



PhenoGO : A Resource for the Multiscale Integration of Clinical and Biological Data




Lee T. Sam, Eneida A. Mendonça, Carol Friedman,
 Yves A. Lussier, M.D. (Corresponding Author)
 Director, Dept. of Medicine Center for Biomedical Informatics
 Associate Director for Informatics, Cancer Research Center
 Co-Director, Clinical Translational Science Award (CTSA) Informatics Core
 Associate Professor of Medicine, Biological Science Division



A Problem of Specificity: The $Foxn1^{nu}/Foxn1^{nu}$ mouse

Current GO model implies a unicellular organization

MG: $Foxn1^{nu}/Foxn1^{nu}$ →



OMIM (NCBI): In 2 sisters with T-cell immunodeficiency, congenital alopecia

GO 24 annotations for FOXN1
 “ T-lymphocyte function”
 “impaired embryonic morphogenesis”
 “keratinocyte differentiation”

PMID: 16232301
 Nude ($Foxn1^{nu}/Foxn1^{nu}$) mice develop largely normal hair follicles and produce hair shafts. Since hair shafts fail to penetrate the epidermis, macroscopic nudity results ...
 MH - Hair Follicle
 MH - Mice, Nude/anatomy & histology/genetics/physiology



Overview

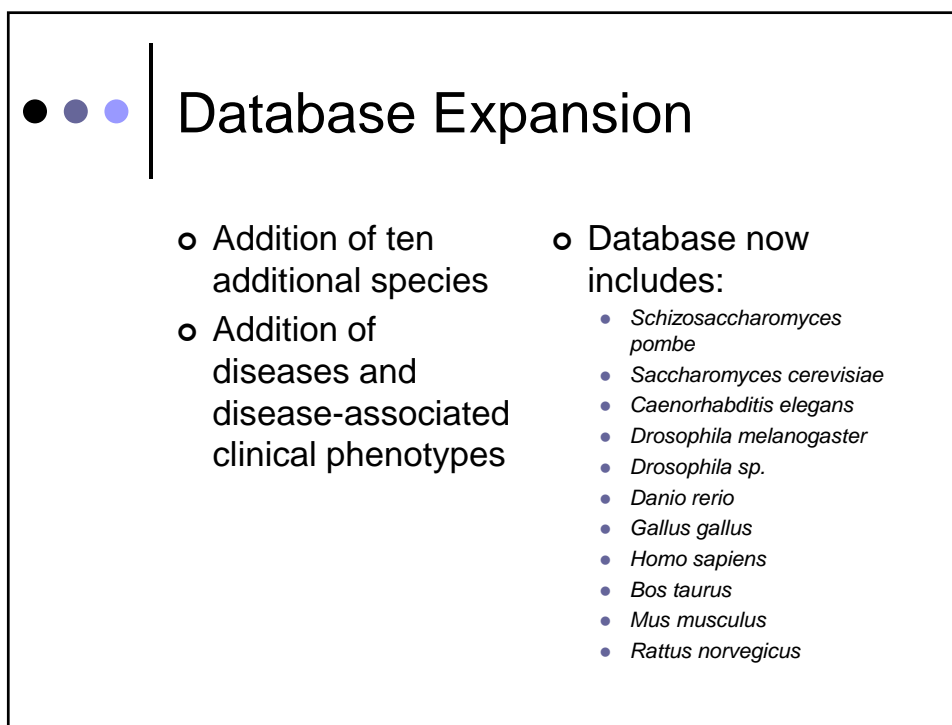
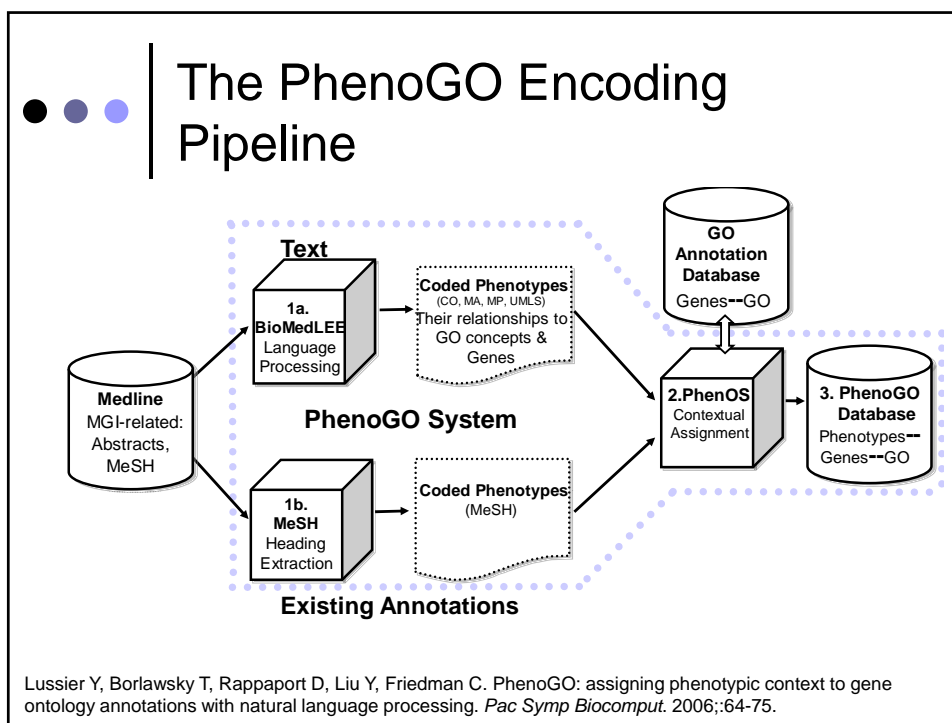
- Introduction/Previous Work
- Updates
- Web Interface
- Evaluation
 - Comprehensive
 - Disease annotation specific
 - Cell annotation specific
- Applications
- Conclusions

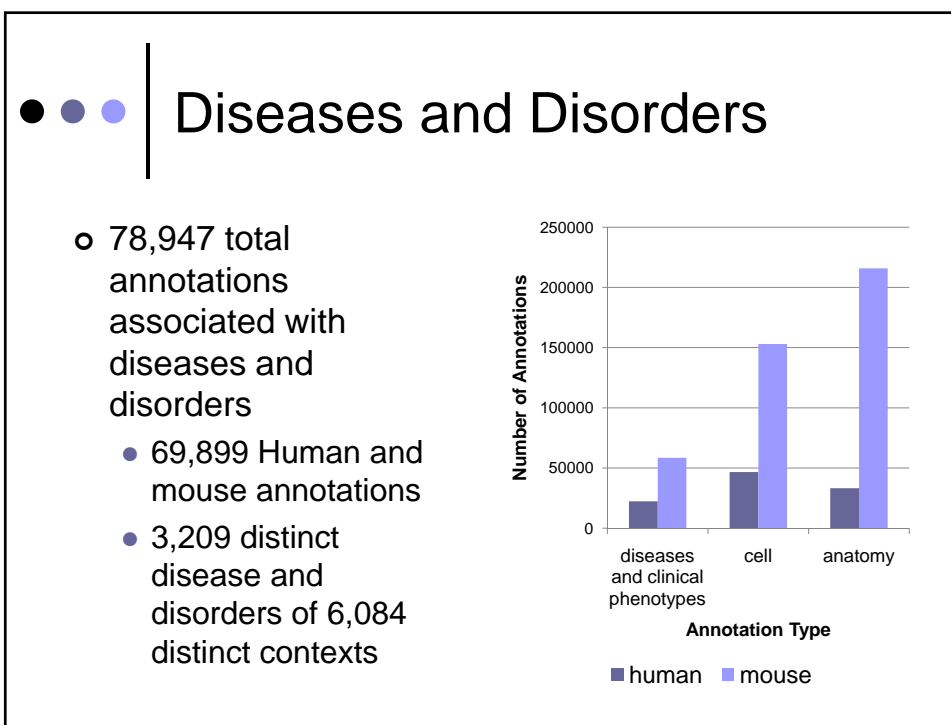
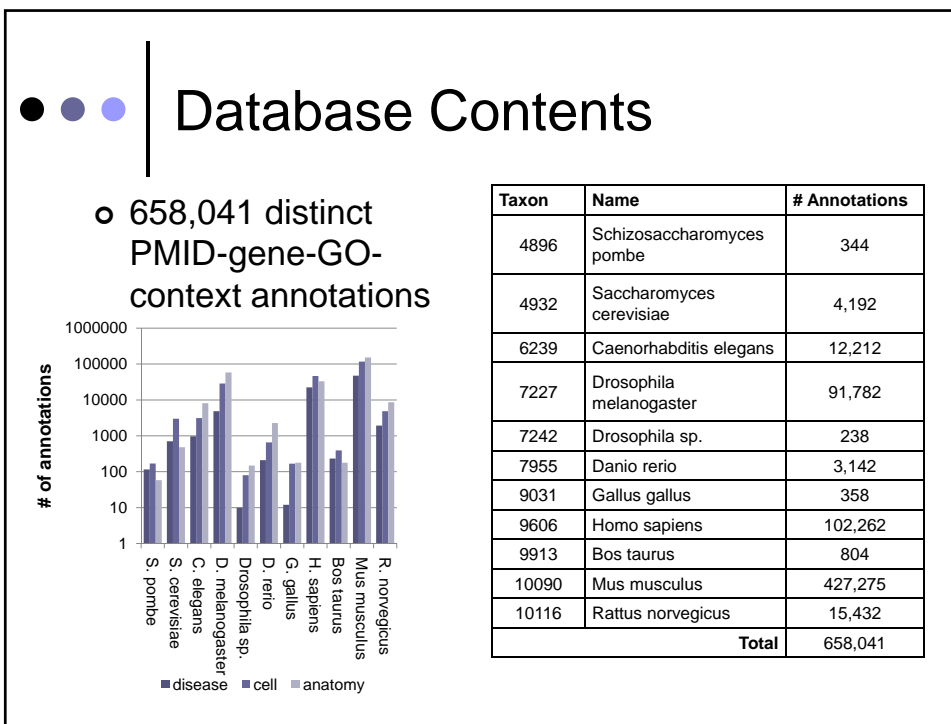


Introduction/Previous Work

- The PhenoGO database – adding context to Gene Ontology Annotations
- Open standards
 - Standardized ontologies and coding systems
 - Genes: MGI, SwissProt, UniGene, etc...
 - Phenotypes/contexts: Mammalian Phenotype Ontology, UMLS, Cell Type Ontology, MeSH, etc...
- 2006
 - 260,049 gene-GO-cell&anatomy context annotations
 - Specifically focused on the mouse

Lussier Y, Borlawsky T, Rappaport D, Liu Y, Friedman C. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput.* 2006;:64-75.





Web Query: Basic

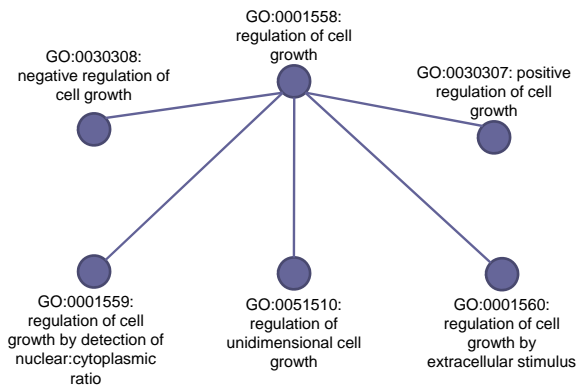
- Runs a query analogous to an SQL OR for the search terms

Web Query: Advanced

- Runs a query analogous to an SQL AND for the search terms

Hierarchical Query

- Query for more specific descendent terms
- Uses a number of ancestor-descendent tables to build a SQL query



Result Formats

- HTML-formatted

result #	gene code	gene name	gene description	GO code	GO name	comment code	comment name	tissue	source
1	137203Z7	MG2-135344	Transposon insertion, gene 3	GO:000239	apoptosis (cellular)	UMLS:CO02388	leuM	10000	MG
2	137203Z7	MG2-135344	Transposon insertion, gene 3	GO:000829	apoptosis (intrinsic)	UMLS:CO02382	leuM	10000	MG
3	137203Z7	MG2-181903	Transposon insertion, gene 3	GO:000558	apoptosis (intrinsic)	UMLS:CO01593	leuM	10000	MG
4	137203Z7	MG2-181903	Transposon insertion, gene 3	GO:004252	apoptosis (intrinsic)	UMLS:CO01593	leuM	10000	MG
5	137203Z7	MG2-181903	Transposon insertion, gene 3	GO:000588	apoptosis (intrinsic)	UMLS:CO01593	leuM	10000	MG
6	137203Z7	MG2-181903	Transposon insertion, gene 3	GO:000558	apoptosis (intrinsic)	UMLS:CO02388	leuM	10000	MG
7	137203Z7	MG2-181903	Transposon insertion, gene 3	GO:004252	apoptosis (intrinsic)	UMLS:CO02388	leuM	10000	MG
8	137203Z7	MG2-181903	Transposon insertion, gene 3	GO:000588	apoptosis (intrinsic)	UMLS:CO02388	leuM	10000	MG
9	137203Z7	MG2-155448	Transposon insertion, gene 3	GO:000683	apoptosis (intrinsic)	UMLS:CO02927	leuM	10000	MG
10	137203Z7	MG2-155448	Transposon insertion, gene 3	GO:017388	apoptosis (intrinsic)	UMLS:CO02927	leuM	10000	MG

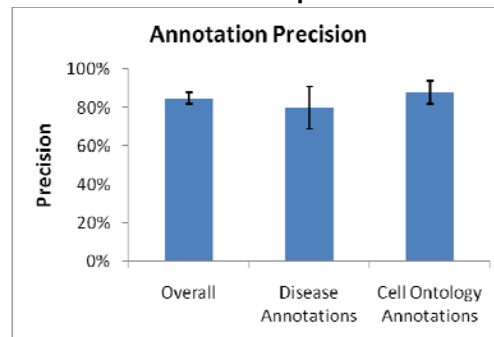
- Tab-delimited text

```

result # gene code gene name gene description GO code GO name comment code comment name tissue source
1 137203Z7 MG2-135344 Transposon insertion, gene 3 GO:000239 apoptosis (cellular) UMLS:CO02388 leuM 10000 MG
2 137203Z7 MG2-135344 Transposon insertion, gene 3 GO:000829 apoptosis (intrinsic) UMLS:CO02382 leuM 10000 MG
3 137203Z7 MG2-181903 Transposon insertion, gene 3 GO:000558 apoptosis (intrinsic) UMLS:CO01593 leuM 10000 MG
4 137203Z7 MG2-181903 Transposon insertion, gene 3 GO:004252 apoptosis (intrinsic) UMLS:CO01593 leuM 10000 MG
5 137203Z7 MG2-181903 Transposon insertion, gene 3 GO:000588 apoptosis (intrinsic) UMLS:CO01593 leuM 10000 MG
6 137203Z7 MG2-181903 Transposon insertion, gene 3 GO:000558 apoptosis (intrinsic) UMLS:CO02388 leuM 10000 MG
7 137203Z7 MG2-181903 Transposon insertion, gene 3 GO:004252 apoptosis (intrinsic) UMLS:CO02388 leuM 10000 MG
8 137203Z7 MG2-181903 Transposon insertion, gene 3 GO:000588 apoptosis (intrinsic) UMLS:CO02388 leuM 10000 MG
9 137203Z7 MG2-155448 Transposon insertion, gene 3 GO:000683 apoptosis (intrinsic) UMLS:CO02927 leuM 10000 MG
10 137203Z7 MG2-155448 Transposon insertion, gene 3 GO:017388 apoptosis (intrinsic) UMLS:CO02927 leuM 10000 MG
    
```

Evaluations

- Comprehensive Evaluation
- Cell Ontology-specific Evaluation
- Disease Annotations-specific Evaluation



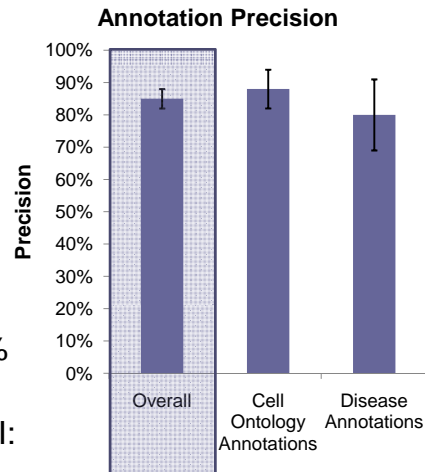
Evaluation Methodology

- Precision:
 - A set of randomly extracted entries was extracted and each was evaluated for correctness manually
- Recall
 - Randomly extracted sentences were presented to evaluators and each was checked to exist in the database if it made sense



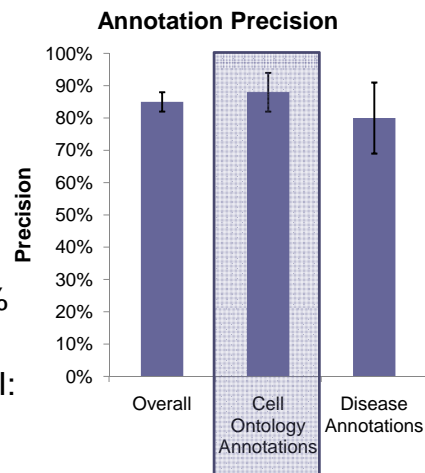
Comprehensive Evaluation

- 2 Experienced Evaluators with biological backgrounds
- n = 300 phenotypic annotations evaluated independently
- Precision: 85% (95% CI: 82%-89%)
- Recall: 76% (95% CI: 69-83%)



Cell Ontology-Specific Evaluation

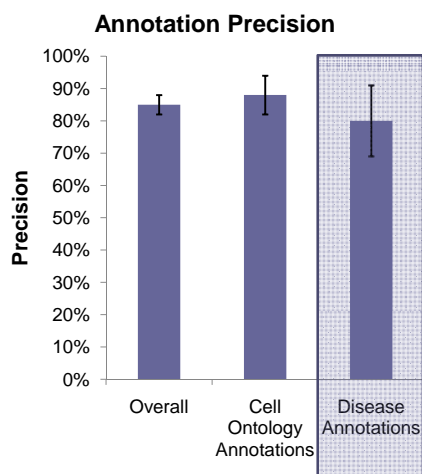
- 2 Experienced Evaluators with biological backgrounds
- n = 50 cell ontology-specific annotations
- Precision: 88% (95% CI: 82%-94%)
- Recall: 79% (95% CI: 69%-89%)





Disease-Specific Evaluation

- 2 Experienced Evaluators with biological background
- n = 50 disease and disease-associated clinical phenotypes evaluated independently
- Precision : 80% (95% CI:69%-90%)



Applications

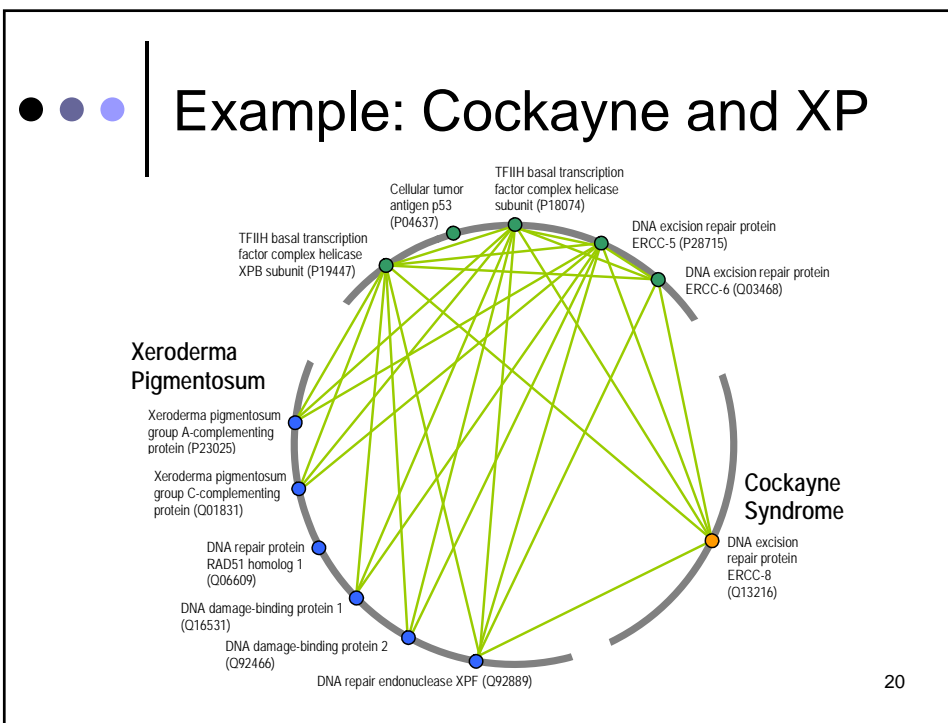
- Gene-Phenotype annotations play a central role in the study of the networks associated with human disease
- Approaches
 - Manual annotation
 - Krauthammer M, et al., *PNAS* (2004)
 - Goh, et al., *PNAS* (2007)
 - NLP
 - Lage, et al., *Nature Biotech.* (2007)

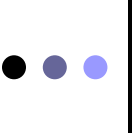
Discovery of Shared Phenotype-Associated Protein Interaction Subnetworks

- o Integration of:
 - The Reactome protein-protein interaction network
 - Human-specific PhenoGO gene-disease associations
- o Resulting in 11,703 disease-disease comparisons

UMLS ID	Disease 1	UMLS ID	Disease 2	Corrected pvalue
C0009207	Cockayne Syndrome	C0043346	Xeroderma Pigmentosum	8.5e-18
C0043346	Xeroderma Pigmentosum	C0085390	Li-Fraumeni Syndrome	4.9e-06
C0007001	Carbohydrate Metabolism, Inborn Errors	C0002514	Amino Acid Metabolism, Inborn Errors	6.2e-05
C0009404	Colorectal Neoplasms	C0950123	Genetic Diseases, Inborn	5.0e-05
C0085390	Li-Fraumeni Syndrome	C0009207	Cockayne Syndrome	1.9e-04

Sam L, Liu Y, Li J, Friedman C, Lussier YA. Discovery of protein interaction networks shared by diseases. *Pac Symp Biocomput.* 2007;:76-87.

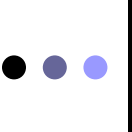




Top Results

UMLS ID	Disease 1	UMLS ID	Disease 2	P-PI (#)	pvalue
C0009207	Cockayne Syndrome	C0043346	Xeroderma Pigmentosum	38	7.3e-22
C0043346	Xeroderma Pigmentosum	C0085390	Li-Fraumeni Syndrome	24	6.7e-11
C0007001	Carbohydrate Metabolism, Inborn Errors	C0002514	Amino Acid Metabolism, Inborn Errors	9	8.3e-10
C0009404	Colorectal Neoplasms	C0950123	Genetic Diseases, Inborn	16	6.7e-10
C0085390	Li-Fraumeni Syndrome	C0009207	Cockayne Syndrome	16	2.7e-09
C0009404	Colorectal Neoplasms	C0015625	Fanconi's Anemia	8	1.5e-05
C0009404	Colorectal Neoplasms	C0085413	Polycystic Kidney, Autosomal Dominant	8	1.5e-05
C0024141	Lupus Erythematosus, Systemic	C0004364	Autoimmune Diseases	4	9.3e-05
C0024314	Lymphoproliferative Disorders	C0004364	Autoimmune Diseases	6	1.3e-04
C0024314	Lymphoproliferative Disorders	C0024141	Lupus Erythematosus, Systemic	6	1.3e-04

21



Results after multiple-testing adjustment

<p><u>Dunn-Sidak correction</u></p> $p' = 1 - (1 - p)^r$ <ul style="list-style-type: none"> ○ 5 significant results ○ Bonferroni-type method ○ Less severe, but still very conservative ○ Study $r = 11,703$ comparisons 	<p><u>Permutation Resampling</u></p> <ul style="list-style-type: none"> ○ 162 significant results ○ Uncorrected $p < .092$ for 5% error rate ○ How much does the p-value depend on which gene-phenotype associations exist? ○ Build background sample from randomized gene-phenotype data ○ 1000 iterations
--	--

22



Conclusions

- Substantial additions to a high-quality, wide-ranging gene-GO-phenotype resource drawn from the biomedical literature and existing knowledge bases
- It's a fantastic resource and I urge everyone to make use of it
- <http://www.phenogo.org>



Acknowledgements

- **University of Chicago**
 - Tara Borlawsky (now at Ohio State)
 - Yang Liu
 - Jianrong Li
- **Grants**
 - NLM, 1K22 LM008308-01 (YL)
 - NLM, R01 LM007659-01 (CF)
 - NLM 1U54 CA121852-01A1 National Center: Multi-Scale Study of Cellular Networks