

Utilization of Existing Bio-Ontologies for the Annotation of Phenotype Data

Matt Mailman, NCBI

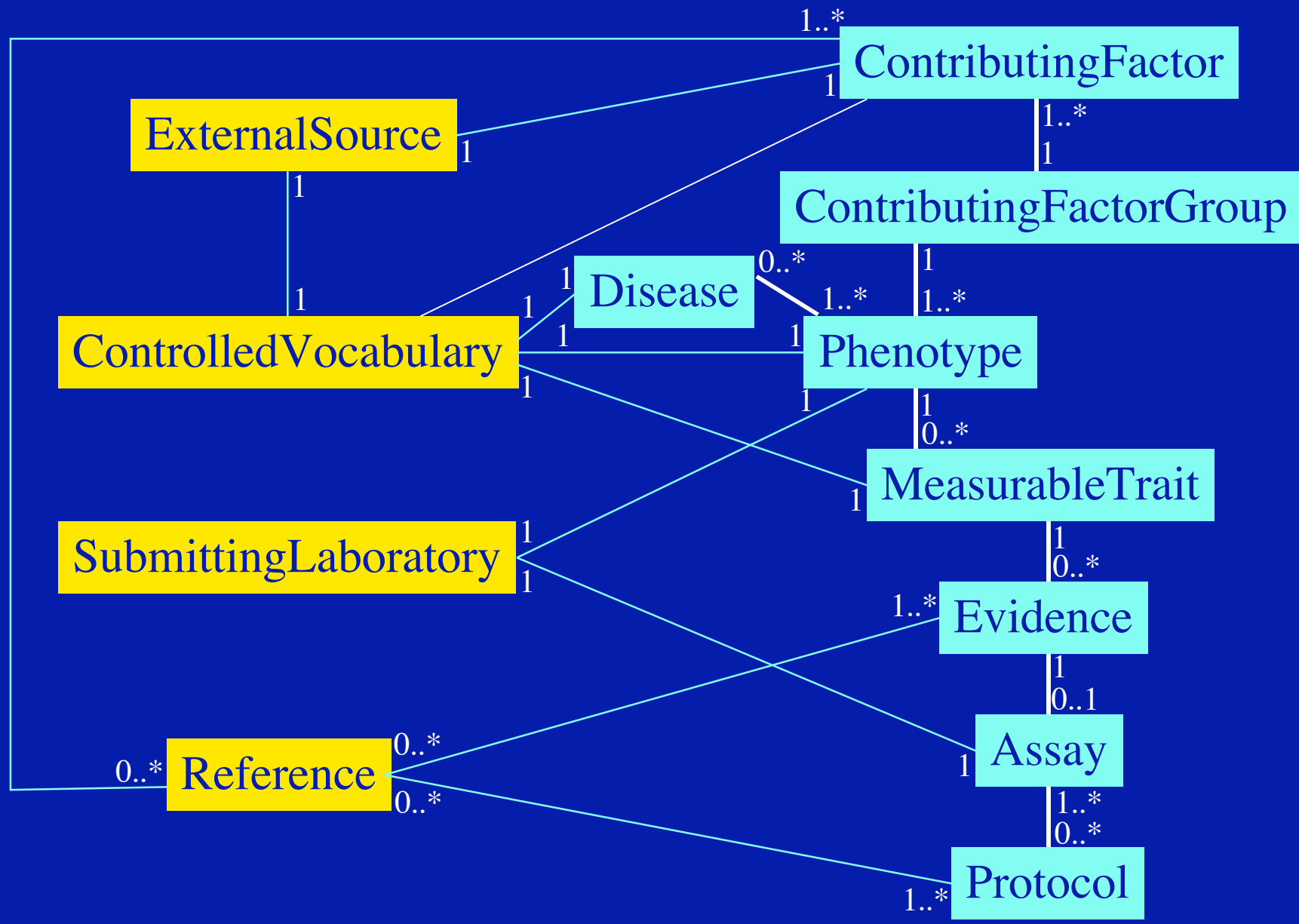
October 26, 2004

SOFG

Goals

- Model phenotype data in a structured manner to allow for powerful queries and hypothesis generation
- Establish an international public repository that allows for the comparison of phenotypes across studies and across species
- “De-confound” the actual measurable traits correlated with diseases
- Bring together categorical phenotype vocabulary from MGI and the measured value data described using PATO
- Associate phenotype with other NCBI resources such as dbSNP, haplotypes, Entrez Gene, GEO, and OMIM

Schema Design



Use of Controlled Vocabularies

CATEGORY	ORGANISM	VOCABULARY
anatomy	human	FMA
	mouse	MGI or FMA
	rat	MGI or FMA
	fly	FlyBase
phenotype	mammals (others?)	MGI Mammalian Phenotype Ontology, PATO
developmental stage	human	Carnegie stages + custom
	mouse	Theiler stages + custom
	rat	Witschi stages + custom
	fly	FlyBase
inheritance	any	custom (ie: dominant, recessive, codominant, etc)
drug treatment	any	MeSH
epigenetic changes	any	custom (ie: histone acetylation, nucleotide methylation)
disease	mammals	ICD, SNOMED, MeSH, NCI Metathesaurus
biological process	any	GeneOntology
cellular component	any	GeneOntology
molecular function	any	GeneOntology

Example - Human Colon Cancer

- Contributing Factors

- age - 29 years
- gender - male
- SNP - H718Y in MLH1 gene (dbSNP: rs2020873)
- epigenetics - methylation of MLH1 promoter (custom vocabulary)
- family history - affected father, uncle and brother
- diet - *still working on this vocabulary (suggestions?)*
- drugs -
- geographical location - Finland (MeSH: Z01.542.267)

- Disease

- name - hereditary nonpolyposis colon cancer (HNPCC)
- Online Mendelian Inheritance in Man (OMIM) - 114500
- International Classification of Diseases (ICD)
 - malignant neoplasm of digestive organs and peritoneum
 - malignant neoplasm of colon of descending colon (ICD: 153.2)
- MeSH - colorectal neoplasms, hereditary nonpolyposis (D003123)

Phenotype 1 - Abnormal DNA Mismatch Repair

- name - abnormal DNA mismatch repair
- description - expansion of microsatellite repeats
- controlled term - error-prone DNA repair (GO)
- taxonomy - Homo sapiens (NCBI)
- scale - biological macromolecule (custom)
- developmental stage - adult (custom)
- inheritance - recessive (custom)
- anatomy - nuclear deoxyribonucleic acid (FMA)

Phenotype 1 (continued)

- MeasurableTrait 1
 - name - BAT-25 microsatellite
 - unit - number of repeats
 - value - 50
- MeasurableTrait 2
 - name - D17S250 microsatellite
 - unit - number of repeats
 - value - 38
- MeasurableTrait 3
 - name - D5S346 microsatellite
 - unit - number of repeats
 - value - 52

Phenotype 1 (continued)

- Assay
 - name - microsatellite instability analysis
 - description - (details of assay)
- Evidence
 - (not necessary - no statistical test performed)
- Protocol
 - name - polymerase chain reaction
 - description - (details of protocol)
 - reference - Pubmed: 15217520
- SubmittingLaboratory
 - handle, PI, institution, submission date, etc.

Phenotype 2 - Colorectal Adenocarcinoma

- name - colorectal adenocarcinoma
- controlled term - abnormal colon morphology (MGI Phenotype)
- taxonomy - Homo sapiens (NCBI)
- scale - organ part (custom)
- developmental stage - adult (custom)
- inheritance - recessive (custom)
- anatomy - epithelium of ascending colon (FMA)

Phenotype 2 (continued)

- MeasurableTrait 1
 - name - immunohistochemical pathology
 - unit - histopathological observation
 - value - mucinous adenocarcinoma
- Assay
 - name - immunohistochemistry
 - description - (detailed description)
- Evidence - (none required because no statistical test)
- Protocol
 - name - histopathological analysis for mucinous adenocarcinoma
 - description - (detailed description)
 - reference - Kunstmann et al. 2004. BMC Medical Genetics 5:16 (PubMed: 15217520)
- SubmittingLaboratory - handle, PI, institution, submission date

Example Queries

- return all nonsynonymous SNPs or QTLs that are associated with breast cancer in humans and either mouse, rat, or dog
- return all contributing factors related to bristle number in *Drosophila* that have been observed in more than one study
- return all studies associated with cardiovascular disease in which cholesterol level was assayed

Status

- Finished
 - relational database schema
 - controlled vocabularies/ontologies stored
 - proof of concept test data
 - rat phenotype data from RGD
 - existing unformatted phenotype data in LocusLink
 - XML schema
- Current and future work
 - online submission forms
 - collection of new data
 - MGI phenotype
 - expecting millions of Drosophila phenotypes in Spring
 - dog strain disease and morphological phenotype data

Concerns

- reconciling redundant vocabularies

ie: International Classification of Diseases (ICD), SNOMED, Medical Subheadings (MeSH), National Cancer Institute (NCI) Diseases

Use one or more than one vocabulary?

Specialize – ie use NCI vocabulary for cancers?

- vocabulary for quantitative traits

ie: LDL cholesterol level, tumor mass, number of microsatellite repeats

– would be an difficult to develop and maintain

– PATO

Acknowledgements

Steve Sherry (NCBI - dbSNP)

Donna Maglott (NCBI - Entrez Gene, LocusLink)

Contact Information

Matt Mailman

email: mmailman@ncbi.nlm.nih.gov

phone: 301-402-5123