

---

# Automatic Annotation of Gene Lists from Literature Analysis

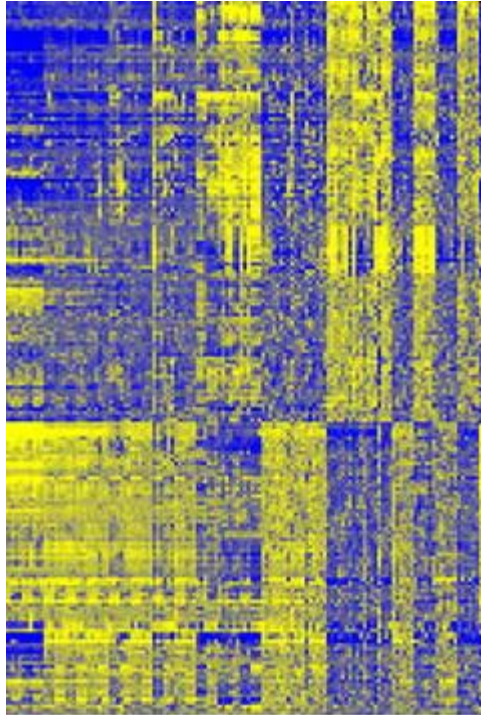
---

Xin He

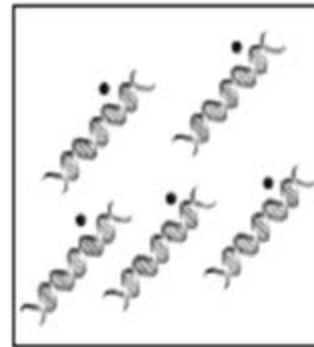
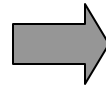
Beespace Annual Workshop

05/21/2009

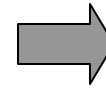
# Annotating Gene Lists



Gene expression profiles for 132 individual dissected honey bee brains (columns).



Gene group



# Enrichment of Gene Ontology Terms

GO ID	Level	GO Term	In the background		In the given gene list		P-value
			Reference Occ.	Reference Freq.	Dataset Occ.	Dataset Freq.	
GO:0007424	5	<a href="#">tracheal system development (sensu Insecta)</a>	72	0.0095	4	0.0851	0.0008918
GO:0006811	6	<a href="#">ion transport</a>	381	0.0501	8	0.1702	0.0016169
GO:0007165	4	<a href="#">signal transduction</a>	1144	0.1503	15	0.3191	0.0018011
GO:0007154	3	<a href="#">cell communication</a>	1451	0.1906	17	0.3617	0.0027301
GO:0016055	6	<a href="#">Wnt receptor signaling pathway</a>	52	0.0068	3	0.0638	0.0036685
GO:0050801	4	<a href="#">ion homeostasis</a>	17	0.0022	2	0.0426	0.0046440
GO:0030003	7,6	<a href="#">cation homeostasis</a>	17	0.0022	2	0.0426	0.0046440
GO:0009586	8,9,10	<a href="#">rhodopsin mediated phototransduction</a>	17	0.0022	2	0.0426	0.0046440
GO:0006873	6,5	<a href="#">cell ion homeostasis</a>	17	0.0022	2	0.0426	0.0046440
GO:0007222	7	<a href="#">frizzled signaling pathway</a>	18	0.0024	2	0.0426	0.0051935
GO:0006101	3,6	<a href="#">malate metabolism</a>	1	0.0001	1	0.0213	0.0061753
GO:0030001	7	<a href="#">monovalent inorganic cation homeostasis</a>	1	0.0001	1	0.0213	0.0061753
GO:0006885	10,9	<a href="#">regulation of pH</a>	1	0.0001	1	0.0213	0.0061753
GO:0050918	8,7	<a href="#">positive chemotaxis</a>	1	0.0001	1	0.0213	0.0061753
GO:0046341	7,8	<a href="#">CDP-diacylglycerol metabolism</a>	1	0.0001	1	0.0213	0.0061753
GO:0030641	9,8	<a href="#">hydrogen ion homeostasis</a>	1	0.0001	1	0.0213	0.0061753

Ckl1<sup>alpha</sup>  
nkd

Enrichment test based on these numbers



---

# Limitations of GO Analysis

- GO annotations of all genes involve substantial manual efforts
  - Rapid growth of literature: constantly add new functions to existing genes
  - Coverage is not even in all areas. E.g. ecology and behavior; medicine; anatomy and physiology; etc.
-

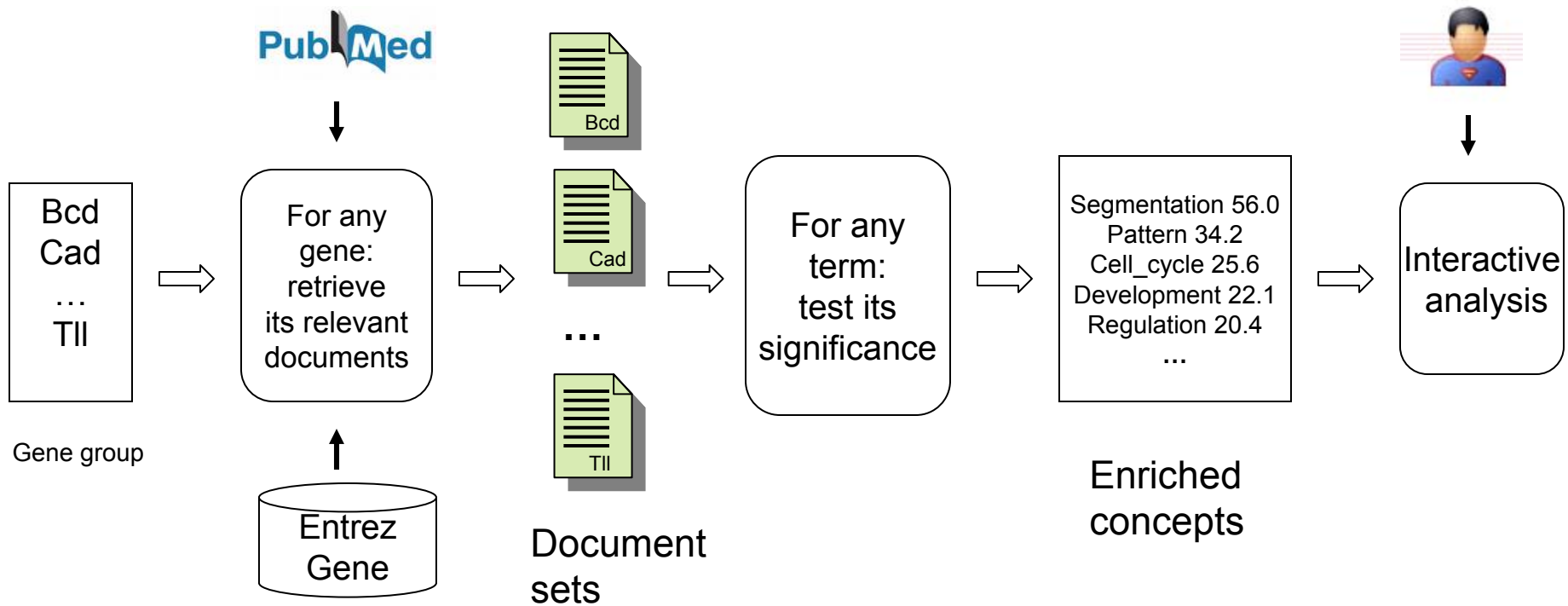
# Literature-based Analysis

- Gene-term matrix: the count of terms in the documents of a gene.

Gene	TPI1	GPM1	PGK1	TDH3	TDH2
protein_kinase	0	0	2	0	0
→ decarboxylase	10	0	10	7	6
● protein	39	26	65	44	33
→ stationary_phase	2	7	3	4	2
→ energy_metabolism	4	5	5	8	0
oscillation	0	0	0	0	1

- Enrichment of terms: if a term is associated with many genes in the input list, this term is likely important for this list.
- Need to account for the expected term occurrences by chance: a term may occur in a gene, but not important.

# Overview of Gene List Annotator





---

# Document Retrieval for Genes

- Input: a list of gene identifiers
    - Yeast: SGD ids
    - Fruit fly: FlyBase ids
    - Mouse: MGI ids
  - Mapping genes to synonyms: use Entrez Gene database (manually created synonyms)
  - Document collection: choose or create one from Beespace
  - Retrieve documents in the collection that match at least one synonym
-

# Statistical Method (I)

**Problem:** given the following information about a term:

$x_1, \dots, x_n$ : the number of occurrences in the documents of each gene;

$d_1, \dots, d_n$ : the length of the document set of each gene;

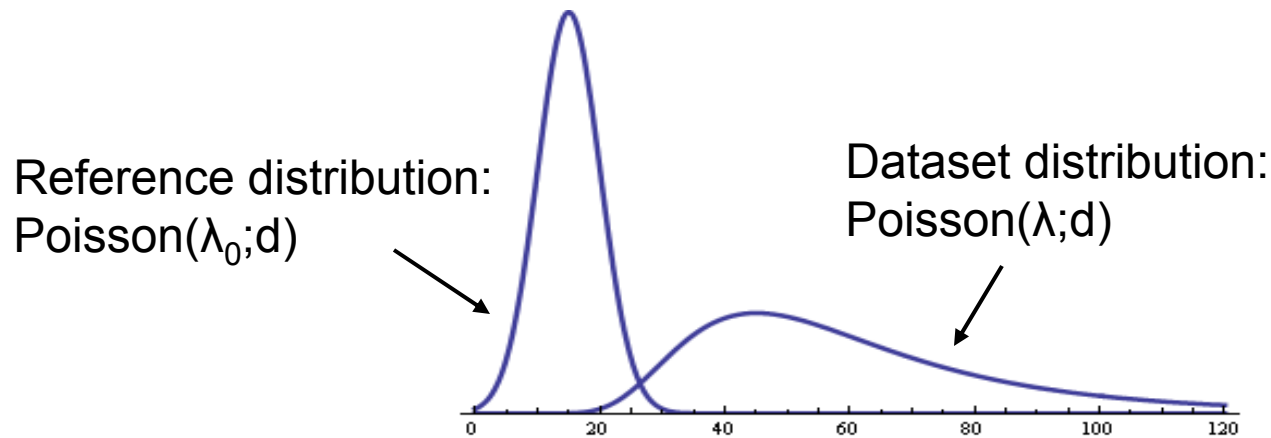
$\lambda_0$ : the frequency of the term in the whole collection (background).

Test the enrichment of this term in the gene list.

**Intuition:**

- 1) For a gene  $i$ , if the term count  $x_i$  is significantly higher than expected by chance (determined by  $\lambda_0$  and  $d_i$ ), then the term may be related to the gene  $i$ ;
- 2) If there are many genes related to the term, then this term is enriched in the given gene list.

# Statistical Method (II)



**Model:** whether a gene is related to the term is unknown, so assume the term count  $x_i$  follows the mixture of two Poisson distributions.

**Likelihood ratio test:** on the observed term counts, mixture distribution vs null distribution (reference distribution only)

# Interactive Analysis (I)

Output control

**Gene Annotation Results v4**

Filter Genes:  No Gene Filtering  Keep Genes  Remove Genes  
Filter:  No Phrase  Keep  Remove Phrases  
Phrases: Filtering Phrases  
Sort By:  Significance  Ratio  Concept  Cluster

Note: Ratio is the proportion of genes in the input list that the concept is associated with. Significance is the likelihood ratio score of the concept; larger significance corresponds to lower p-value. Only the concepts with p-values less than .05, after Bonferroni correction, are shown.

Significant Concepts					Genes Found			
<input type="checkbox"/>	#	<u>Concept</u>	<u>Ratio</u>	<u>Significance</u>	<u>Cluster</u>	Name	Gene	Found
<input type="checkbox"/>	0	<a href="#">mw</a>	0.08053	153.6	1	Fasn	95485	36
<input type="checkbox"/>	1	<a href="#">er</a>	0.09297	65.2	4	Tg	98733	20
<input type="checkbox"/>	2	<a href="#">grp78</a>	0.10062	60.5	1	Psme3	1096366	15
<input type="checkbox"/>	3	<a href="#">midline</a>	0.08671	56.7	5	Mink1	1355329	9
<input type="checkbox"/>	4	<a href="#">hsp70</a>	0.10064	52.2	1	Nxf1	1858330	8
<input type="checkbox"/>	5	<a href="#">hfp</a>	0.07800	51.8	9	Elovl6	2156528	4
						Por	97744	2

Choose concepts

Significant Concepts

Relevant Statistics

Information of Input Genes

# Interactive Analysis (II)

User-selected  
concepts

**Concept-Gene Analysis**

*Note : This is a concept-by-gene co-occurrence matrix. The integer entries represent the number of times that a gene co-occurs with a concept.*

<u>Gene Name</u>	<u>Gene ID</u>	light chain	heavy chain	thick filament	nonmuscle myosin	myosin ii
Mhc	FBgn0002741	<u>201</u>	<u>188</u>	<u>119</u>	<u>56</u>	<u>56</u>
jar	FBgn0011225	<u>39</u>	<u>147</u>	<u>56</u>	<u>22</u>	<u>12</u>
Strn-Mlck	FBgn0013988	<u>89</u>	0	0	<u>2</u>	<u>10</u>
Prm	FBgn0003149	0	<u>6</u>	<u>57</u>	0	0
sls	FBgn0003432	<u>8</u>	<u>1</u>	<u>10</u>	0	0
Bsg	FBgn0011219	<u>33</u>	<u>13</u>	0	<u>9</u>	<u>7</u>
Fas3	FBgn0000636	<u>1</u>	0	0	<u>1</u>	<u>1</u>
Side	FBgn0032741	<u>4</u>	<u>3</u>	<u>2</u>	0	0
Cg25C	FBgn0000299	0	<u>2</u>	<u>2</u>	0	0
Hn	FBgn0001208	0	<u>2</u>	0	0	0
nrv2	FBgn0015777	0	0	0	<u>1</u>	<u>1</u>

Genes containing the selected concepts

Term counts in genes,  
and link to documents

---

# Applications

- Test case 1. bee genes differentially expressed in brain in different species during behavior maturation
    - Broadly consistent with the results from GO enrichment analysis
    - Identify interesting genes
  - Test case 2. bee genes up-regulated in brain by the methoprene treatment (inducing behavior maturation)
    - GO enrichment analysis: no significant terms
    - A theme about myosin is overrepresented: may suggest neuron growth and movement, or remodeling, during behavior maturation
  - See Beespace v4 Demo for details: 1pm, Friday
-



---

# Summary

- Not limited to a controlled vocabulary (GO)
  - Even for concepts covered by GO, a broader notation of term relevance (gene-term co-occurrence in literature)
  - Possible to retrieve the supporting documents for further exploration
  - Not meant to substitute GO-based analysis, but a complementary tool
-

---

# Acknowledgement



Bruce Schatz



Chengxiang Zhai



Gene Robinson

Software support: Xu Ling, Jing Jiang, Brant Chee, David Arcoelo

Biological evaluation: Moushumi Sen Sarma, Amy Toth

---