
Gene Ontology of CNS and Disease Ontology

Giorgio Valle
CRIBI Biotech Centre
University of Padua

Acknowledgements



Padua University



Erika Feltrin



Alessandro Albiero

GlaxoSmithKline



Fabrizio Caldara



GO @ EBI

Jennifer Deegan
(nee Clark)



Midori Harris

CNS GO editing

- Books
- Journals
- Encyclopedia, Vocabularies
- Several Ontologies and Databases
- Discussion with experts**

SF 1262241 neurogenesis

8 months (08/2005-03/2006)
36 pages of discussion
6615 words

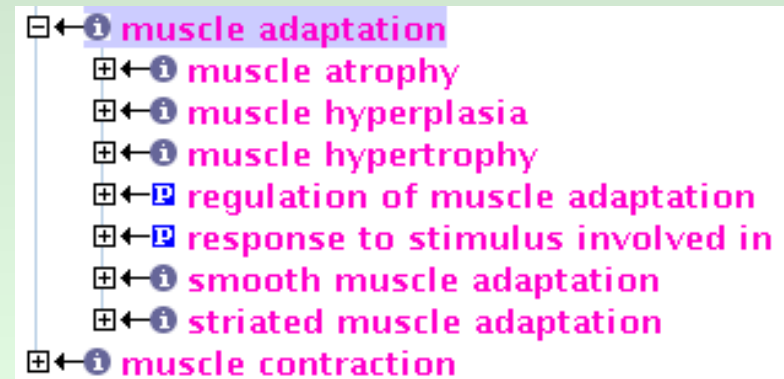
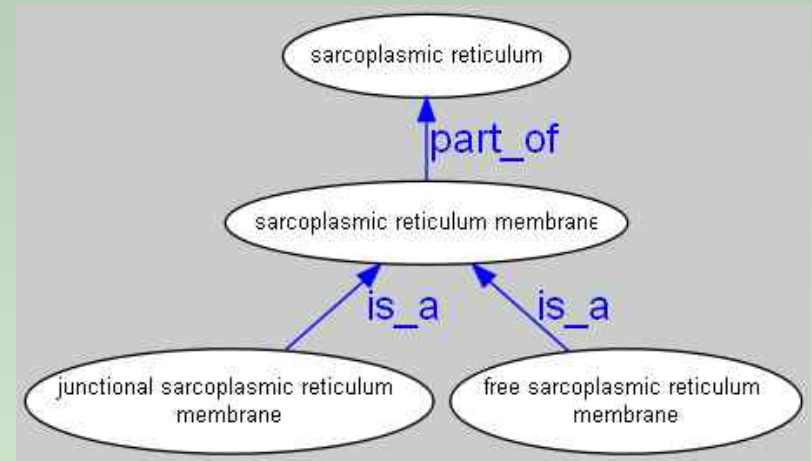
Neurobiology Content Meeting

(organized by David Hill) with experts
of CNS biology **500 terms**

- ▣ ← ⓘ system development
 - ⊕ ← ⓘ endocrine system development
 - ⊕ ← ⓘ exocrine system development
 - ▣ ← ⓘ nervous system development
 - ← ⓘ branching morphogenesis of a nerve
 - ⊕ ← ⓘ central nervous system development
 - ← ⓘ ganglion mother cell fate determination
 - ⊕ ← ⓘ gliogenesis
 - ▣ ← ⓘ neural tube formation
 - ← ⓘ neural crest formation
 - ▣ ← ⓘ primary neural tube formation
 - ← ⓘ neural fold formation
 - ← ⓘ neural plate formation
 - ← ⓘ neural plate shaping
 - ← ⓘ neural rod cavitation
 - ⊕ ← ⓘ neural rod formation
 - ← ⓘ neural tube closure
 - ⊕ ← ⓘ secondary neural tube formation
 - ▣ ← ⓘ neurogenesis
 - ⊕ ← ⓘ neuroblast activation
 - ▣ ← ⓘ neuroblast differentiation
 - ← ⓘ neuroblast development
 - ▣ ← ⓘ neuroblast fate commitment
 - ⊕ ← ⓘ neuroblast fate determination
 - ← ⓘ neuroblast fate specification

Muscle Content Meeting

- Focused mainly on **skeletal** and **smooth** muscles
- Muscle experts
 - from Padua University
 - from Europe
- Improvement in BP and CC ontologies
 - muscle regeneration
 - muscle plasticity
 - muscle development
 - muscle contraction
 - sarcomere and membrane-bounded organelles
- Cross-references between ontologies



Muscle Content Meeting

- *Large-scale additions* to the BP and CC ontologies
 - **156** new terms added
 - **57** existing terms redefined and/or renamed
- A valuable resource for *annotation of gene products* related to muscle biology and involved in neuromuscular disease
- *The interpretation of high-throughput data* in the area of muscle science and muscle disease

[Feltrin *et al.* submitted to BMC Medical Genomics]

Muscle Biology Community Annotation Wiki

172 genes associated with **muscle biological process** and **muscle disease**

- to annotate gene products to the new muscle terms
- to review the existing GO annotations for any gene of interest
- to provide information about any aspect of the biology of a gene from any species
- to facilitate community participation in the Gene Ontology project and speed up the annotation process



**TRAIT (TRAnscript Integrated Table):
 a knowledgebase of human skeletal muscle
 transcripts**

Stefano Toppo*, Nicola Cannata, Paolo Fontana,
 Chiara Romualdi, Paolo Laveder, Emanuela Bertocco,
 Gerolamo Lanfranchi and Giorgio Valle*

Cellular component

- mitochondrial membrane
- mitochondrial outer membrane
- mitochondrial outer membrane translocase complex
- mitochondrial ribosome
- mitochondrion**

Chromosome /map

- | | | |
|---------------------------------------|---------------------------------------|---------------------------------------|
| <input checked="" type="checkbox"/> 1 | <input checked="" type="checkbox"/> 2 | <input checked="" type="checkbox"/> 3 |
| 1p12 | 2cen-q13 | 3p11-q11 |
| 1p13 | 2cen-q24 | 3p12-q12 |
| 1p13-q23 | 2p11.1 | 3p13-q13.33 |
| 1p13.1 | 2p11.1-q11.1 | 3p14 |
| 1p13.1-q21.3 | 2p11.2 | 3p14.3 |

Protein domains

- AND OR
- Ribonuclease_BN (Ribonuclease BN-like family)
 - ribonuclease_T2 (Ribonuclease T2 family)
 - ribonuc_red_lg (Ribonucleotide reductase, all-alpha domain)
 - ribonuc_red_lgC (Ribonucleotide reductase, barrel domain)
 - Ribosomal_L1 (Ribosomal protein L1p/L10e family)**

Orthologs

- | | | | |
|-----------------|--------------------------------------|-------------|---------------------------|
| C. elegans | <=< <input type="radio"/> | e-10 | >=> <input type="radio"/> |
| | | e-30 | |
| | | e-50 | |
| D. melanogaster | <=< <input checked="" type="radio"/> | e-10 | >=> <input type="radio"/> |
| | | e-30 | |
| | | e-50 | |
| S. cerevisiae | <=< <input checked="" type="radio"/> | e-10 | >=> <input type="radio"/> |
| | | e-30 | |
| | | e-50 | |

features query results on TRAIT entries

ranking 4/5 - 40 records
 ranking 3/5 - 376 records
 ranking 2/5 - 1194 records
 ranking 1/5 - 1644 records

SHOW THE HITS

Please, select the results to show from the menu on the left and then click the button on the right

ranking 4/5 - 40 records

chr.	Sacc. cer.	Dro. mel.	cell. com.	domain	
					strait-6 C Homo sapiens succinate dehydrogenase complex, subunit B, iron sulfur (Ip) (SDHB), nuclear gene encoding mitochondria
					strait-8 C Homo sapiens ribosomal protein L7 (RPL7), mRNA, complete cds
					strait-114 C H. sapiens Uba80 mRNA for ubiquitin
					strait-123 C ribosomal protein S14
					strait-258 C Homo sapiens ribosomal protein S20 (RPS20), mRNA
					strait-394 C Homo sapiens ribosomal protein S18 (RPS18), mRNA
					strait-690 C Homo sapiens ribosomal protein, large, P0 (RPLP0), mRNA
					strait-779 C Homo sapiens ribosomal protein S7 (RPS7), mRNA
					strait-790 C HSU59309 Human fumarase precursor (FH) mRNA, nuclear gene encoding mitochondrial protein, complete cds
					strait-812 C Homo sapiens hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase
					strait-903 C Homo sapiens ribosomal protein S26 (RPS26), mRNA
					strait-997 C Homo sapiens ribosomal protein S16 (RPS16), mRNA
					strait-1016 C Homo sapiens ATP synthase, H+ transporting, mitochondrial F0 complex, subunit b, isoform 1 (ATP5F1), mRNA
					strait-1041 C Homo sapiens cytochrome P450, subfamily XXVIIA (steroid 27-hydroxylase, cerebrotendinous xanthomatosis), polypep
					strait-1166 C Homo sapiens ribosomal protein L7a (RPL7A), mRNA
					strait-1404 C Homo sapiens ribosomal protein S15 (RPS15), mRNA
					strait-1444 C Homo sapiens voltage-dependent anion channel 1 (VDAC1), mRNA
					strait-1457 C Homo sapiens ribosomal protein S8 (RPS8), mRNA
					strait-1538 C Homo sapiens ribosomal protein L18a (RPL18A), mRNA
					strait-1545 C Homo sapiens ribosomal protein S13 (RPS13), mRNA
					strait-1573 C Homo sapiens ribosomal protein L32 (RPL32), mRNA
					strait-1661 C Homo sapiens ribosomal protein L35a (RPL35A), mRNA
					strait-1665 C Homo sapiens solute carrier family 25 (mitochondrial carrier, Aralar), member 12 (SLC25A12), mRNA
					strait-1666 C Homo sapiens ribosomal protein L11 (RPL11), mRNA
					strait-1715 C Homo sapiens laminin receptor 1 (67kD, ribosomal protein SA) (LAMR1), mRNA

Detecting differentially expressed genes

Comparison of tissues expression levels

This tool allows the comparison of gene expression levels and the identification of genes whose expression levels within selected tissues (selected from the numerator sample) are greater/lower n times (threshold) than other tissues (selected from denominator sample).

Please select tissues from the numerator and denominator set, select the threshold and press ENTER.

Numerator

- Brain
- Testis
- Uterus
- Placenta
- Thymus
- Liver
- Aorta
- Heart
- Adult Skeletal Muscle 21 years
- Adult Skeletal Muscle 40 years

Threshold

4

ENTER

RESET

Display only overexpressed genes

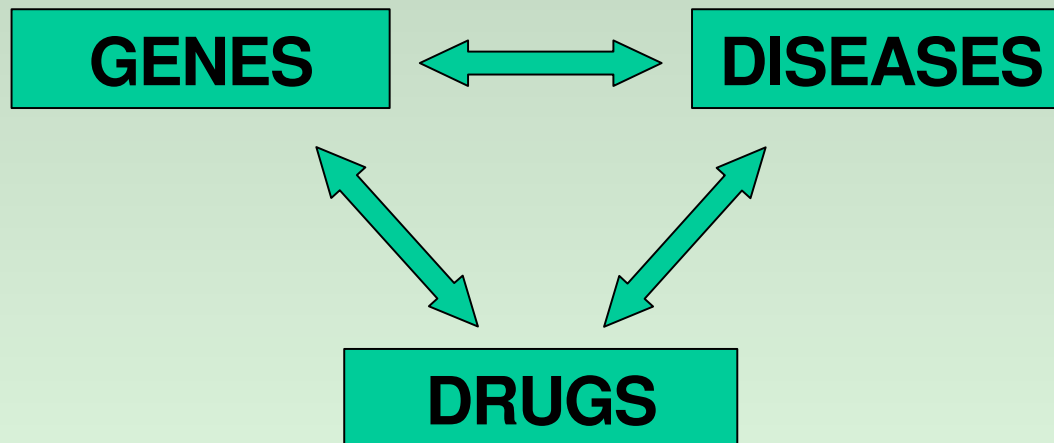
Denominator

- Brain
- Testis
- Uterus
- Placenta
- Thymus
- Liver
- Aorta
- Heart
- Adult Skeletal Muscle 21 years
- Adult Skeletal Muscle 40 years

Gene differentially expressed at 4-fold level

Strait	Locus	Link	Level	Description
1627			13.38	PROVISIONAL: not matching CDS
4749			13.38	Anonymous sequence
784			4.11	PROVISIONAL: not matching CDS
2230			7.82	Anonymous sequence
3126			5.00	Anonymous sequence
1271	26548		5.54	Homo sapiens integrin beta 1 binding protein (melusin) 2 (ITGB1BP2), mRNA
830	1605		4.22	Homo sapiens dystroglycan 1 (dystrophin-associated glycoprotein 1) (DAG1), mRNA
770	1831		5.50	AF153603 Homo sapiens TSC-22 related protein (TSC-22R) mRNA, complete cds
37	4625		4.36	Homo sapiens myosin, heavy polypeptide 7, cardiac muscle, beta (MYH7), mRNA
38	4620		4.36	Homo sapiens myosin, heavy polypeptide 2, skeletal muscle, adult (MYH2), mRNA
39	4619		4.36	Homo sapiens myosin, heavy polypeptide 1, skeletal muscle, adult (MYH1), mRNA
1309	4626		4.36	Homo sapiens myosin, heavy polypeptide 8, skeletal muscle, perinatal (MYH8), mRNA
2415	57644		4.36	AB040945 Homo sapiens mRNA for KIAA1512 protein, partial cds
2951			21.50	Anonymous sequence
999	5204		14.50	Homo sapiens prefoldin 5 (PFDN5), mRNA
594	2170		6.00	fatty acid binding protein 3, muscle and heart (mammary-derived growth inhibitor)
834	6193		9.00	Homo sapiens ribosomal protein S5 (RPS5), mRNA
4527			9.00	Anonymous sequence
988			8.67	PROVISIONAL: not matching CDS
941	1622		6.33	diazepam binding inhibitor (GABA receptor modulator, acyl-Coenzyme A binding protein)
2863			12.70	Anonymous sequence
578	5224		5.00	phosphoglycerate mutase 2 (muscle)
4571			5.00	Anonymous sequence
2006	10345		4.50	Homo sapiens triadin (TRDN), mRNA
4786			4.50	Anonymous sequence
2708			12.12	Anonymous sequence
544			7.32	PROVISIONAL: not matching CDS
2208			4.08	Anonymous sequence
277	84265		88.00	Homo sapiens hypothetical protein MGC3200 (MGC3200), mRNA
308			36.50	PROVISIONAL: not matching CDS
933			9.28	Homo sapiens chromosome 19, BAC 41195 (CIT-B-31c15)
856	23197		4.92	Homo sapiens mRNA for KIAA0887 protein, partial cds
1554	23052		186.01	AB020637 Homo sapiens mRNA for KIAA0830 protein, partial cds
1609	8409		6.47	Homo sapiens ubiquitously-expressed transcript (UXT), mRNA

**To include the information about diseases
and drugs would be extremely interesting !**



Diseases

Disease Ontology is an excellent starting point

...

But if we want to link diseases to genes we need to use also other sources such as OMIM and Genetic Association database (GAD)

...

And what we found is that each resource uses different synonyms for diseases, and sometimes also for genes.

Disease Ontology

The Disease Ontology (DO) is a controlled medical vocabulary, modelled on GO, developed at the Bioinformatics Core Facility in collaboration with the NuGene Project at the Center for Genetic Medicine (Chicago, US). It was designed to facilitate the mapping of diseases and associated conditions to particular medical codes such as ICD9CM, SNOMED and others. Disease Ontology is implemented as a directed acyclic graph (DAG)

OMIM

The Online Mendelian Inheritance in Man (OMIM) is a comprehensive, authoritative and regularly updated knowledgebase of human genes and genetic disorders compiled to support human genetics research and education and the practice of clinical genetics. OMIM data are organised in two different files: the 'gene map' and the 'morbid map' available from the FTP site. The OMIM Gene Map is a single file, in tabular format, listing genes that are described in OMIM. Not all OMIM entries are included in the Gene Map, but only those for which a cytogenetic location has been published in the cited references. Each entry is a list of fields such as gene location, gene symbol, MIM number, disorders and reference. The OMIM Morbid Map is an alphabetical list of diseases used in the database and their corresponding cytogenetic locations.

Genetic Association Database

The Genetic Association Database (GAD) is a publicly available NIH based database of published gene-based genetic association studies which contains records of over 5,000 human genetic association studies. The database is gene centered and provides a standardized molecular nomenclature by including official HUGO gene symbol. Each record refers to a gene or a marker and is annotated with links to molecular databases (LocusLink, GeneCards) and references databases (PubMed, CDC) among others. The goal of this database is to allow the user to rapidly identify medically relevant polymorphism from the large volume of polymorphism and mutational data.

PharmGKB

The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) is a public resource that contains genomic, phenotype and clinical information collected from ongoing research and from the literature. It is devoted to cataloguing information about pharmacogenes such as those genes involved in modulating the response to drugs. These genes are pharmacogenes because they are involved in the pharmacokinetics (PK) of a drug (how the drug is absorbed, distributed, metabolised and eliminated) or the pharmacodynamics (PD) of a drug (how the drug acts on its target and its mechanisms of action). The aim is to capture the relationships between drugs, diseases/phenotypes and genes including several other types of information such as literature annotations, primary data sets, PK and PD pathways, and expert-generated summaries of PK/PD relationships.

Method

Gene Association Database

125160	Y	alopecia areata	IMMUNE	6	6p21.3	HLA-A	1157051	1160393	Xiao, F. L. et al. 2005	16185849
125147	Y	hepatitis C,	INFECTION	6	6p21.3	HLA-A	1157051	1160393	McKiernan, S. et al. 2004	15239092
125143	Y	psoriasis	IMMUNE	6	6p21.3	HLA-A	1157051	1160393	Zhang, X. J. et al. 2003	14527733

GAD vs DO term comparison

OMIM morbid map

3-methylglutaconic aciduria, type I, 250950
 (3) |AUH|600529|Chr.93q21q26 syndrome (1)
 |EVI1|165215|3q26
 6-mercaptopurine sensitivity, 610460
 (3) TPMT|187680|6p22.3
 ACAT2 deficiency (1)|ACAT2|100678|6q25

OMIM vs DO term comparison

Disease Ontology

gallbladder disease
 maturation disease
 body growth disease
 cell growth disease
 spleen disease
 Pigmentation Disorders
 Monarticular juvenile
 Rheumatoid arthritis
 Non-Neoplastic Eyelid
 Disorder Polyarthritis

GKB synonyms

Sialic Acid Storage Disease	Infantile Sialic Acid Storage Disease
Sialic Acid Storage Disease	Salla Disease
Sialic Acid Storage Disease	Sialic Acid Storage Disease, Finnish
Sialic Acid Storage Disease	Sialic Acid Storage Disease, Infantile
Sialic Acid Storage Disease	Sialuria
Sialic Acid Storage Disease	Sialuria, Finnish Type
Sialic Acid Storage Disease	Sialuria, Infantile Form
Sialic Acid Storage Disease	Finnish Type Sialuria
Sialic Acid Storage Disease	Finnish Type Sialurias
Sialic Acid Storage Disease	Infantile Form Sialuria

GKB vs DO term comparison

Disease Vocabulary

Schizophrenia	Psychotic disorder	Schizoaffective Disorder
Alzheimer disease	Senile dementia	Syn2b.....
Depression	Syn1c	Syn2c.....
Dis_d	Syn1d	Syn2d.....
Dis_e	Syn1e	Syn2e.....
Dis_f	Syn1f	Syn2f.....

DO Gene Annotation

DO: Anemia, Megaloblastic*Syn: Megaloblastic anemia- Norwegian type*Gene: AMN*
 DO: Anemia, Megaloblastic*Syn: Megaloblastic anemia- Finnish type*Gene: CUBN*
 DO: Anemia, Megaloblastic*Syn: Megaloblastic anemia- Finnish type*Gene: IFCR*
 DO: Anemia, Megaloblastic*Syn: Anemia megaloblastic due to DHFR deficiency*Gene: DHFR*
 DO: Anemia, Megaloblastic*Syn: Thiamine-responsive megaloblastic anemia syndrome*Gene: SLC19A2*
 DO: Anemia, Megaloblastic*Syn: Thiamine-responsive megaloblastic anemia syndrome*Gene: THTR1*

Criteria to pick up synonyms

Similar definitions are considered synonyms based on the validation of the following empirical rule:

$$I \geq \text{int}(K/2) \text{ must be true; } K=T-N$$

I=Identities, T=total words in the DO definition, N= words not relevant)

Order of terms does not influence score

Some exceptions to the rule

(e.g. for $T=1 \rightarrow I=K$)

Non-relevant words

Short words:

to, of, the, with,
and, or, in

Some long words:

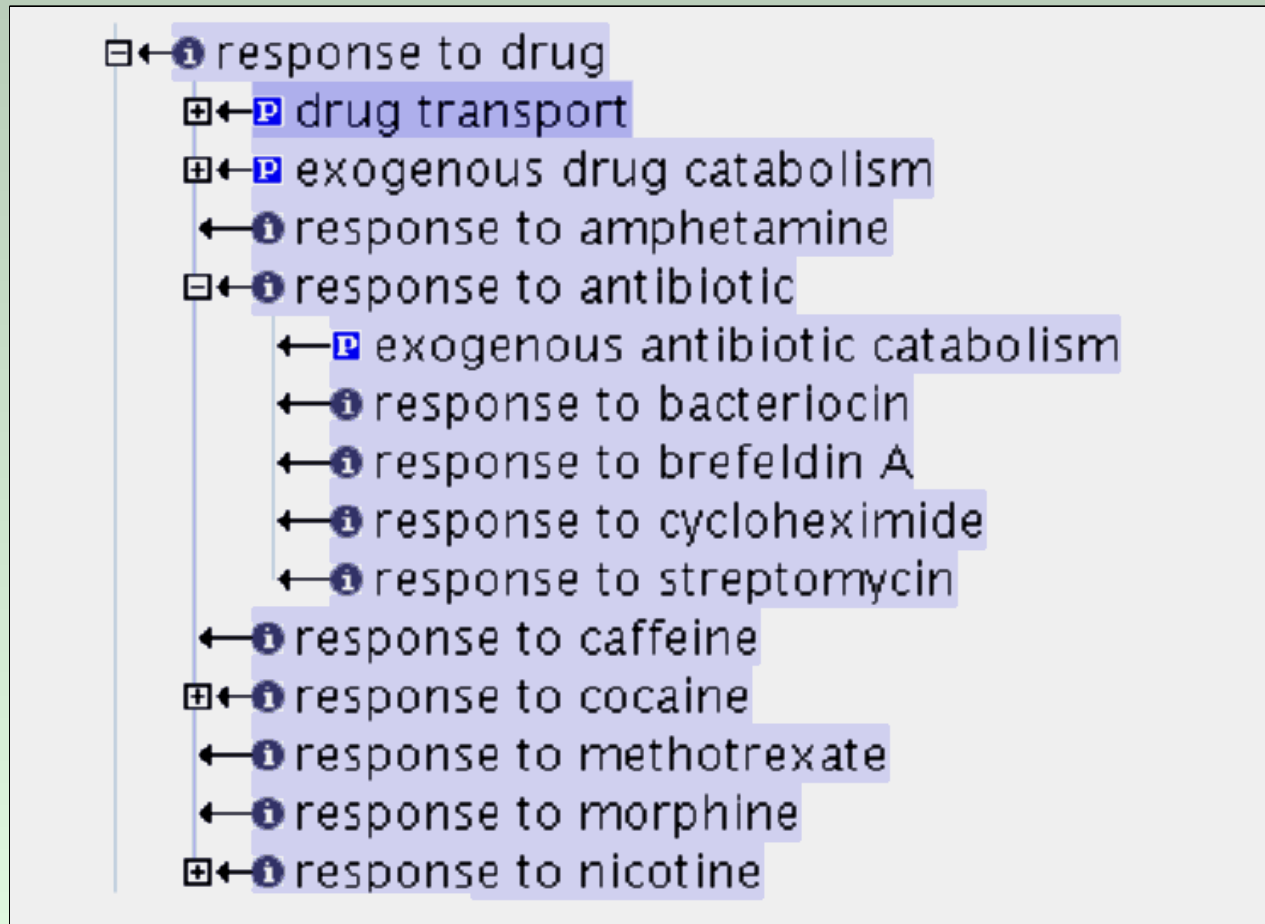
susceptibility,
disease, system,
chronic, syndrome,
storage, disorder,
acute, type, form

Finding synonyms

	Total matches (column C+D)	Identity matches	Manual curated matches
PharmGKB (3,998)	2,866	2,633	233
GAD (5,635)	2,700	424	2,276
OMIM (4,121)	2,084	184	1,900

Drugs

The response to drugs can be defined as a Biological Process in GO, but it is not an ideal solution to the problem.



Drugs

Which nomenclature should we use for drugs ?
(ChEBI + synonyms)

Then, we want to make two associations:

Drugs to genes (PharmGKB, DrugBank)

Drugs to diseases (PharmGKB, DrugBank)

PharmGKB

The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) is a public resource that contains genomic, phenotype and clinical information collected from ongoing research and from the literature. It is devoted to cataloguing information about pharmacogenes such as those genes involved in modulating the response to drugs. These genes are pharmacogenes because they are involved in the pharmacokinetics (PK) of a drug (how the drug is absorbed, distributed, metabolised and eliminated) or the pharmacodynamics (PD) of a drug (how the drug acts on its target and its mechanisms of action). The aim is to capture the relationships between drugs, diseases/phenotypes and genes including several other types of information such as literature annotations, primary data sets, PK and PD pathways, and expert-generated summaries of PK/PD relationships.



Chemical entities of biological interest

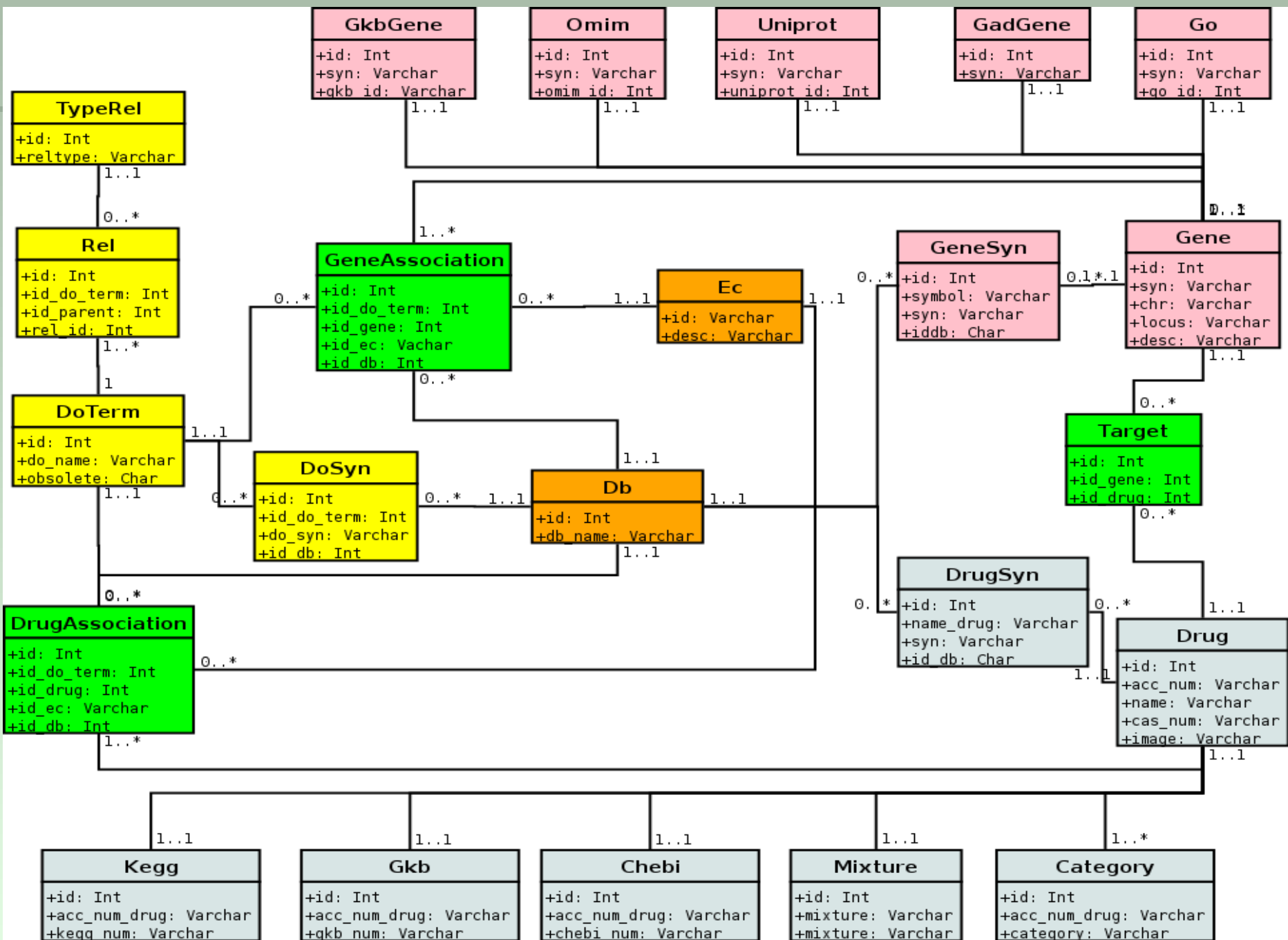
Is a standard terminology

it should finally be used in biological databases

- data in ChEBI is **manually curated**
- 4 sub-ontologies:
 - Molecular Structure
 - Biological Role
 - Application
 - Subatomic Particle

DrugBank

The DrugBank is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. This includes physical property data, structure and image files, pharmacological and physiological data about thousands of drug products as well as extensive molecular biological information about their corresponding drug targets. Each DrugCard contains more than 80 data fields with half of the information being devoted to drug/ chemical data and the other half devoted to drug target or protein data.



<http://bioinformatics.cribi.unipd.it/disease>