision Disorders [414 patients, 196 terms]
 - Language Disorders [1 patients, 2 terms]
 - Learning Disorders [2 patients, 2 terms]
- Dependence [24 patients, 67 terms]
- Substance Withdrawal Syndrome [59 patients, 4 terms]
- Tobacco Use Disorder [90 patients, 5 terms]

ICD-9 Term(s) to Find:

Terms ANDED

34882: Hearing problem

27634: Vision Disorders

OR

Terms ANDED

BUT NOT

Terms Excluded

ICD-9 Codes (233)	ICD-9 Codes (0)	ICD-9 Codes (0)	Unique Patients 67
036.81 094.84 250.80 253.5 264.5 360.21 362.85 363.05 367			Unique Samples

Save Query

Name for Query:

Project Name:

Category:

Comments:



GeneRIF mining using DO

Using MMTx to mine GeneRIFs with DO, UMLS or MeSH

DO, UMLS and MeSH result in low false positive,
DO has a lower false negative rate

Using MMTx to mine OMIM with DO, UMLS or MeSH

Complex full sentences with compound ideas are hard to parse with standard text mining techniques – Mining OMIM did not perform as well as mining GeneRIFs

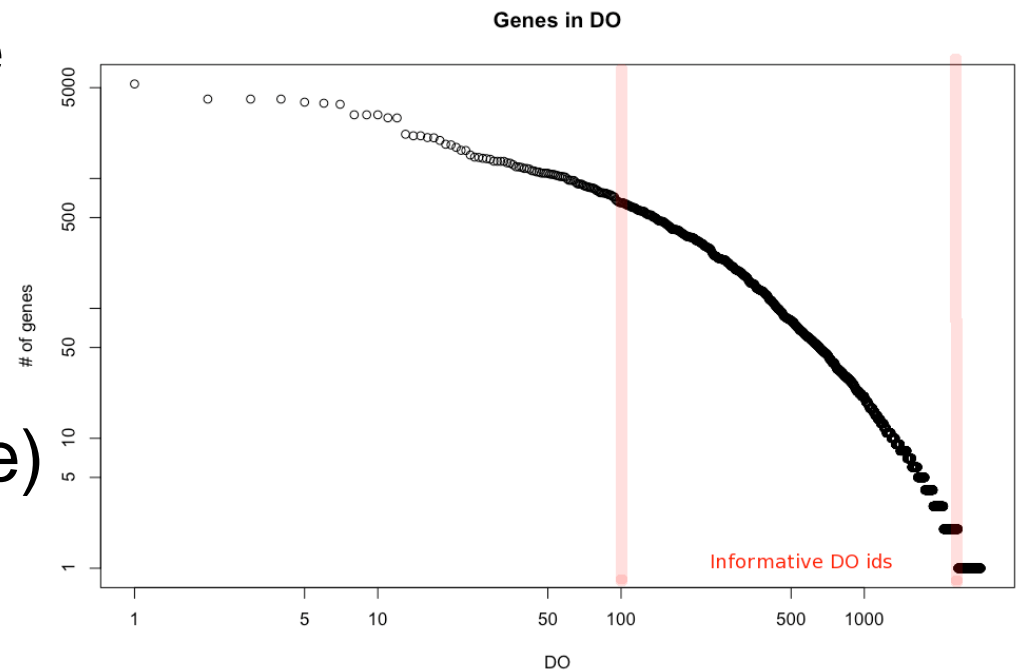
Releasing geneRIF mappings

- Goal: Build a standard Bioconductor package to retrieve and analyze information based on Disease Ontology for a given genelist.
 - Information retrieval between genes and DO
 - Analysis of DO category for a given genelist
 - Hypergeometric Mean Visualization

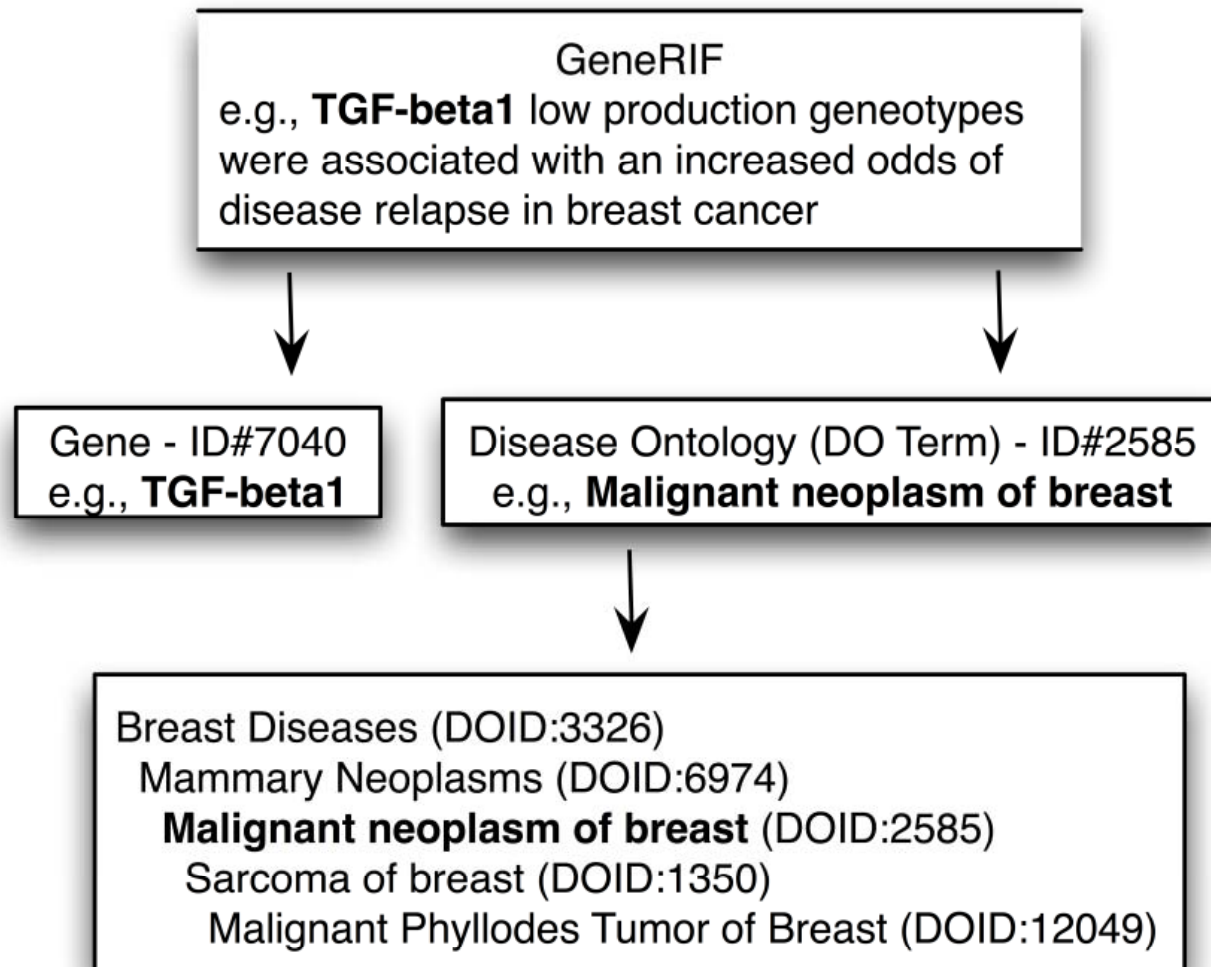
$$f(k; N, m, n) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}.$$

Informative DO ids

- Remove top-node, non-specific DO categories
 - Sort DO ids based on the amount of genes in each category
 - Remove top 100 DO ids and last 1237 DO ids (containing only one gene)



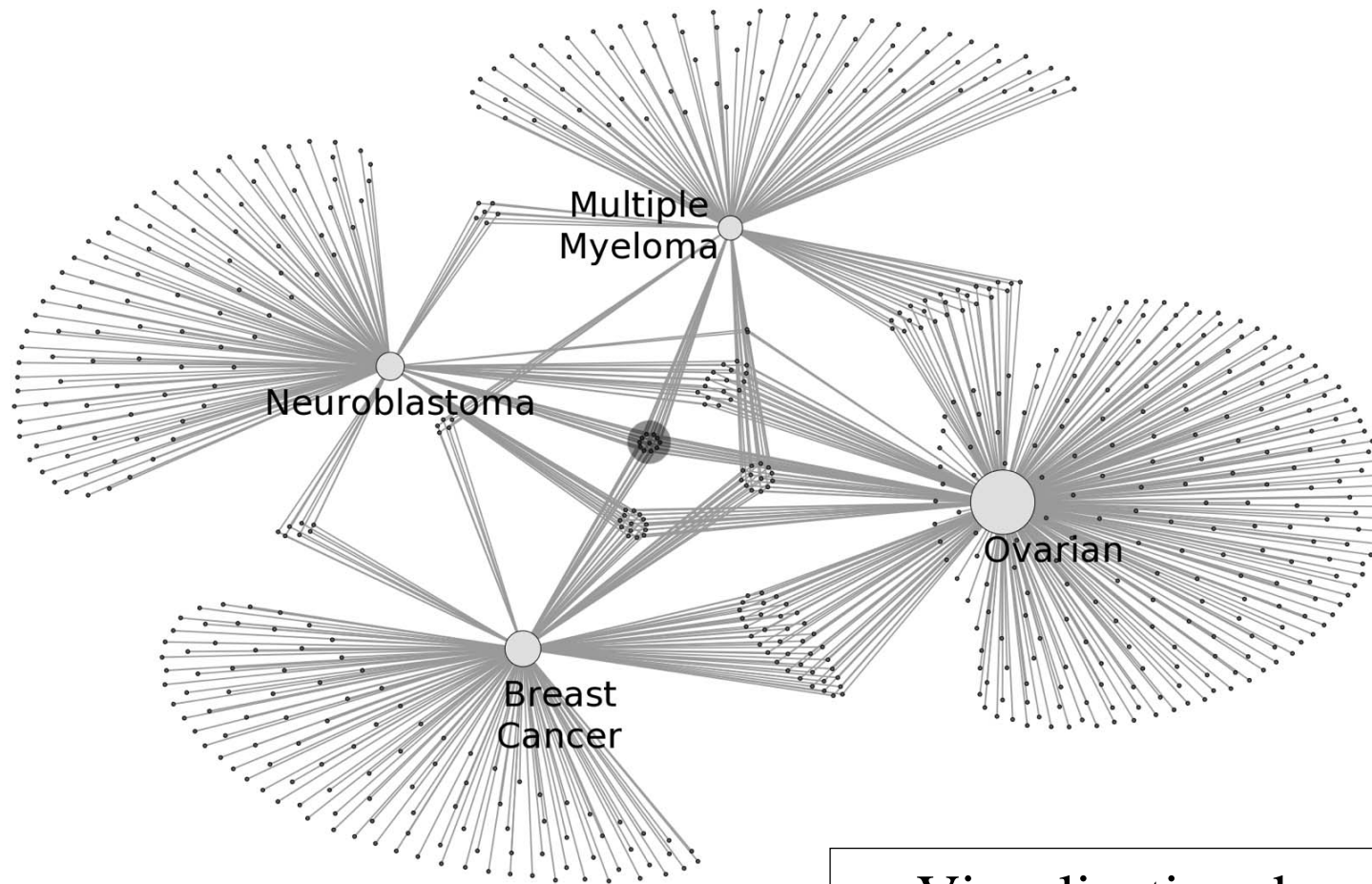
Annotating the human genome with DO



Estimation of recall and precision of disease annotation

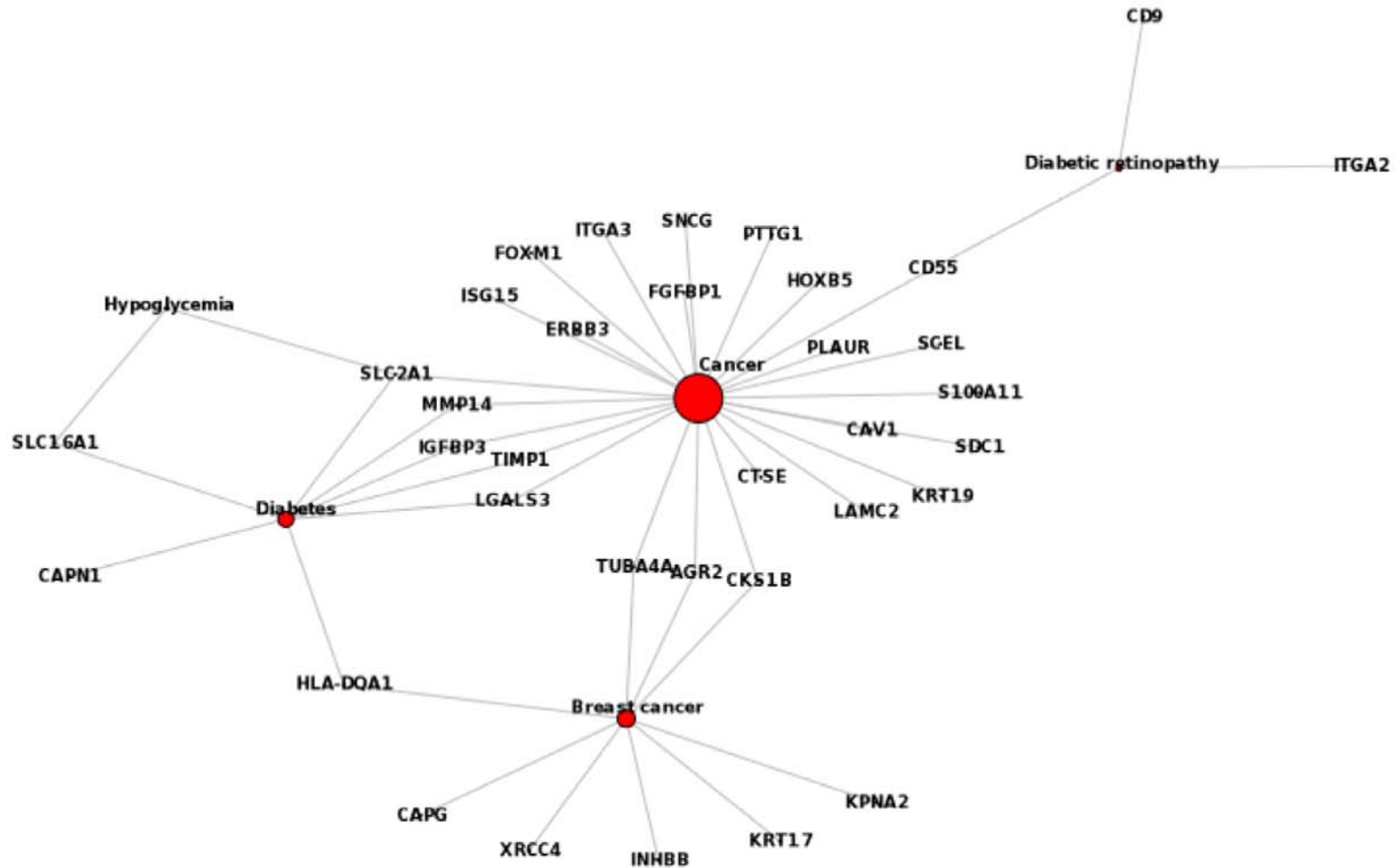
	OMIM	GeneRIF
Recall	21.85	90.76
Precision	98.46	96.66

Disease-gene network



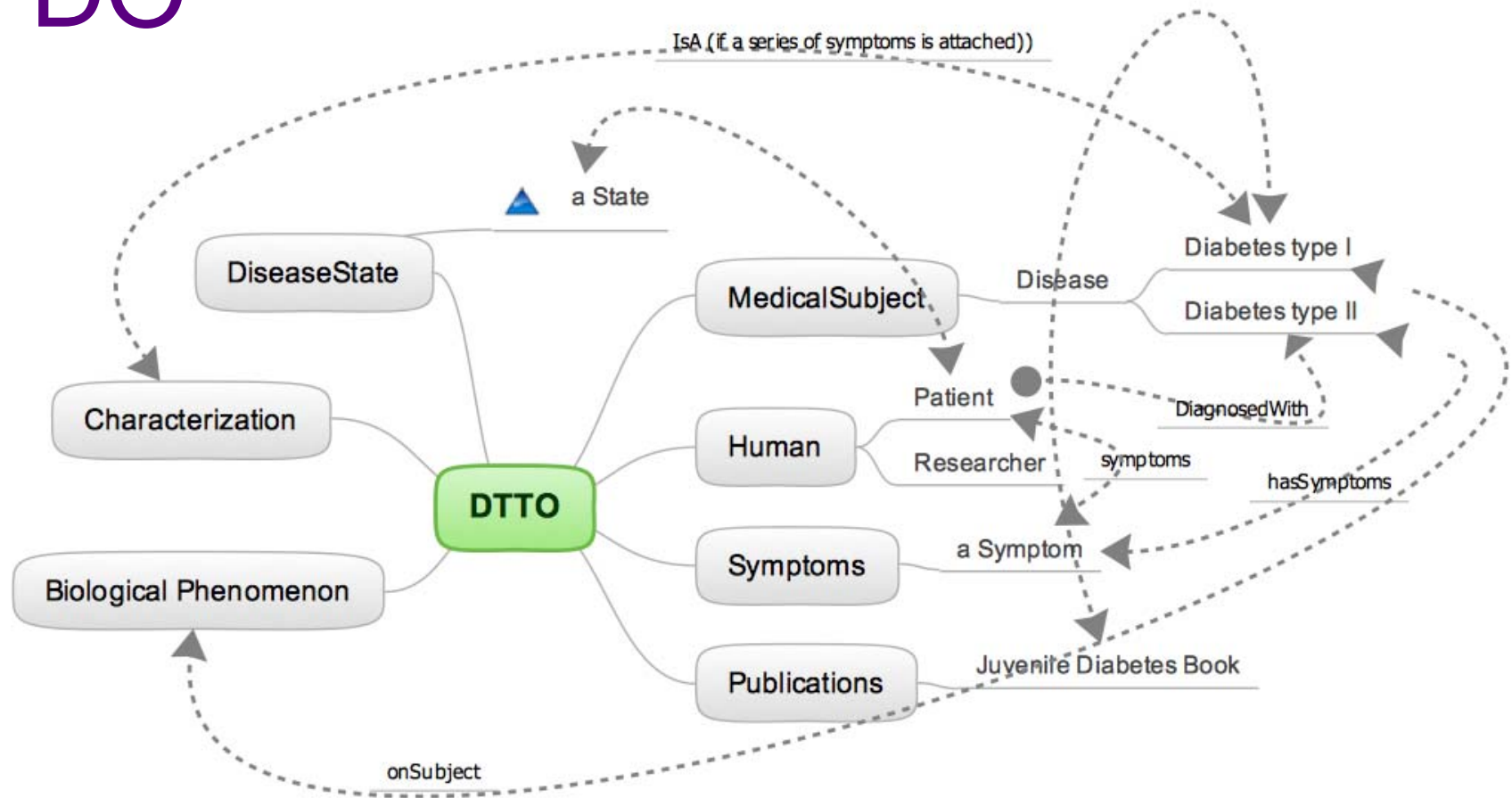
Visualization done using
Cytoscape by Gilbert Feng

FunDO



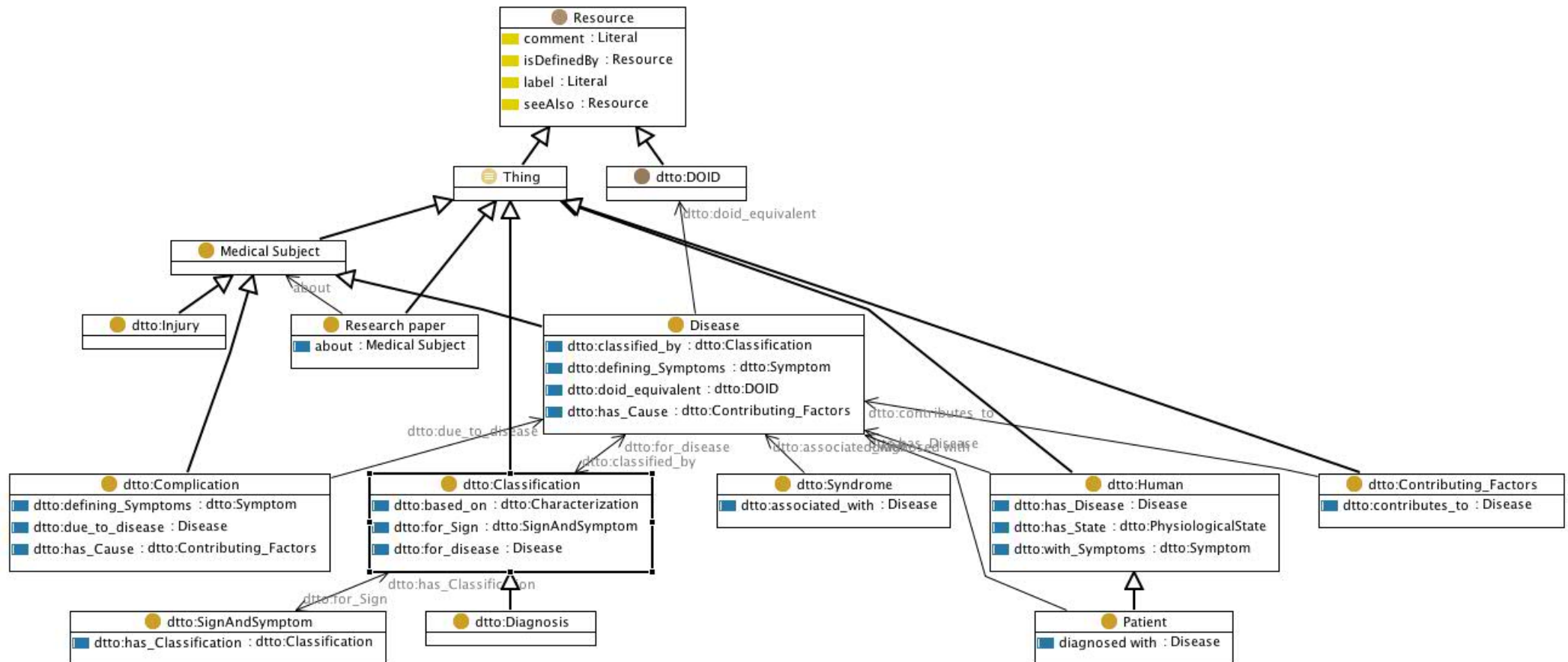
<http://projects.bioinformatics.northwestern.edu/fundo>

Diabetes Type 2 Ontology (DTTO) and DO



Eric Neumann, Pfizer

DTTO and DO



Eric Neumann, Pfizer

DO and PATO

- A critical aspect in applying DO to medical literature and medical records is that ability to walk between "Signs and Symptoms" attached to a patient record or the medical literature and the disease
- PATO will describe the phenotype, possibly with a 'Signs and Symptoms' view, and DO will describe the disease concept(s) linked with those signs and symptoms
- When we map from ICD-9 to DO, for instance, we find that collections of signs and diagnoses are associated with disease, and that it is the collection, rather than a single association, that enables the inference of disease from a set of observables



DO and model systems for disease

- DO is human-centric
- Is DO sufficient to link disease concepts between organisms?
- What happens when the underlying mechanisms appear to have changed between the organisms?
 - Epilepsy in Dogs and Human
 - Viral Diseases
 - Rat and Mouse models and Cancer in Humans

How to get involved in DO

Ontology Tools	URLs
Wiki	http://do-wiki.nubic.northwestern.edu/
DO Listserv	https://lists.sourceforge.net/lists/admin/db/diseaseontology-discussion
DO Sourceforge Tracker for term submission/definitions	http://diseaseontology.sourceforge.net/#projects
Subversion	http://svn.sourceforge.net/viewvc/diseaseontology/trunk/HumanDO.obo?view=log

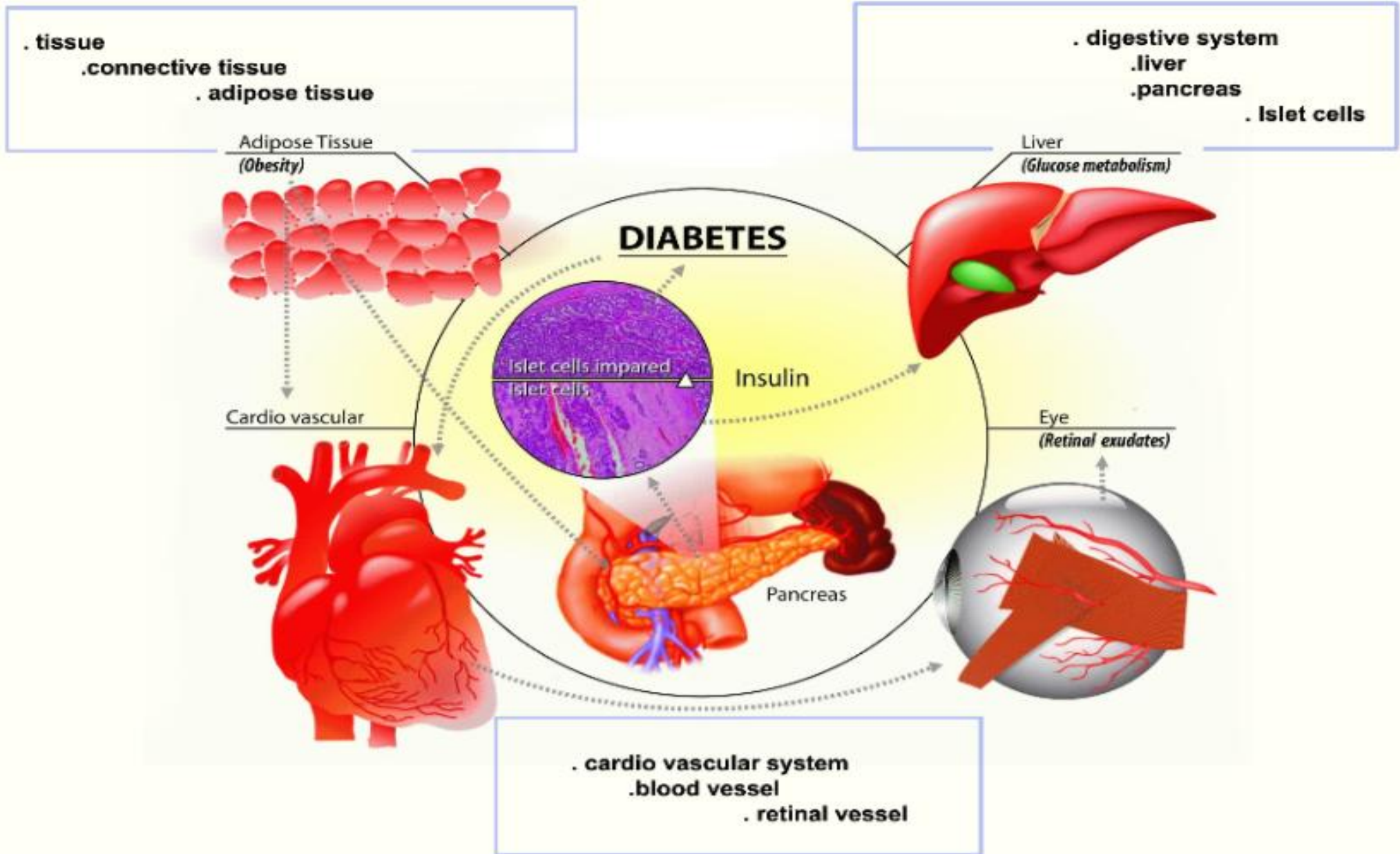


DO Community Engagement

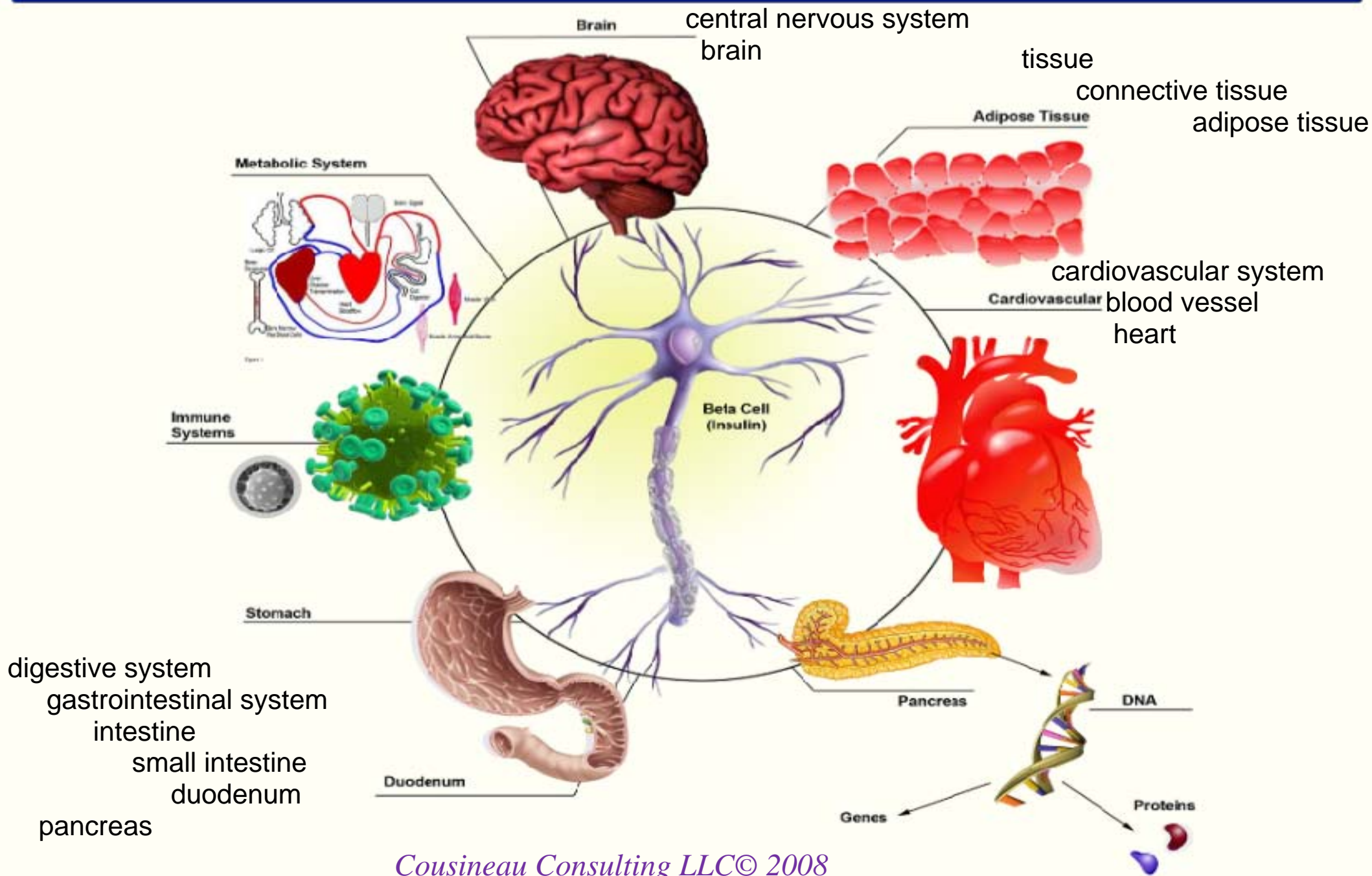
How DO is Working Toward Engaging the Community:

- This meeting
- More active role in OBO Foundry: attending OBO Foundry meeting in July, 2008
- Participating in other related Ontology efforts: Infectious DO, Human Phenotype Ontology and Symptom Ontology
- Medical Ontology Efforts (Medical Knowledge Research Group) Steve Roessingh, Leo Cousineau, Robert Baud
- Engaging the Neurogenomics community through Maryann Martone and BIRN
- CASIMIR mouse community (Paul Schofield), MGI (Janan Eppig, Judy Blake), eMERGE (NCRR), caBIG[®]

Medical Ontology : Relationships between diseases, disorders, & systems, organs and tissues



Biomedical Ontology : Neuronal interaction between diseases, systems, organs, substances, tissues, cells, proteins and genetics





Invitation for participation

For Disease Ontology to succeed, it needs buy-in from the community. We hope you will join us in fixing and extending Disease Ontology to meet your needs.

Acknowledgements

Steering Committee	Contributors to this talk
Rex Chisholm, DO Founder, NU	John Osborne
Michael Ashburner, EBI	Simon Lin
Barry Smith,	Jared Flatow
Richard Scheuerman, UT Southwestern	Julie Zhu
Suzi Lewis, UC Berkeley	Gilbert Feng
Maryann Martone, UC San Diego	Pan Du
Chris Mungall, Lawrence Berkeley Lab	Wendy Wolf
	Abel Kho
	Steve Roessingh
	Dong Fu
	Eric Neumann