# Functional Gene Clustering by Latent Semantic Indexing of MEDLINE Abstracts

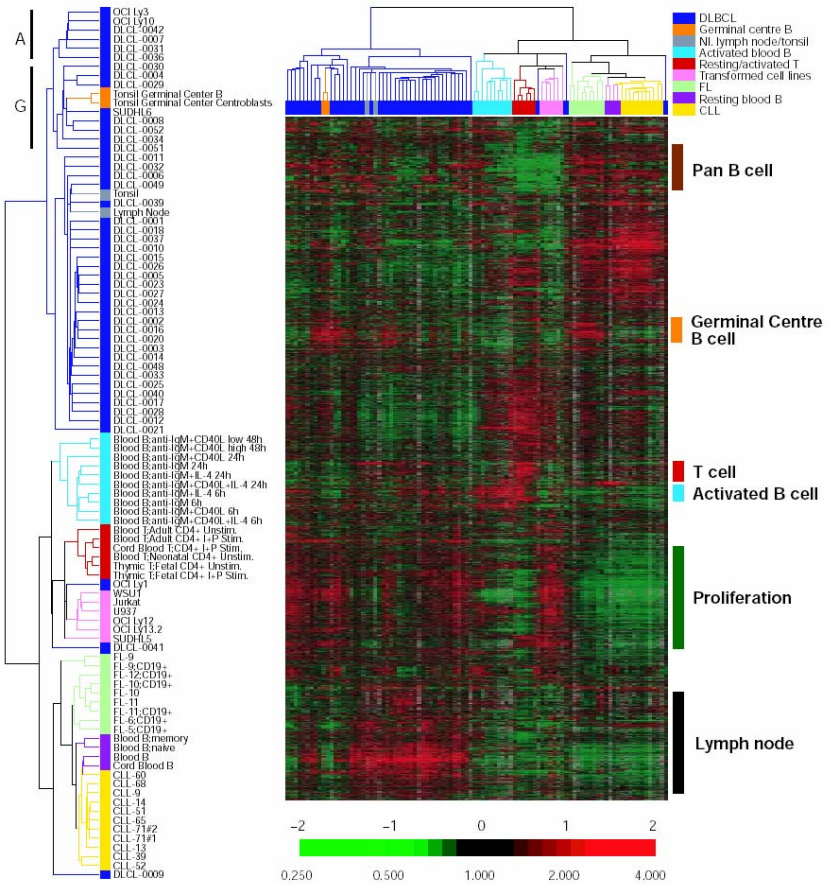
Ramin Homayouni, Ph.D.
Department of Neurology
University of Tennessee Health Science Center

Genomics and Bioinformatics Center of Excellence

CITR

Center *for* Neurobiology *of* Brain Diseases

# Gene Expression Profiling




Alizadeh, et al., (2000) Nature 403:503.

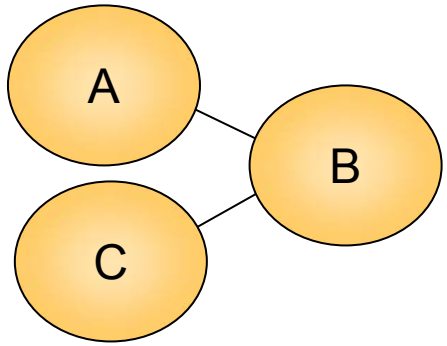
Now What?

# Some Web Resources

- **NCBI Sites**
  - **OMIM** http://www.ncbi.nlm.nih.gov/Literature/index.html
  - **LocusLink** http://www.ncbi.nlm.nih.gov/LocusLink/
  - **PubMed** http://www.ncbi.nlm.nih.gov/entrez/

- **Others**
  - **HAPI** http://array.ucsd.edu/hapi/
  - **GenMAPP** http://www.genmapp.org/
  - **GO Tree Machine** http://genereg.ornl.gov/gotm/
  - **PubGene** http://www.pubgene.org
  - **Arrowsmith** http://arrowsmith.psych.uic.edu/
  - **Chillibot** http://www.chilibot.net/

# Defining Functional Relationships between Genes



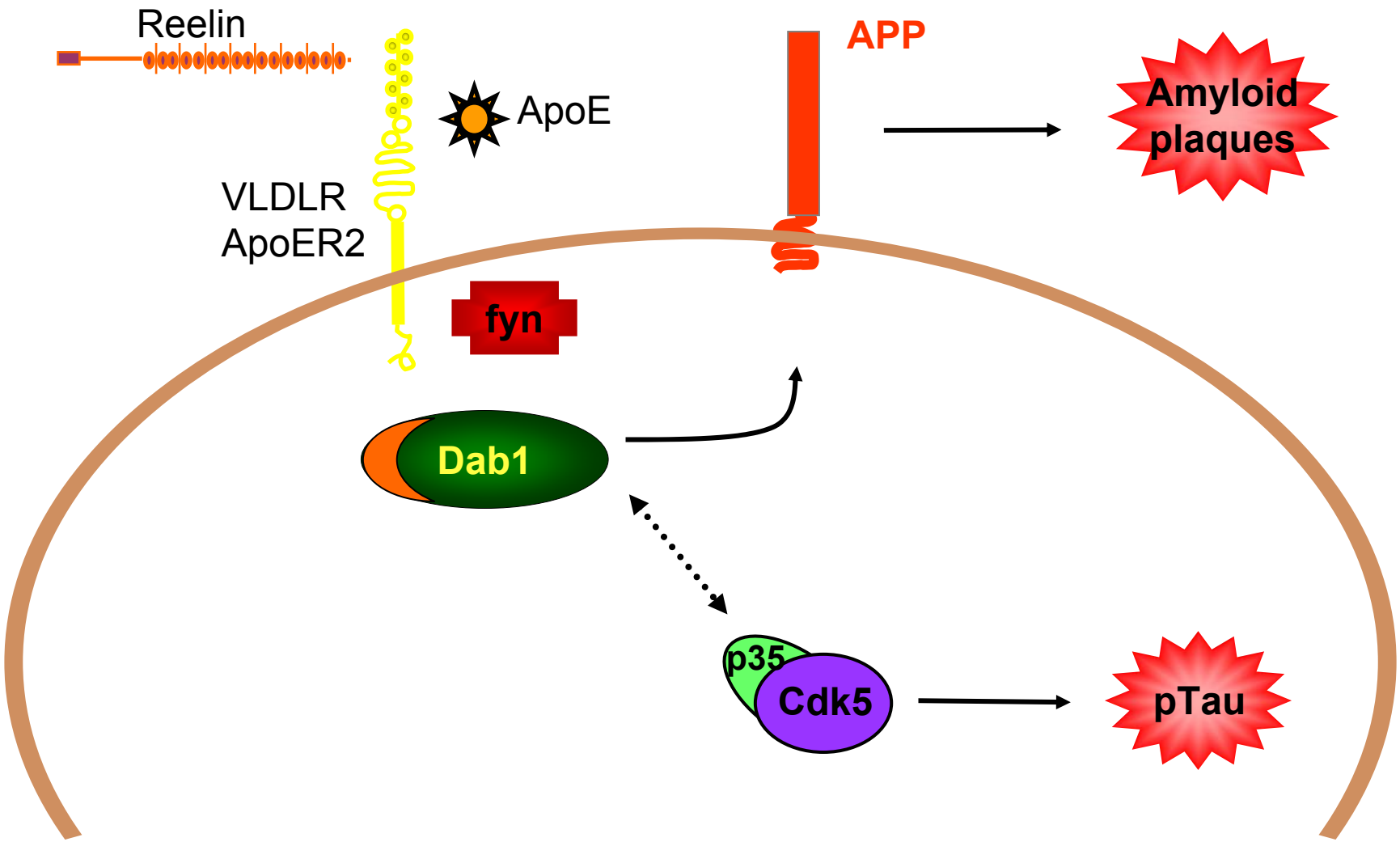

- **Direct Relationship**

  **Gene relationships already known (e.g., A-B or B-C)**
  - **Term co-occurrence**
    - **Gene symbol: PubGene (*Jenssen et al., Nature Genetics 2001 28:21)***
    - **Gene names (synonyms and aliases) – biochemical**

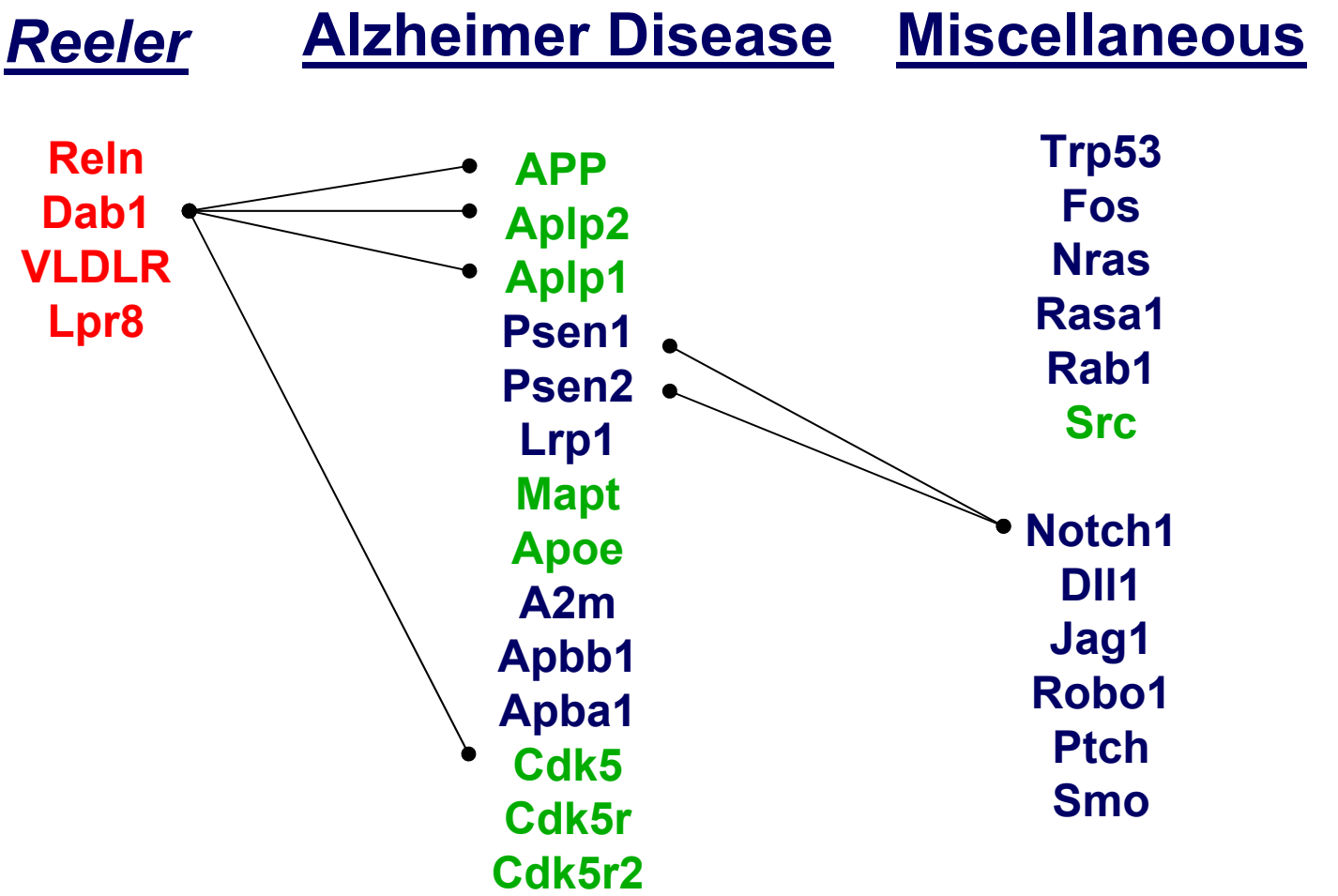
- **Indirect Relationship**

  **Gene relationships unknown (e.g., such as A-C)**

# Reelin Signaling Pathway



# Gene Document Test Set

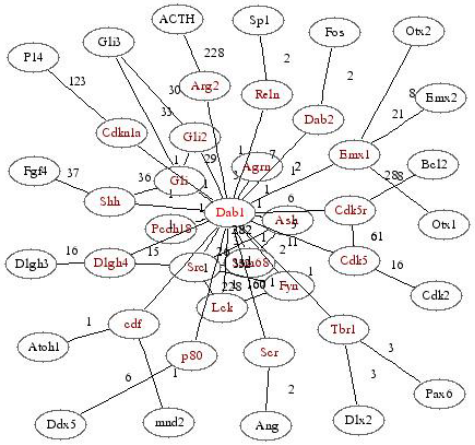


# PubGene Query: Dab1

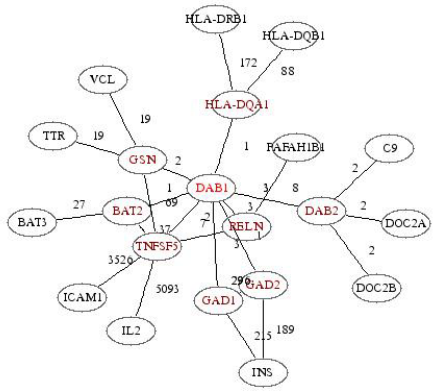
http://www.pubgene.org/

## Mouse

- Reln 7 times
- Cdk5r 6 times
- Cdk5 5 times
- Gli2 3 times
- Src 3 times
- Dab2 2 times
- Fyn 2 times
- Sam68 1 times
- Cdkn1a 1 times
- Tbr1 1 times
- Gli 1 times
- Scr 1 times
- Shh 1 times
- cdf 1 times
- Ash 1 times
- Dlgh4 1 times
- p80 1 times
- Lck 1 times
- Emx1 1 times
- Pcdh18 1 times
- Agrn 1 times
- Arg2 1 times




## Human



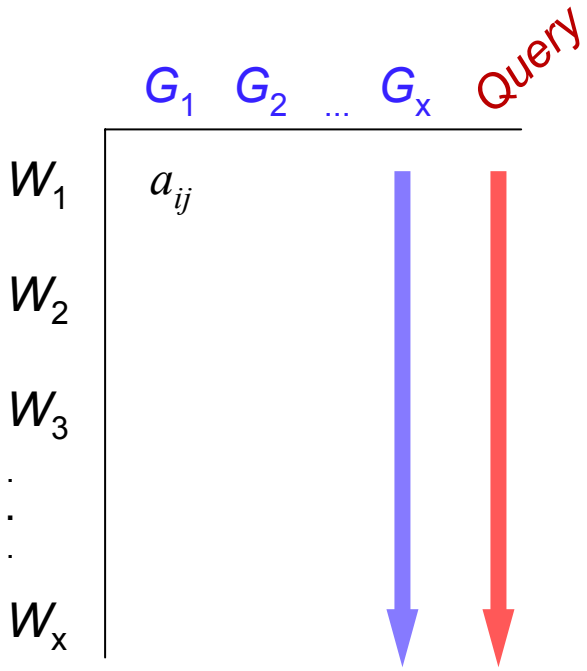

- DAB2 3 times
- GAD1 3 times
- RELN 3 times
- GSN 2 times
- TNFSF5 2 times
- HLA-DQA1 1 times
- BAT2 1 times
- GAD2 1 times

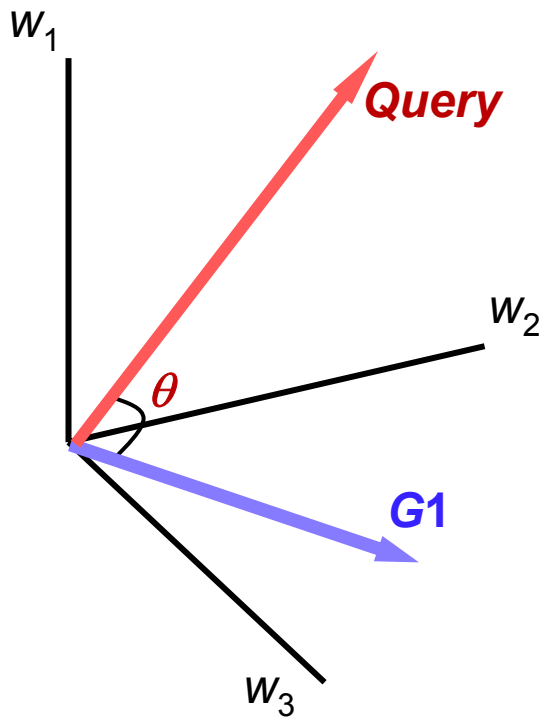
**PubMed Query: Dab1 AND Reln = 10**

**PubMed Query: Dab1 AND reelin = 57 !**

# Vector Space Model:

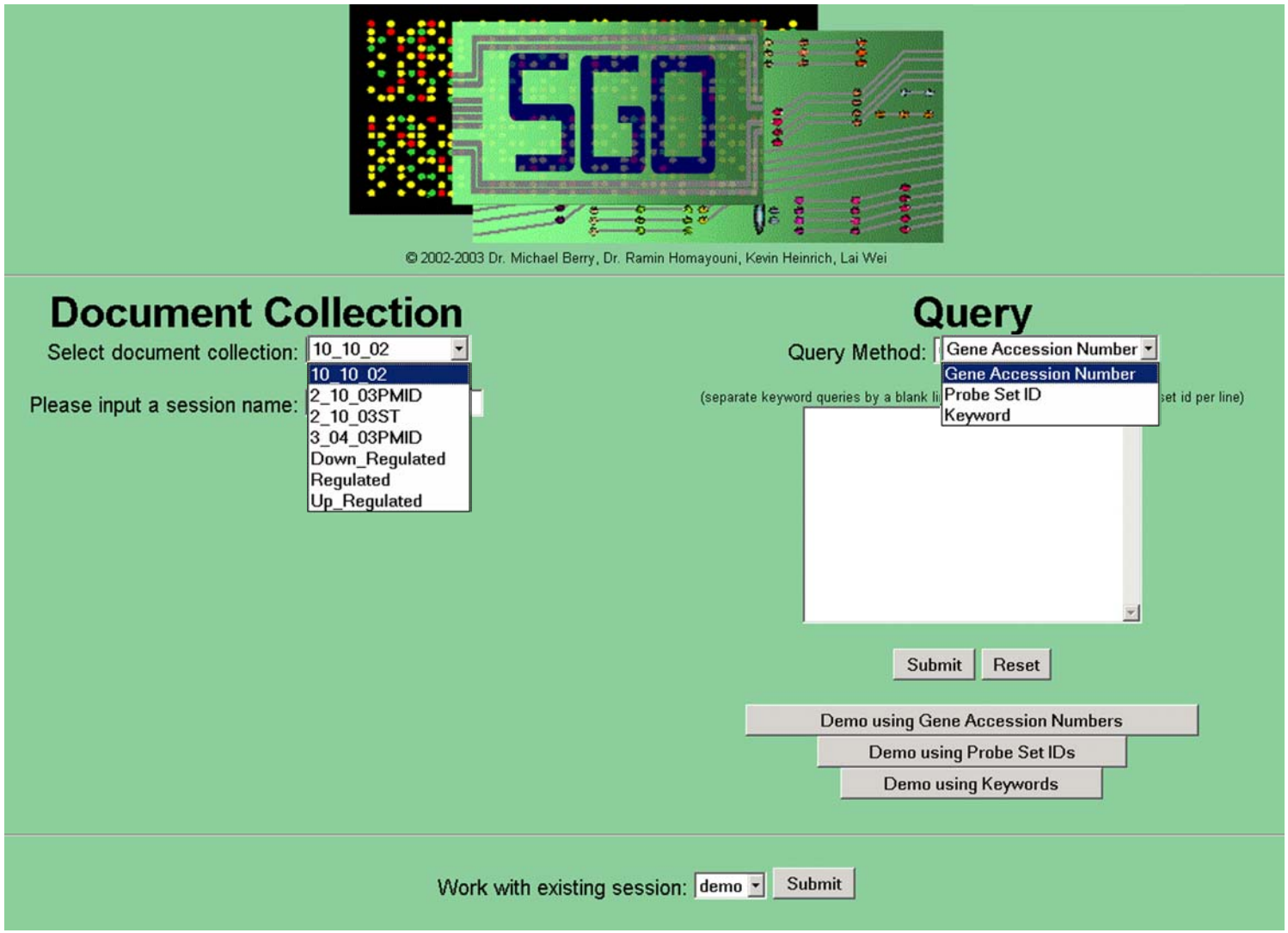
## Latent Semantic Indexing



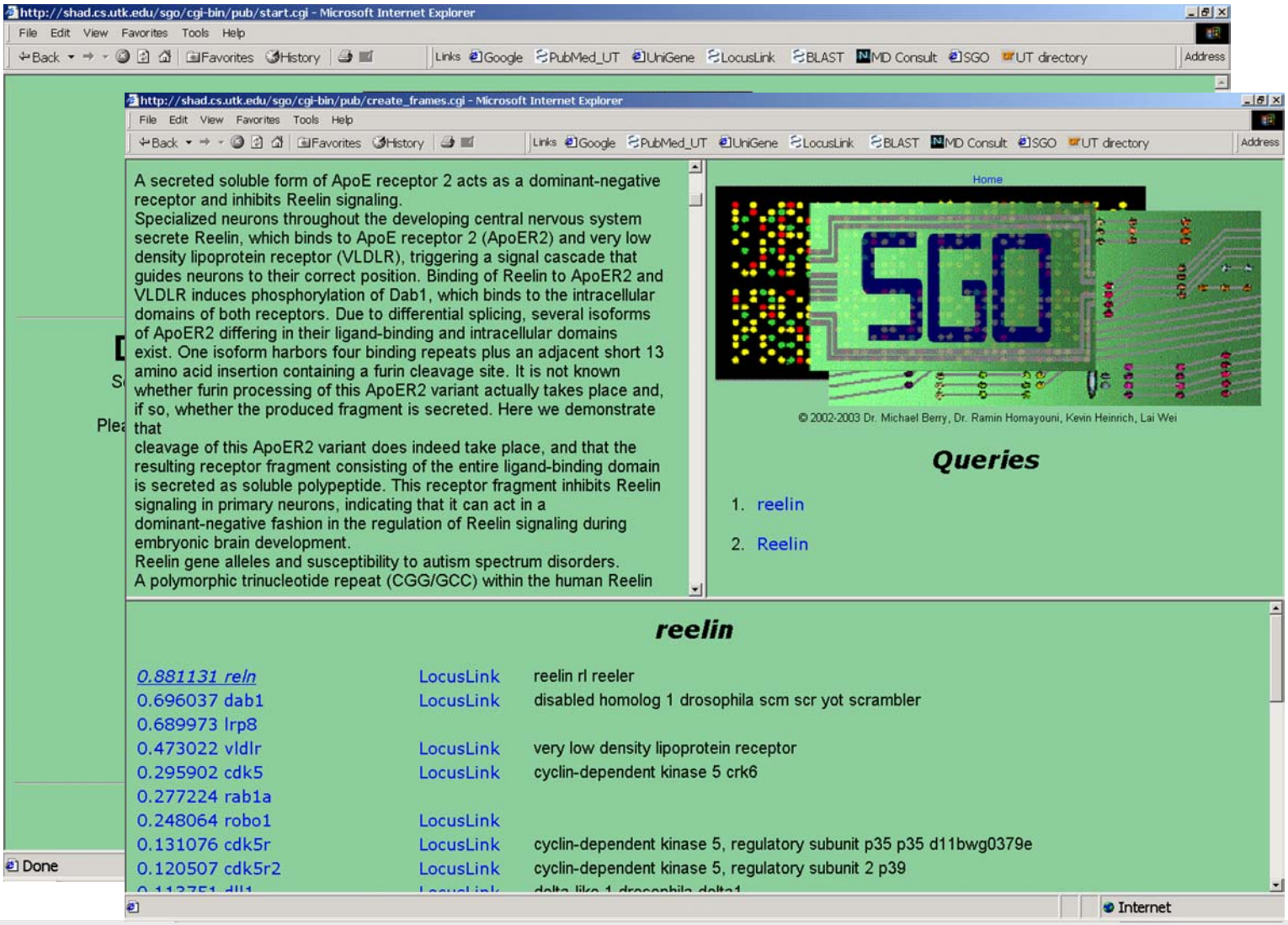

$$a_{ij} = l_{ij}\, g_i$$

# Semantic Gene Organizer© User Interface



# Reelin Accession # Query

# Reelin Keyword Query



# Evaluation of Results

- **Recall (completeness)**

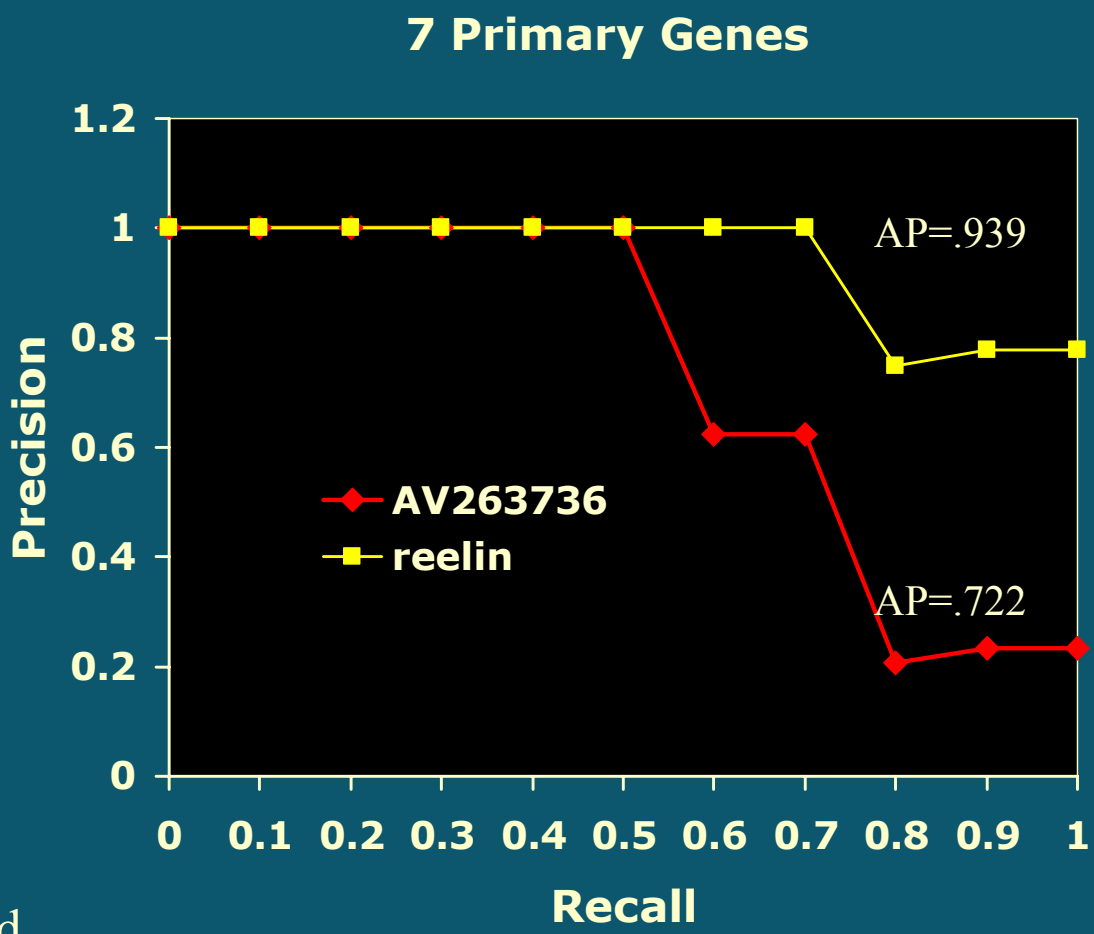
$$\text{Recall} = \frac{\text{Relevant documents retrieved}}{\text{Total \# relevant documents}}$$

- **Precision (accuracy)**

$$\text{Precision} = \frac{\text{Relevant documents retrieved}}{\text{Total \# documents retrieved}}$$

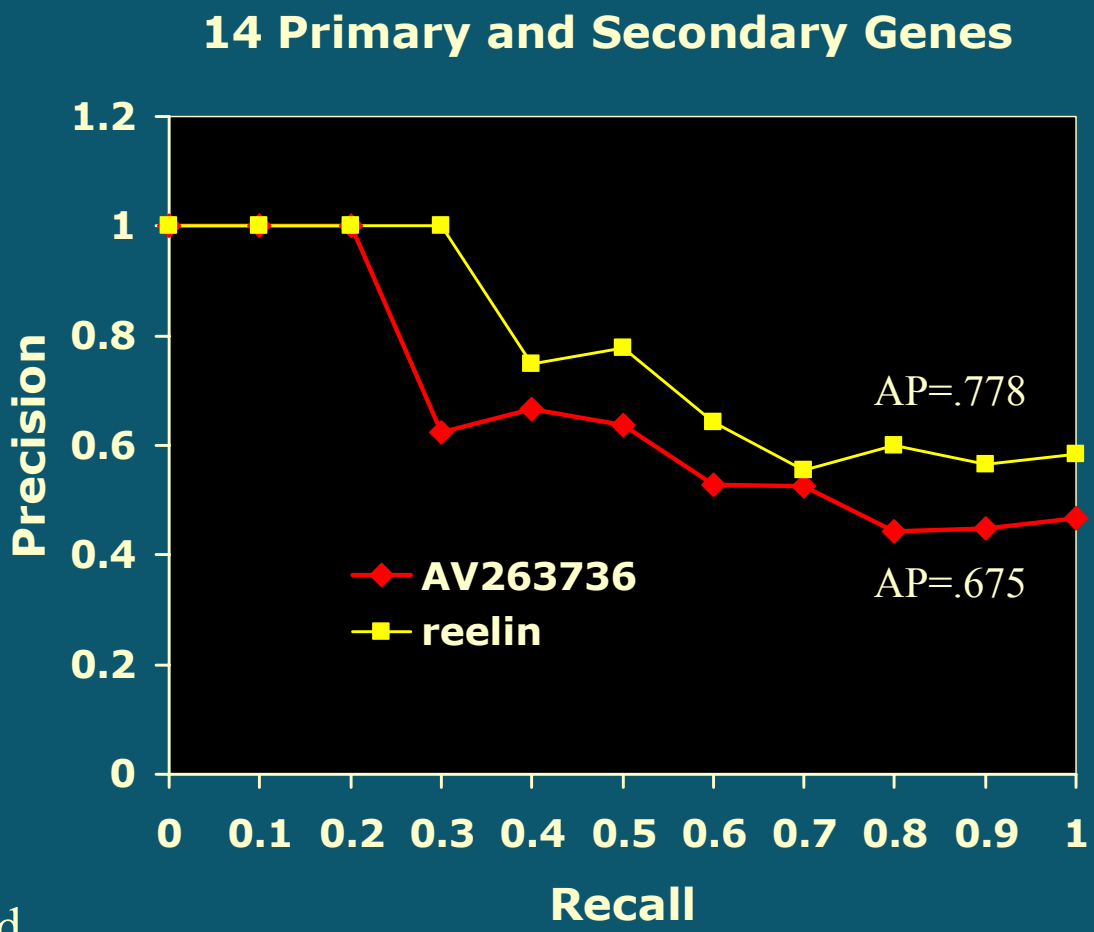
- **Recall and precision are inversely related.**

# SGO Performance on Reelin



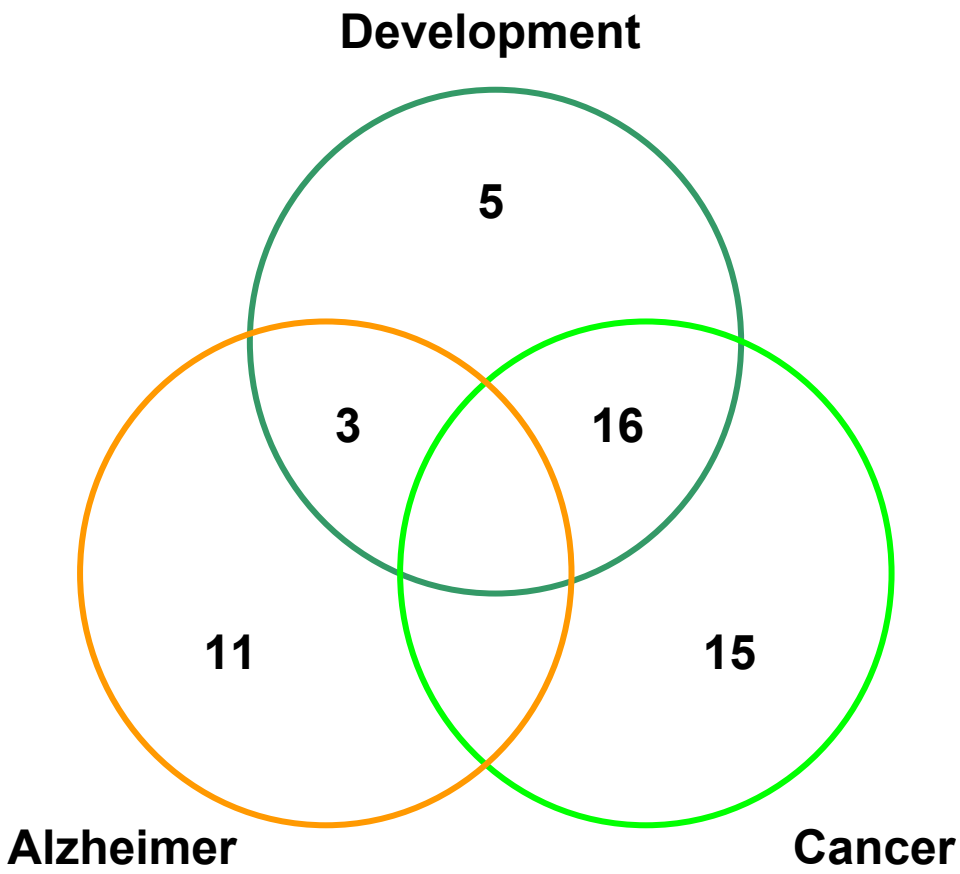

11-pt interpolated average precision

# SGO Performance on Reelin

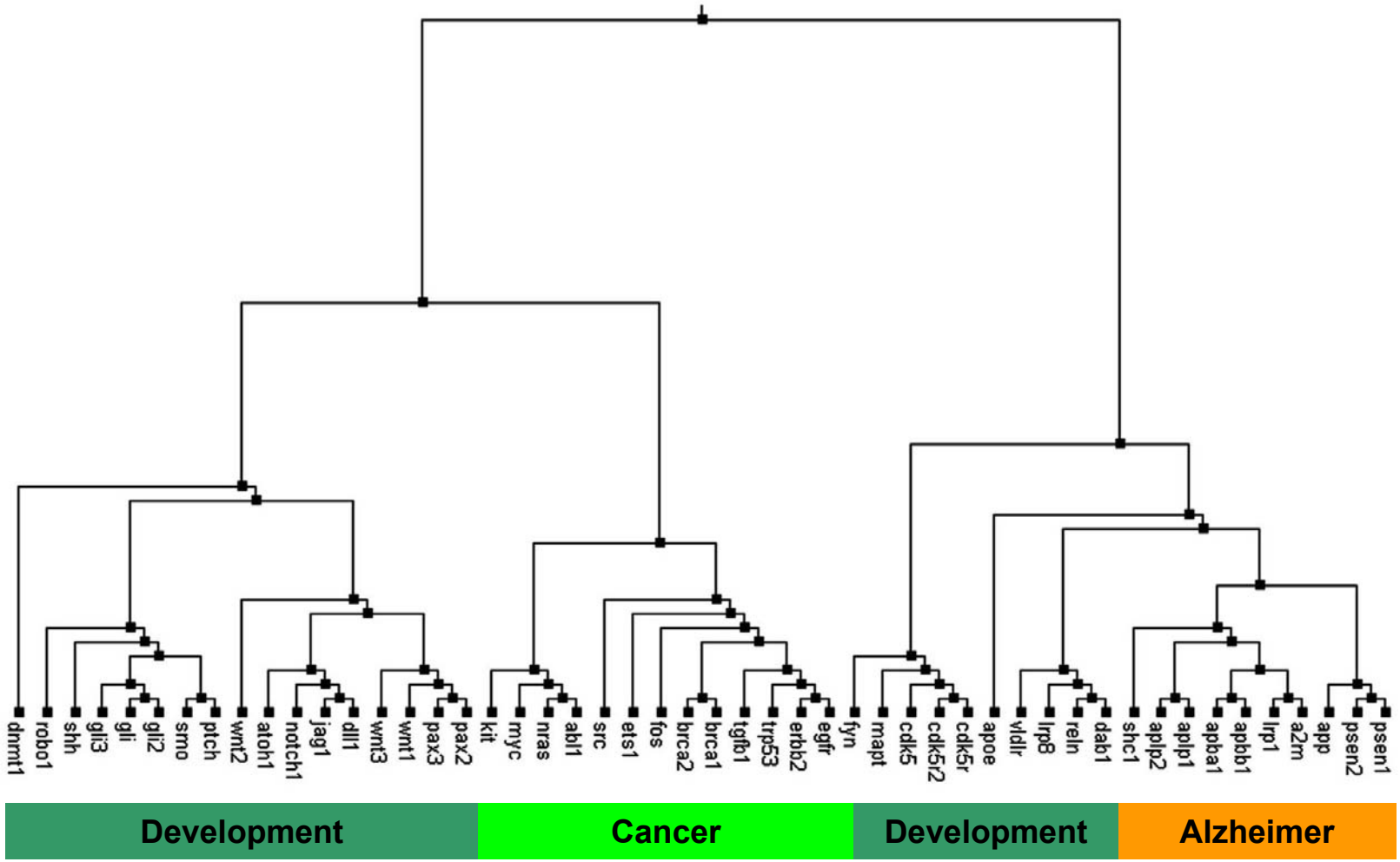



11-pt interpolated average precision

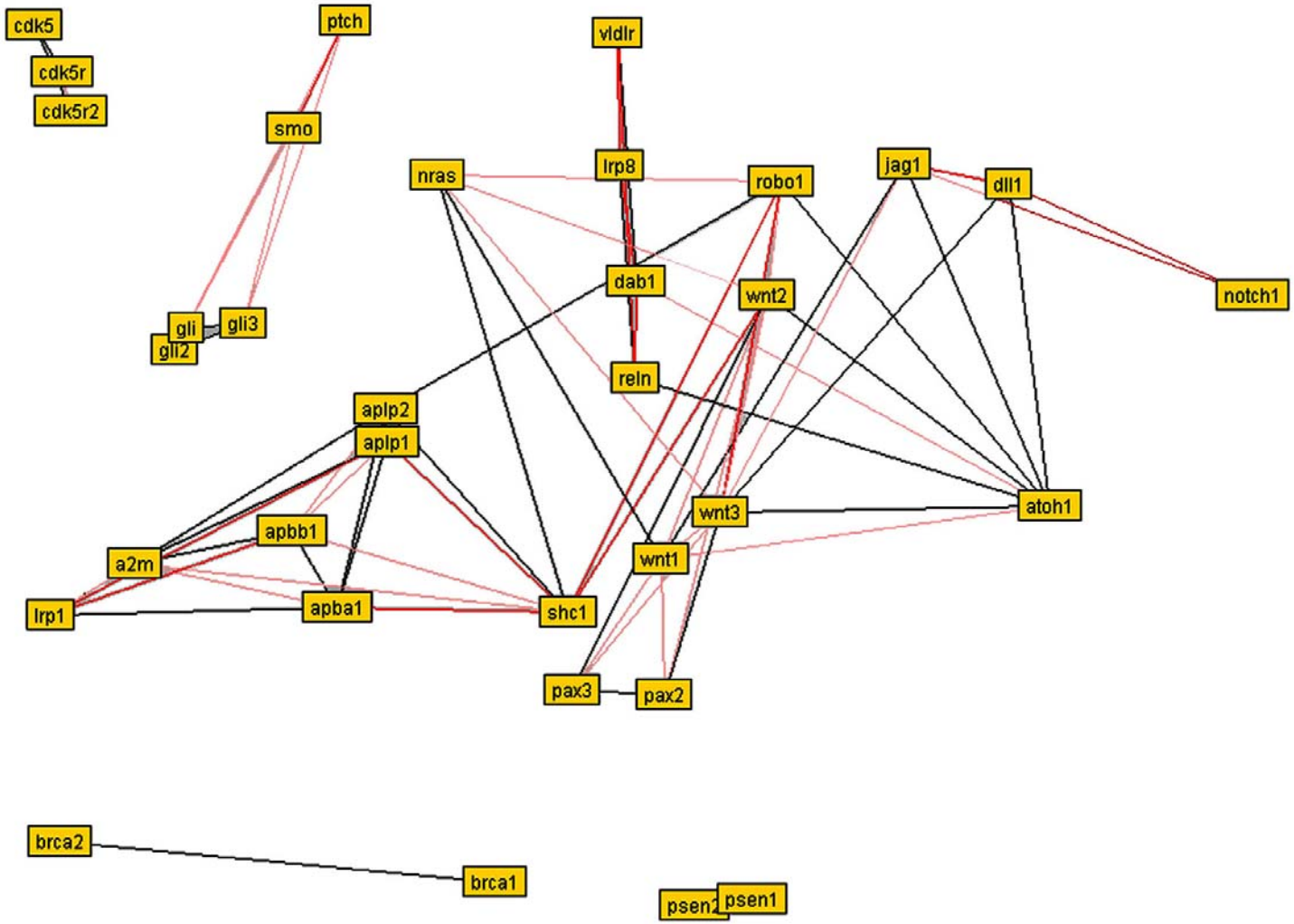
# 50-Gene Document Collection



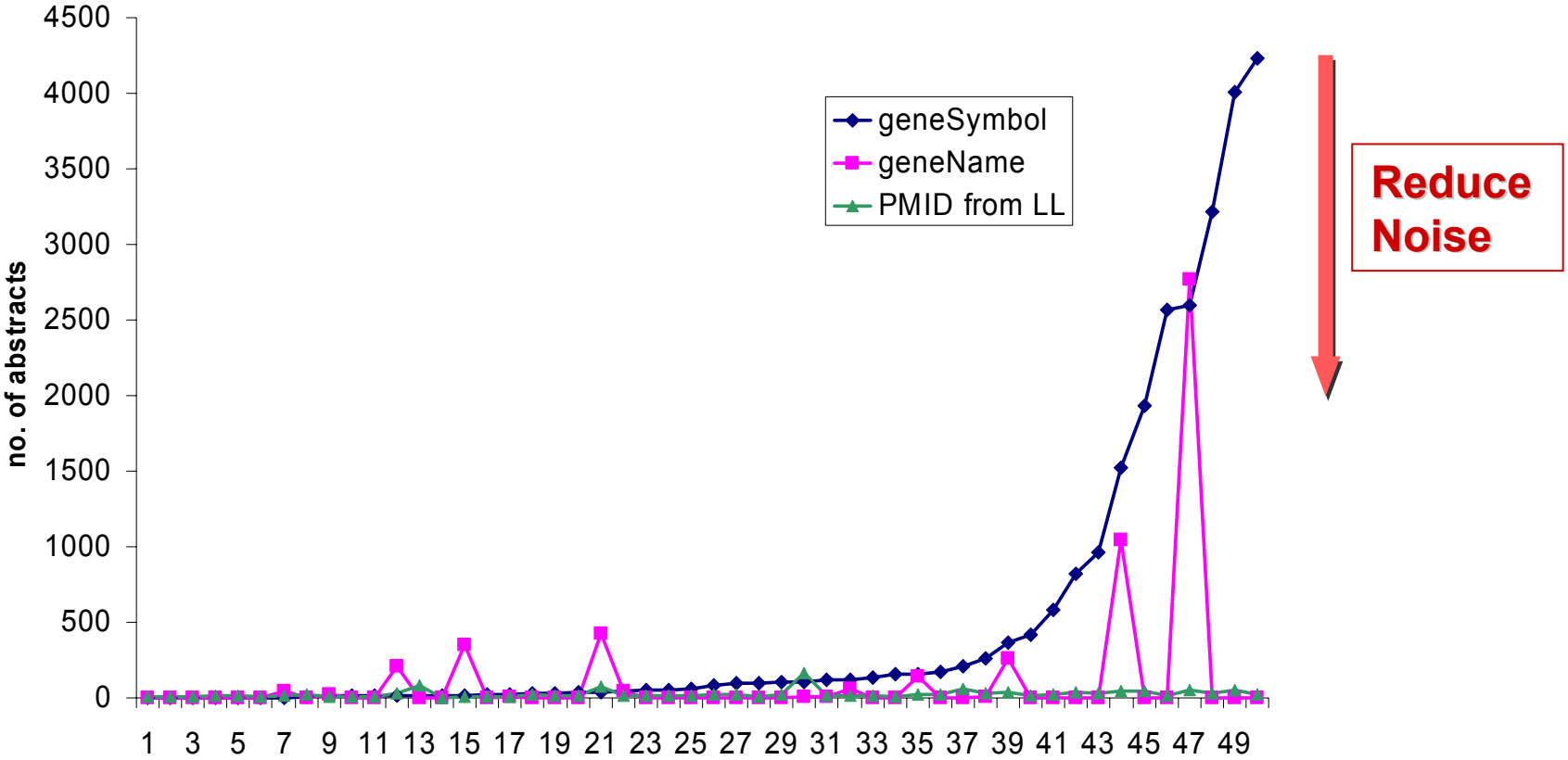
# Hierarchical Tree



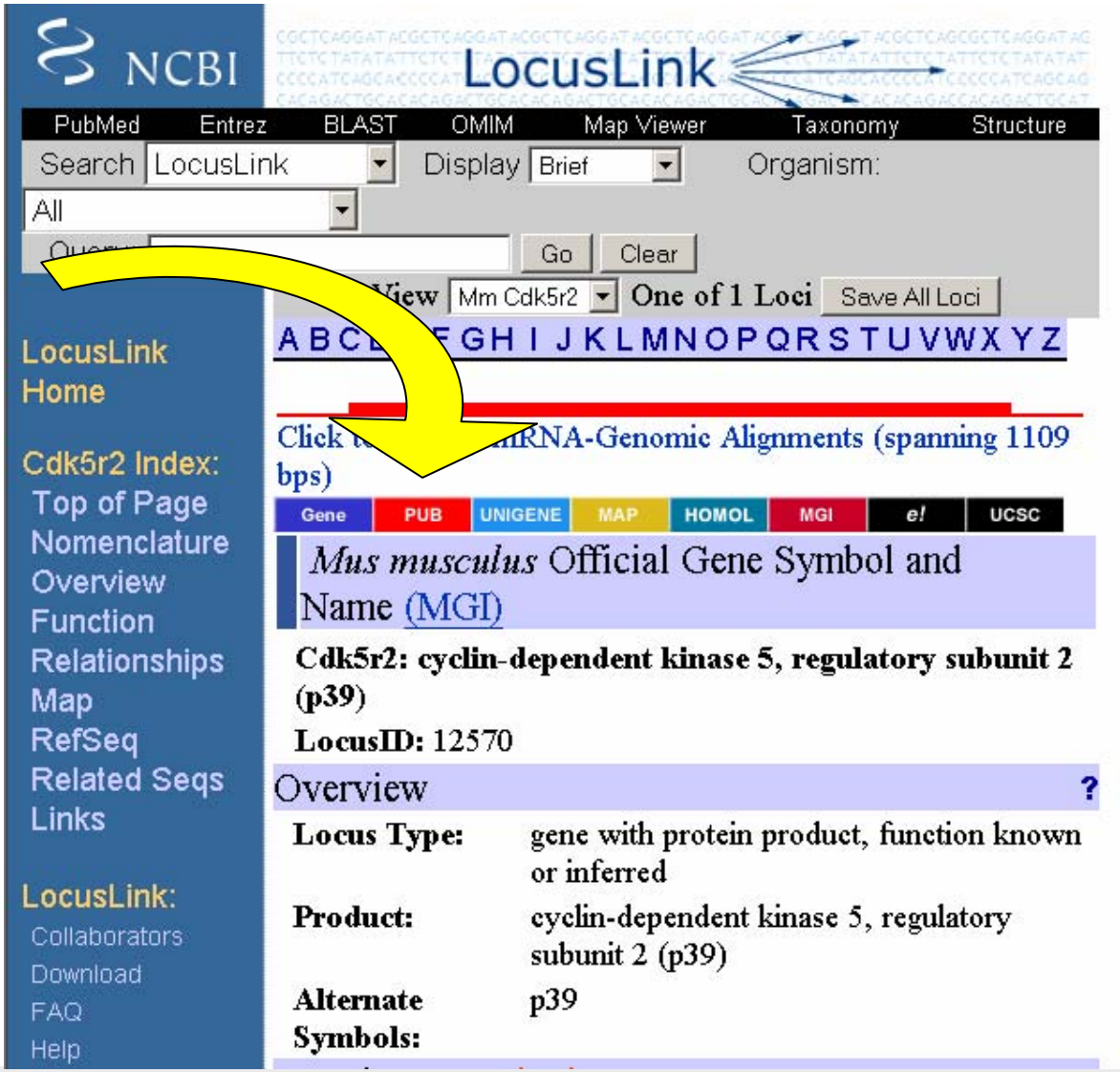
# Unrooted Tree



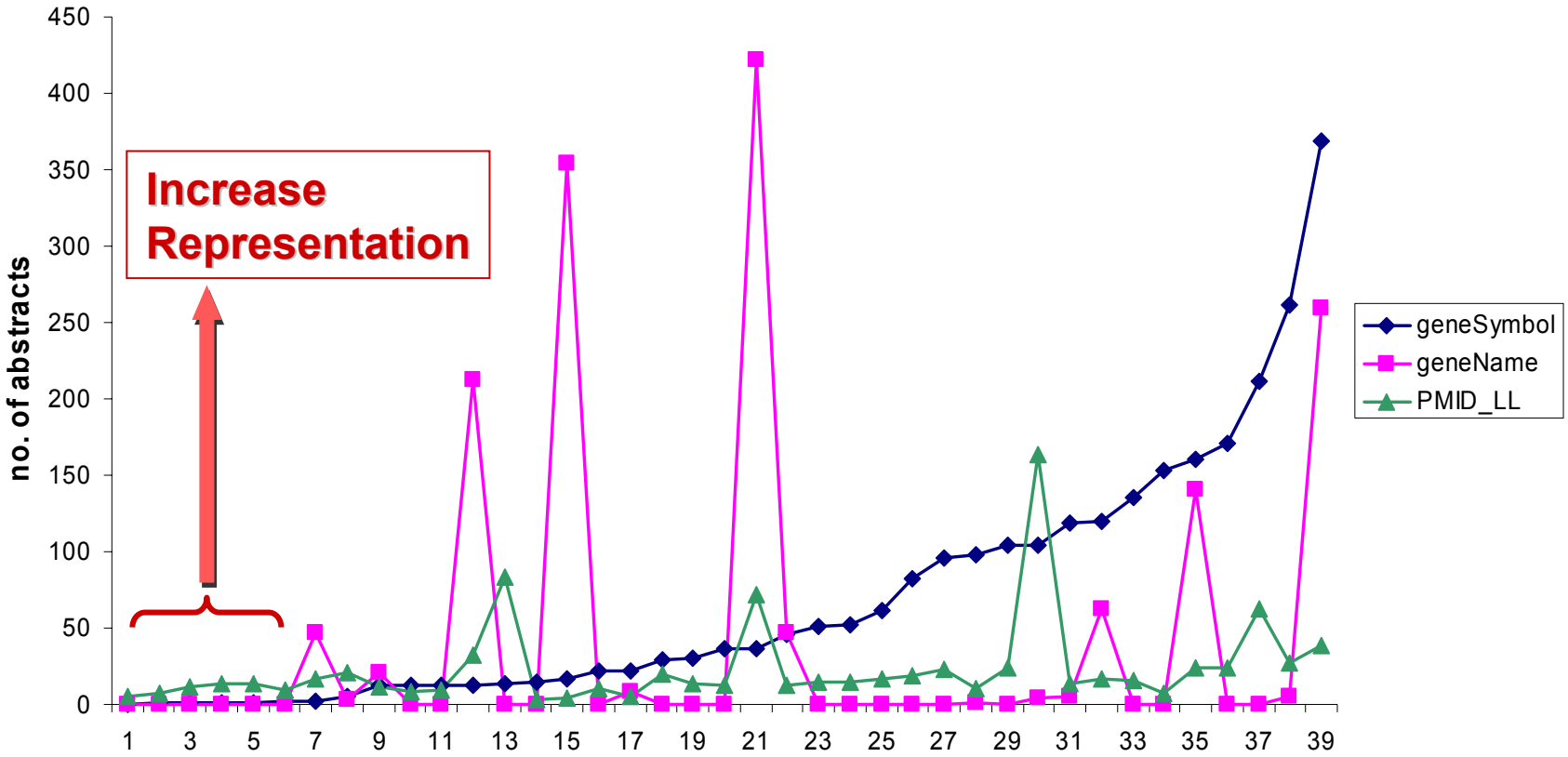
# Variation in Abstract Representation



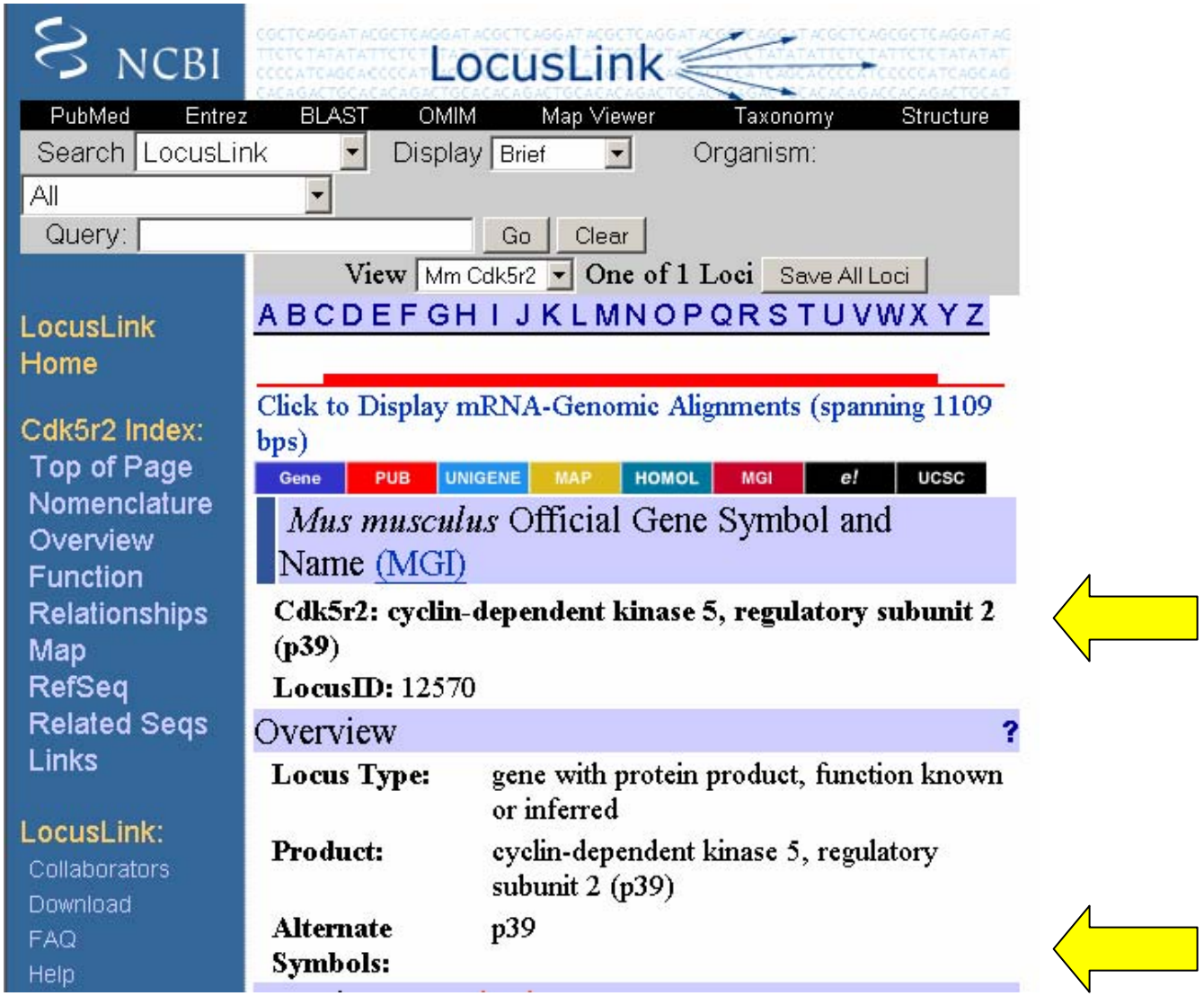
# Abstract References in LocusLink



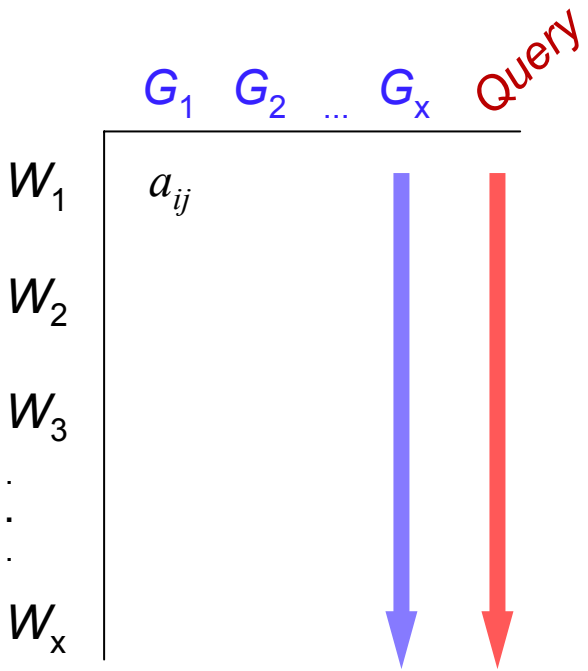
# Gene symbols and names that are not used in the literature



# Alternate Names and Aliases



# Log-entropy Term Weighting



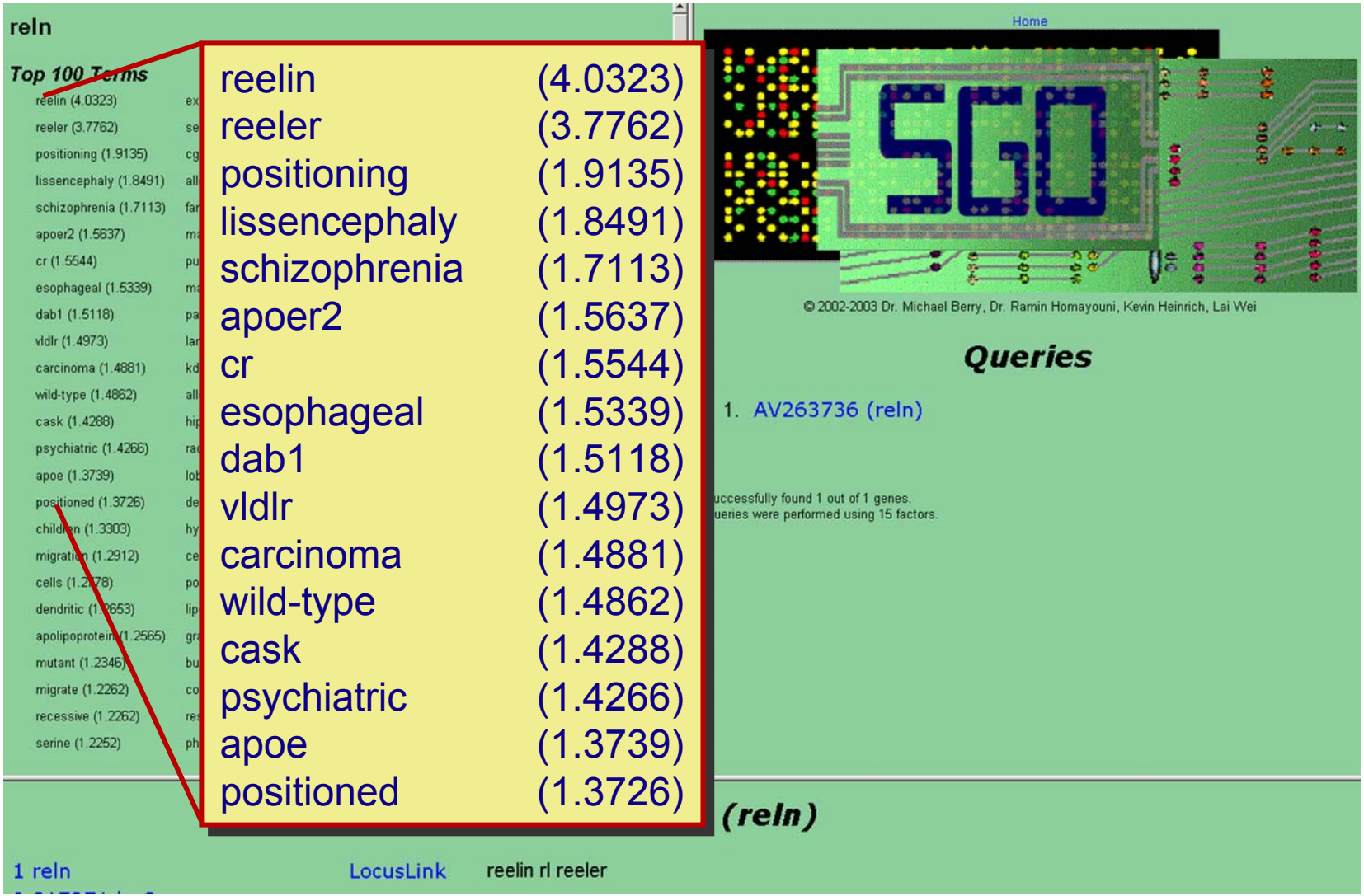

$$a_{ij} = l_{ij}\, g_i$$

$$l_{ij} = \log_2\left(1 + f_{ij}\right)$$

$$g_i = 1 + \left( \frac{\sum_j \left(p_{ij} \log_2 \left(p_{ij}\right)\right)}{\log_2 n} \right)$$

$$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}}$$

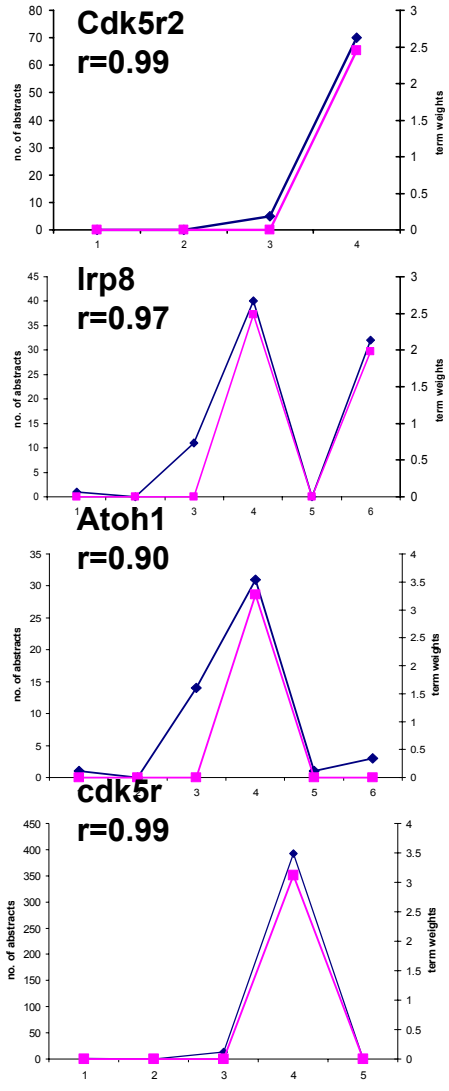
# Top Terms in Gene Document



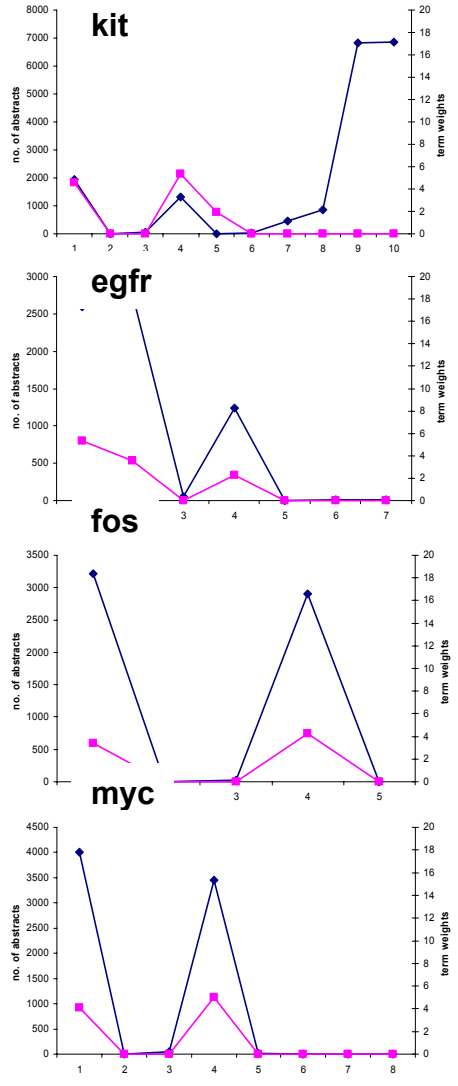

| | |
|---|---|
| reelin | (4.0323) |
| reeler | (3.7762) |
| positioning | (1.9135) |
| lissencephaly | (1.8491) |
| schizophrenia | (1.7113) |
| apoer2 | (1.5637) |
| cr | (1.5544) |
| esophageal | (1.5339) |
| dab1 | (1.5118) |
| vldlr | (1.4973) |
| carcinoma | (1.4881) |
| wild-type | (1.4862) |
| cask | (1.4288) |
| psychiatric | (1.4266) |
| apoe | (1.3739) |
| positioned | (1.3726) |

# Terms with high log-entropy weights retrieve more abstracts by PubMed query

## Under-represented Genes

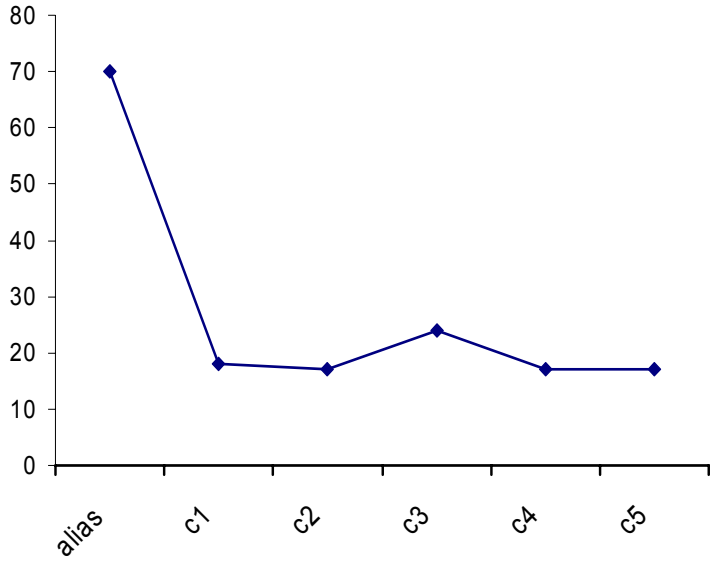


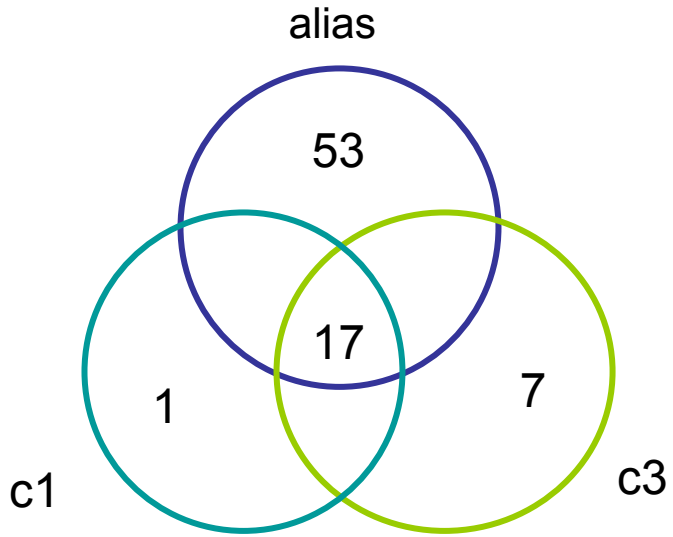

## Over-represented Genes



# Abstract retrieval by combining weighted terms in gene name, symbol or aliases

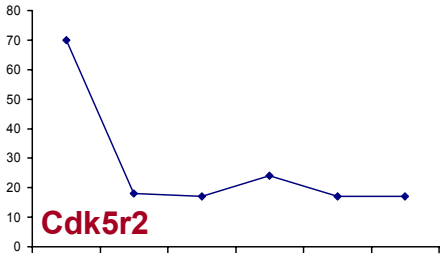
| Query | Description | # abstracts |
|---|---|---|
| symbol | Cdk5r2 | 0 |
| alias | p39 | 70 |
| name | cyclin-dependent kinase 5, regulatory subunit 2 | 0 |
| c1 | p39 AND cdk5 | 18 |
| c2 | p39 AND cyclin-dependent | 17 |
| c3 | p39 AND kinase | 24 |
| c4 | p39 AND cdk5 AND cyclin-dependent | 17 |
| c5 | p39 AND cdk5 AND cyclin-dependent AND kinase | 17 |

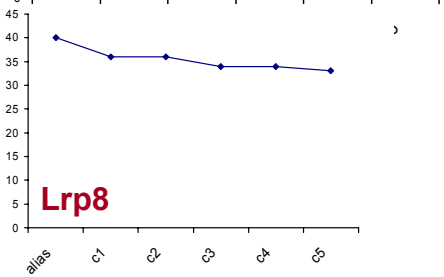


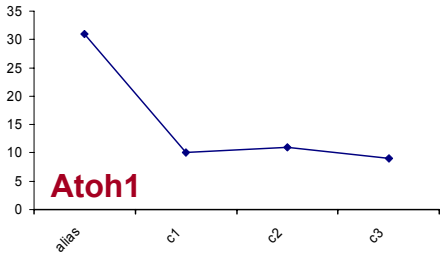



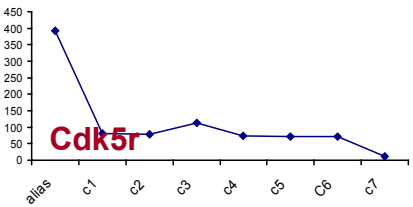
# Weighted Combinatorial Queries

## Under-represented Genes

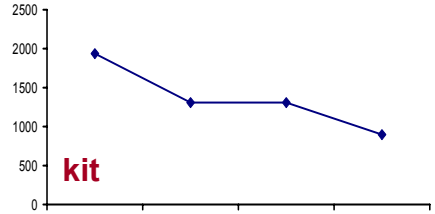


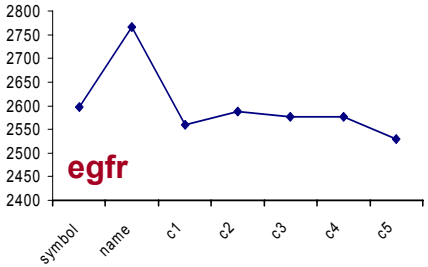



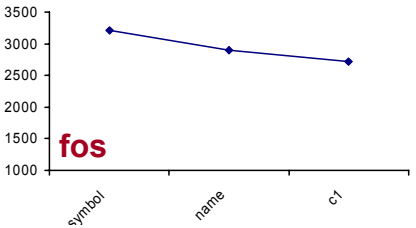



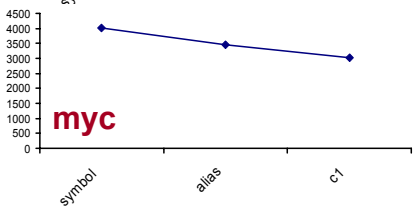




## Over-represented Genes

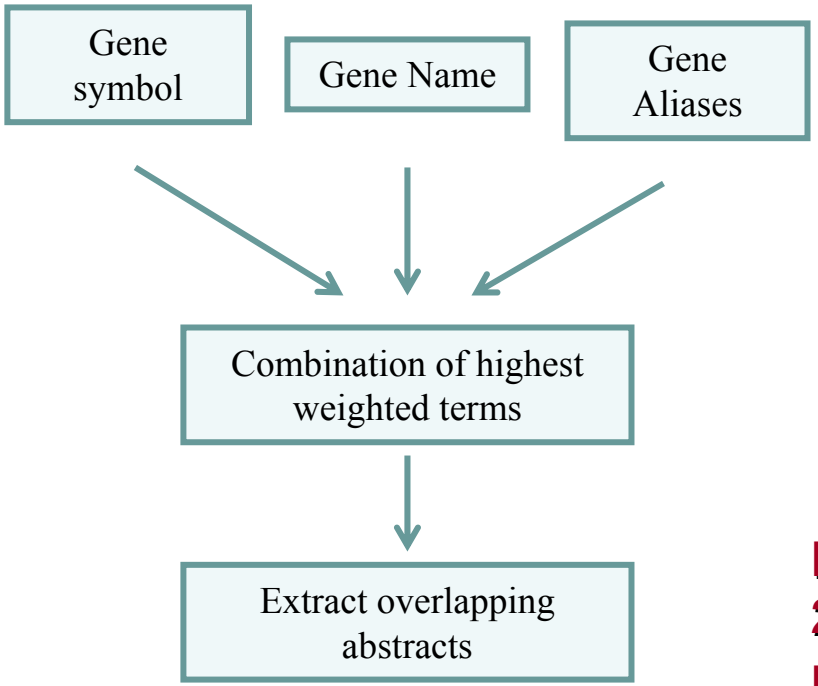


# Combinatorial Query Algorithm




**RESULTS:**
**2-59 fold increase in the number of abstracts associated with genes compared to those referenced in LL**

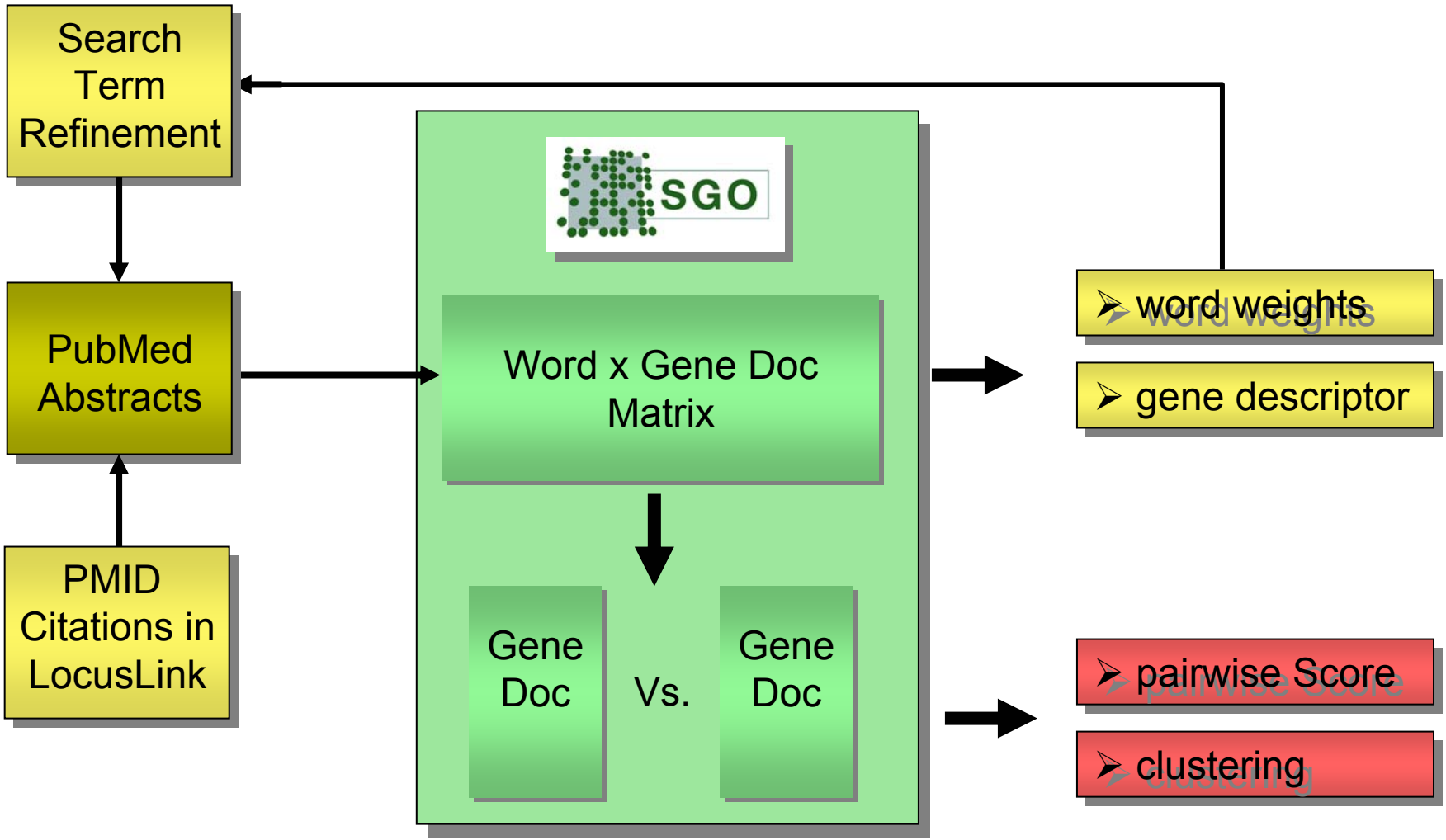
# Summary and Conclusions

Log-entropy weighting identifies descriptive or ‘useful’ aliases for genes.

Weighted Combinatorial Querying increases abstracts for under-represented genes and decreases abstracts for over-represented genes with high specificity.

This automated method improves gene abstract assignment 2 to 59 fold beyond those assigned by LocusLink indexers.

# SGO overview



# Acknowledgments


## UT Memphis

**Neurology**

Lijing Xu, M.S.

Lai Wei, M.D.

**Molecular Sciences**

Yan Cui, Ph.D.

Mi Zhou, M.S.

## UT Knoxville

**Computer Science**

Michael Berry, Ph.D.

Kevin Heinrich