Standards in Proteomics
Protein-Protein Interactions

Christopher W.V. Hogue, Ph.D.
Scientist, Samuel Lunenfeld Research Inst. Mount Sinai Hospital
Associate Professor, Dept. of Biochemistry, U. of Toronto

MOUNT SINAI HOSPITAL

U of T

NUS National University of Singapore

Blueprint

Outline

- OECD Declaration – Data Access/Archiving
- About Blueprint and BIND
- On Standards, Components and Systems
- Reverse Salients
- Communities
- User Behavior
- OntoGlyphs - Visualizing Complex Information
- BIND PICKS – Scoring Yeast Interactions
- New BIND Record Types

OECD

Organisation for Economic Co-operation and Development (oecd.org)

DECLARATION ON ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING

adopted on 30 January 2004 in Paris

The governments (1) of Australia, Austria, Belgium, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Russian Federation, the Slovak Republic, the Republic of South Africa, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States

Declare their commitment to

- Work towards the establishment of access regimes for digital research data from public funding in accordance with the following objectives and principles:

- Openness
- Transparency
- Legal Conformity
- Formal Responsibility
- Professionalism
- IP Protection
- Interoperability
- Quality and Security
- Efficiency
- Accountability

Q13. Archive? Form? Where?

- OECD Declaration in Canada was disseminated top-down to the granting agencies in Nov from the government.
- Mandate to sort out the problem of archiving all scientific research data.
- Timeline in 5 years – inter-agency coordination required.

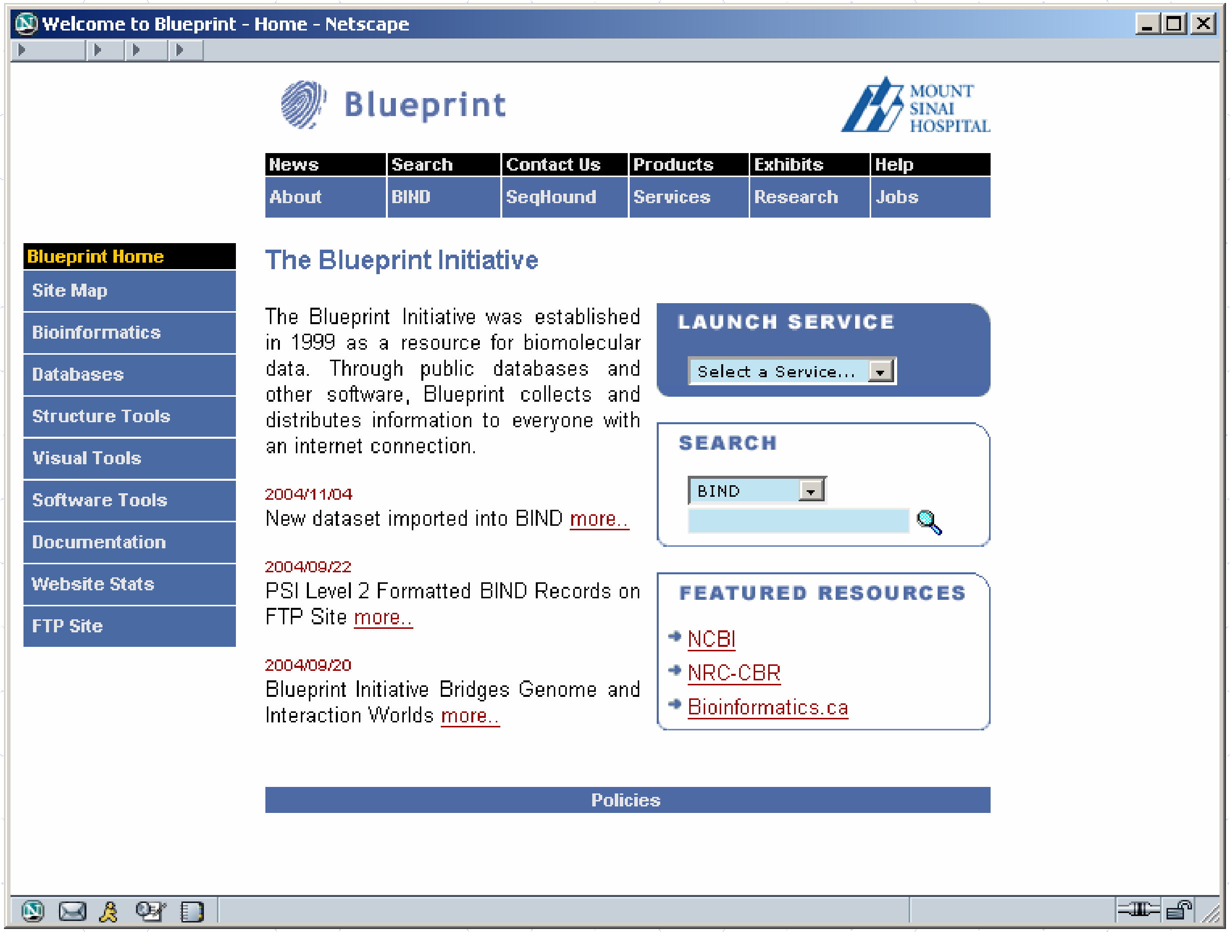
A tea cup in a rainstorm…

- 2000 elemental observations (facts) about molecular assembly (interactions) published in the literature every month.
- By 1965 standards - 10 textbooks full of figures, descriptions, mechanisms every year
- Estimate 200,000 facts sitting in the literature on library shelves, not validated, not useable.
- Blueprint’s mandate is to capture this information in a machine readable database called BIND.

The Blueprint Initiative

- Develop, curate and maintain the Biomolecular Interaction Network Database (BIND) and related tools
- Carry out bioinformatics research in support of our vision.

blueprint.org



The Blueprint Initiative

The Blueprint Initiative was established in 1999 as a resource for biomolecular data. Through public databases and other software, Blueprint collects and distributes information to everyone with an internet connection.

2004/11/04
New dataset imported into BIND more..

2004/09/22
PSI Level 2 Formatted BIND Records on FTP Site more..

2004/09/20
Blueprint Initiative Bridges Genome and Interaction Worlds more..

MOUNT SINAI HOSPITAL

Blueprint

Biomolecular Interaction Network Database

v3.6
[Spec: v3.2]

info@bind.ca

January 4, 2005

▶ BIND Home

- Blueprint Home
- >BIND
- About
- Curation
- Development
- High Throughput
- FTP/Downloads
- Resources
- News
- Featured Submitter
- Credits
- Publications
- FAQ
- Help

Search:

Field Specific Search | BINDBlast | PreBIND | SMID

Browse the contents of the BIND database.

Search BIND data using an identifier, such as PubMed Id, GenInfo Id, PDB Id, GO Id, and more...

Search BIND data using a simple text query.

Search BIND data by building a field specific query.

Submit your interaction data to BIND.

Submit your molecular complex data to BIND.

BIND database general and detailed statistics.

Contact us with your questions and comments.

BIND Statistics

Interactions:	121808
Complexes:	2416
Pathways:	8

Detailed Statistics...

Recently Added Records

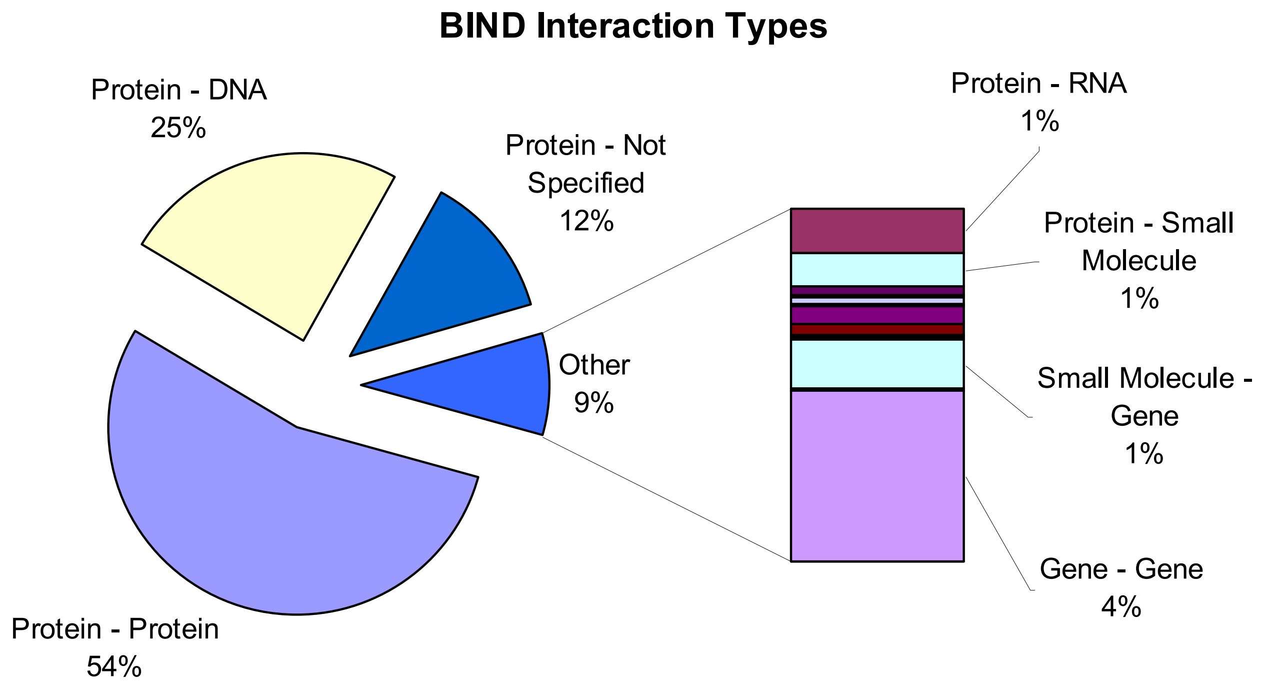
View Today's New Records	18
View Yesterday's New Records	4
View This Week's New Records	93
View This Month's New Records	1420

Identifier Search

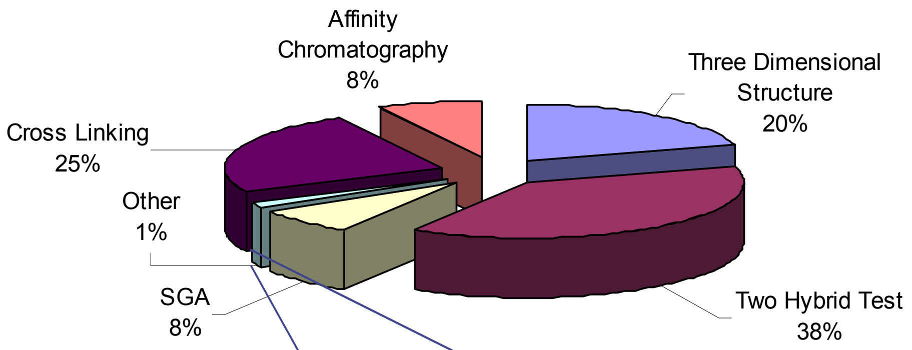
BIND Id

More Options...

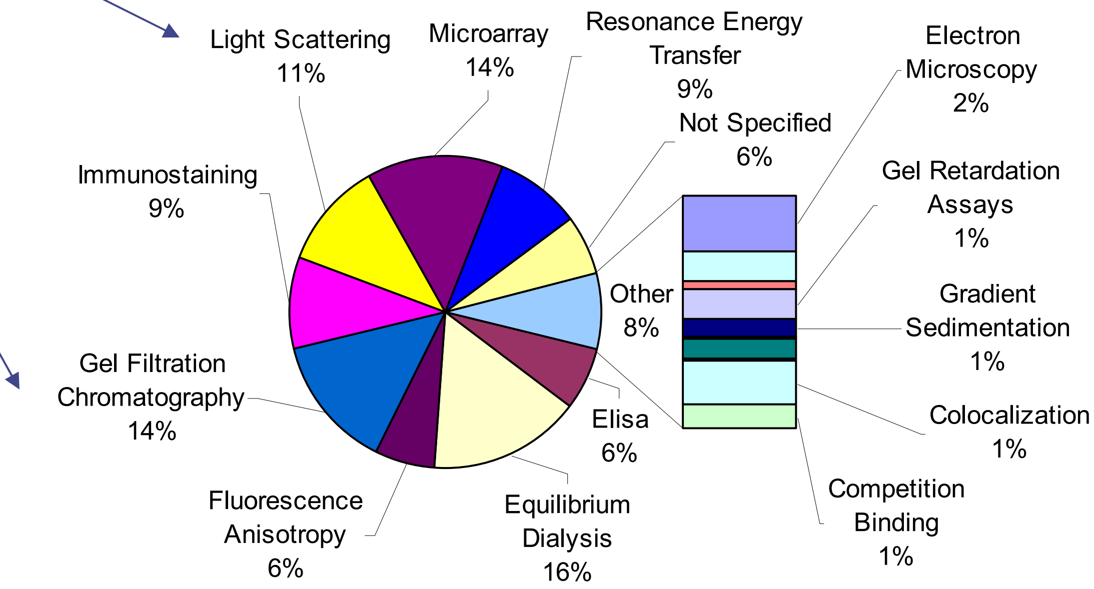
BIND stores molecular interaction data…



Interaction Experimental Evidence Captured



Interaction Experimental Evidence Captured



Standards in Proteomics?

- Software Systems Components (OSI Layers…)
    - Human Interfaces
    - Application Programming Interfaces
    - Communications Protocols
    - Content Structure ← ?
    - Database (ODBC/JDBC compliant MySQL)
    - Document Structures (XML)
    - Architectures (Compatible orchestration of the above)
    - Platforms (Runs the above: Windows, Linux, Unix)

What is in a "Standard" A Historical Perspective

- Standards emerge from successful systems.
- Which one is "the standard" The light bulb – or the electrical grid?
- Lamps were the original killer app.
    - (bye-bye candles, gas lamps, oil lamps)
  - Other Apps: Motors, Heaters, Toasters
  - Unexpected Apps: radio, TV, transformers, computers, rechargables
- Entire "systems" become standards via ad-hoc and popular use – snowball effect.

Emergence and evolution of technological systems…

- Systems emerge across broad frontiers
    - Lots of small inventions are responsible for emerging techologies.
- Portions of the frontier that are held back become the focus of intense innovation
    - Called a "reverse salient" by students of technology
    - An inadequately functioning or accessible component in a complex system of of components
    - Opportunities for invention and replacement

Reverse Salient – AC/DC Example

- Edison’s DC standard lit up Wall Street in Manhattan
- High-level buy-in for DC.
- AC was too complicated, could kill a person!
- Edison’s DC system only worked over short-range.
- This flaw is the “reverse salient”.
- Westinghouse/Stanley/Tesla saw the flaw in this standard
- AC technology raced to fill the gap
- Light bulbs work with both AC or DC.
- Motors required re-invention
- E.S. Rogers “batteryless” radio

Reverse Salient – AC/DC Example

- Edison's DC standard lit up Wall Street in Manhattan
- High-level buy-in for DC.
- AC was too complicated, could kill a person!
- Edison's DC system only worked over short-range.
- This flaw is the "reverse salient".
- Result: Cars, Battery based devices emerged with DC.

- Westinghouse/Stanley/Tesla saw the flaw in this standard
- AC technology raced to fill the gap
- Light bulbs work with both AC or DC.
- Motors required re-invention
- E.S. Rogers "batteryless" radio
- Result: The electrical Grid emerged with AC.

NOT A WINNER-TAKE-ALL RESULT!

Reverse Salient Attitudes

- What holds us back?
  - Oversights (didn't think of that!).
  - Shortsightedness (won't ever need that!).
  - Inability (can't do it!)
  - Stubbornness (won't do it!)
  - Prescriptivism (do it like this!)
  - Nationalism, Continentalism, Colonialism
    - (because that's the way we do it here!)
    - 110 vs 220

Q1. Communities – who matters?

- A1. What Standard? What community?
  - Database Communities
    - IMEX (BIND/DIP/INTACT/MINT/MIPS)
    - BioPAX (pathway databases)
    - SBML (>70 software systems collaborating)
    - Cytoscape (collaborating interface developers)
    - NCBI/Blueprint (architecture)
    - Model Organism Databases (GMOD architecture)
  - Journals and Editors
  - Scientific Societies (HUPO)
  - Member and Non-member Scientists

Q7. Integrate/Synergize

- Identify the communities.
- Recognize that communities are disjoint.
- Require funded efforts to show their efforts to collaborate with and integrate across the spectrum of identified communities.

Q3? Standard Improvement?

- Service all communities effectively with a whole system.
- Drive innovation more through use.
- Gain and effectively incorporate user critique.
- Understand user needs, behaviors.

User Behavior

- The problem of too much choice.
  - (M. Lepper @Stanford and S. Iyengar @Columbia)
- Two tables in a supermarket:
  - 24 jars of jam vs 6 jars of jam.
  - 3% vs 30%
- Choice frustration.

User Behavior

- The problem of too much choice.
    - (M. Lepper @Stanford and S. Iyengar @Columbia)
- Two tables in a supermarket:
    - 24 jars of jam vs 6 jars of jam.
    - 3% vs 30%
- Choice frustration.
- Leads to incrementalism
- Essential user criticism is withdrawn.
        - Can’t Debug - This jam is a little bitter compared to
            - the other 6?
            - the other 26?
        - A whole lot of bad jam that nobody wants to buy…

Q4? Main Problems?

- Standards Fatigue
  - Interactions/Pathways since NIH meeting in Nov 1999. Efforts are still not integrated (PSI/IMEX and BIOPAX).
- Data Standards are not an effective goal to achieve results in a timely way
- Information Systems are better goals.
- Wet Lab Scientists are busy people who are (excuse me) trying to write papers.
- Ongoing wishful thinking about latest new technology (the semantic web will fix everything!)

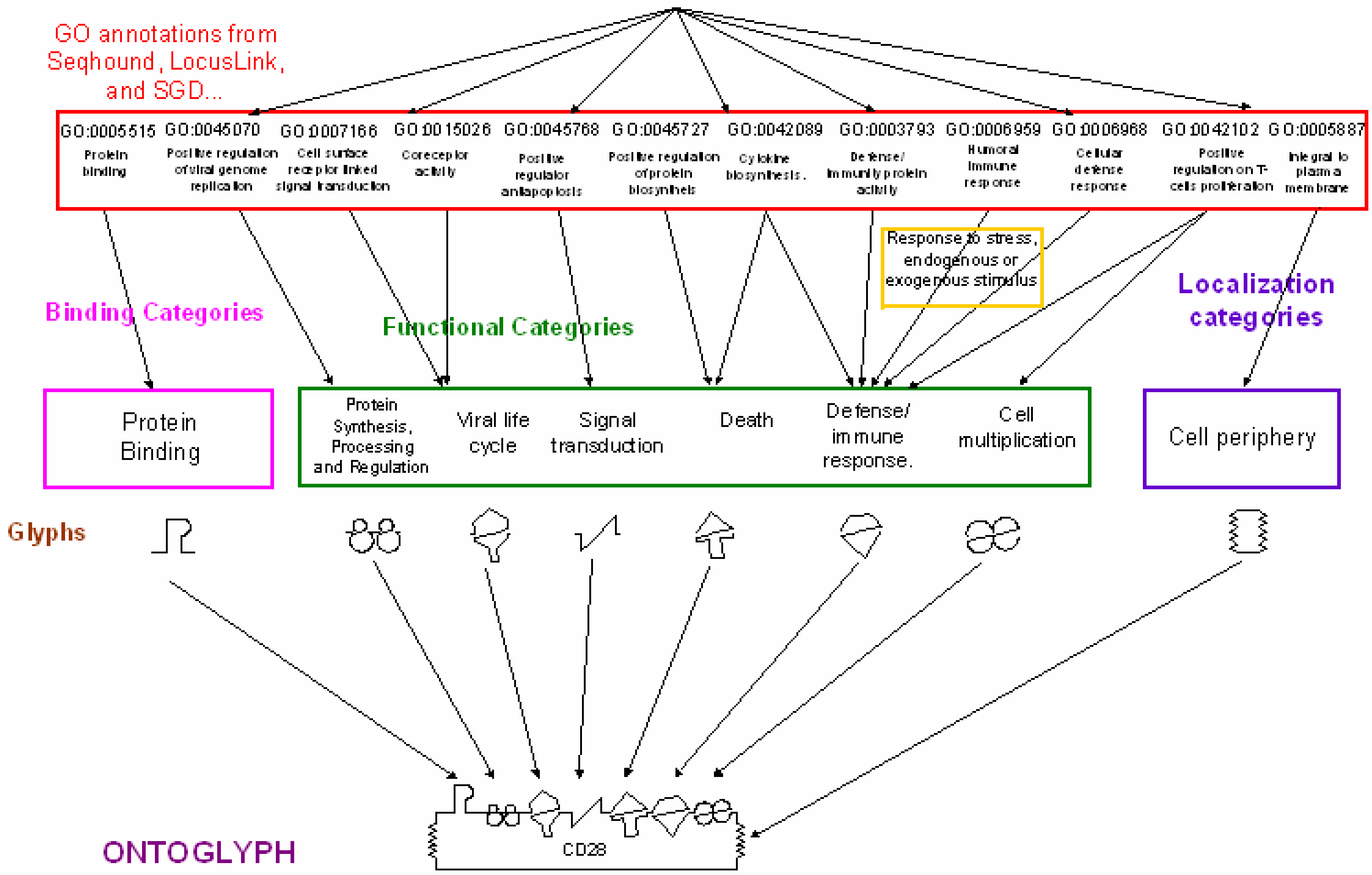
Q8. Proteomic Dictionary (semantics)

- Isn’t that what Gene Ontology is – a collection of terms about proteins?
- >17,000 terms – ½ the size of a pocket dictionary.
- Structured data curation vs term-tagging. Both are required.

OntoGlyphs

- A graphical language
- Derived from Gene Ontology annotation
- The most-used terms/categories
- Simpler – Fewer Choices
- Summarize Long, long long long long long long long long long long lists of data results – like query “cancer”

GI:5453611, Name: CD28



GO annotations from Seqhound, LocusLink, and SGD...

GO:0005515 Protein binding

GO:0045070 Positive regulation of viral genome replication

GO:0007166 Cell surface receptor linked signal transduction

GO:0015026 Coreceptor activity

GO:0045768 Positive regulator antiapoptosis

GO:0045727 Positive regulation of protein biosynthesis

GO:0042089 Cytokine biosynthesis.

GO:0003793 Defense/ Immunity protein activity

GO:0006959 Humoral Immune response

GO:0006968 Cellular defense response

GO:0042102 Positive regulation on T-cells proliferation

GO:0005887 Integral to plasma membrane

Response to stress, endogenous or exogenous stimulus

Binding Categories

Functional Categories

Localization categories

Protein Binding

Protein Synthesis, Processing and Regulation

Viral life cycle

Signal transduction

Death

Defense/ immune response.

Cell multiplication

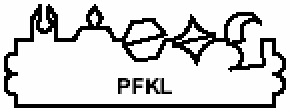
Cell periphery

Glyphs

ONTOGLYPH

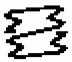
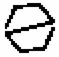
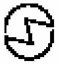
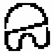
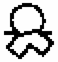
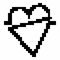
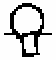
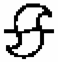
Blueprint

ONTOGLYPHS



Symbols to display protein attributes schematically. A collection of glyphs consisting of 34 functional, 25 binding and 24 location categories populated with Gene Ontology (GO) terms.

FUNCTIONAL CATEGORIES

Category	Glyph
Alcohol & aromatic compound metabolism/transport	
Behaviour	
Binding activity	
Biomembranes: organization and basic functions	
Carbohydrate metabolism/transport	
Cell communication	
Cell motility and structural activity	
Cell multiplication	
Cell physiology	
Cellular organization/biogenesis	
Coenzymes vitamin & hormone metabolism/transport	
Death	
Defence/immune response	
Development	
DNA metabolism	
Energy production/conversion	
General metabolism	
General physiological processes	
General transport	
Ion transport	
Lipid metabolism/transport	
Organic & amino acids & derivatives & amine metabolism/transport	
Pathogenesis & toxin activity	
Phosphorus metabolism	
Protein degradation	
Protein metabolism	
Protein modification	
Protein Synthesis, Processing, and Regulation	
Protein transport	
Response to stress endogenous or exogenous stimulus	
RNA & other nucleic acids metabolism and transport	
Signal transduction	
Transcription	
Viral life cycle	

BINDING CATEGORIES

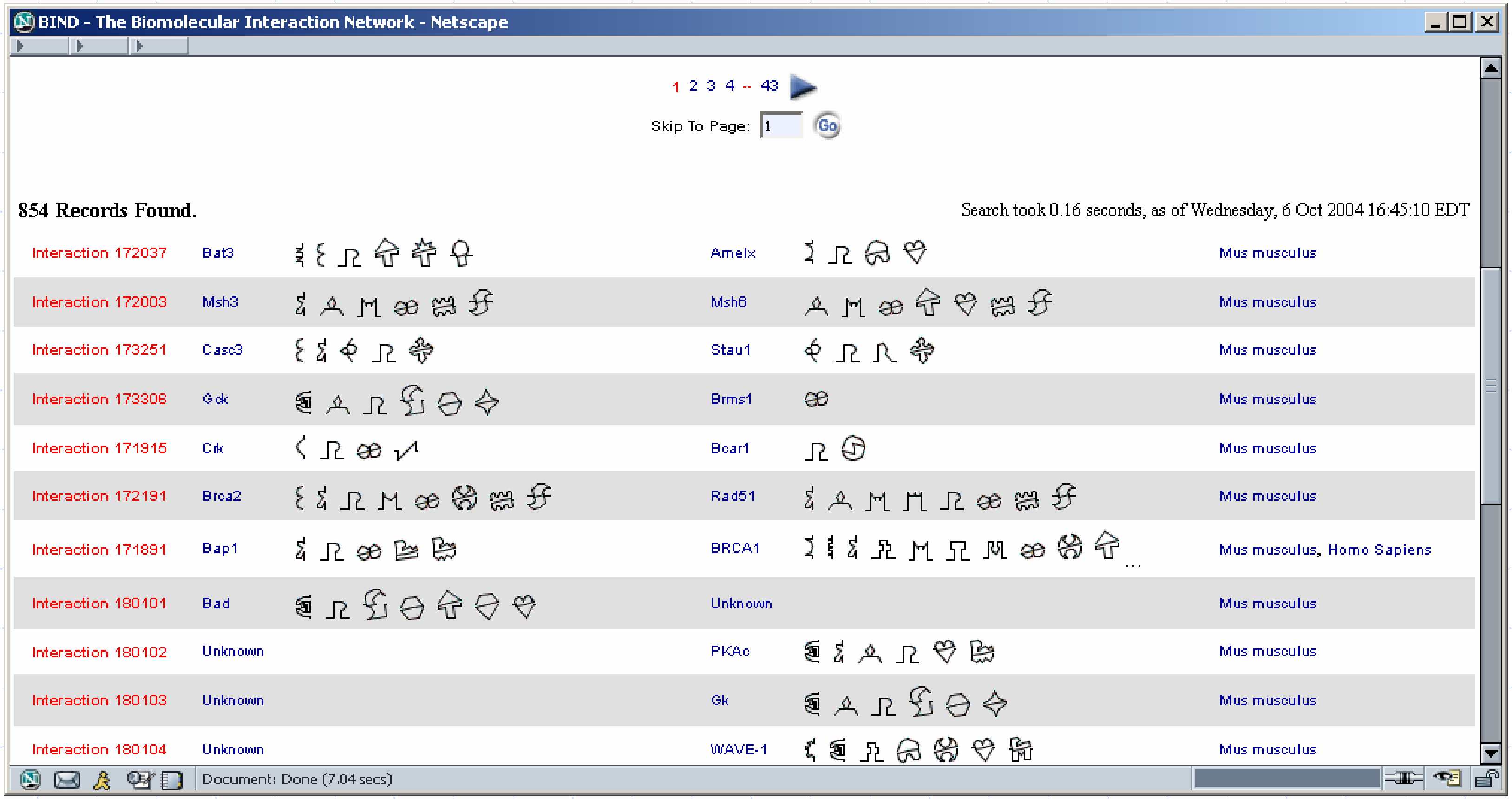
Category	Glyph
Adenyl nucleotide binding	
Antigen binding	
ATP binding	
Calcium ion binding	
Calmodulin binding	
Carbohydrate binding	
Cytokine binding	
Cytoskeletal protein binding	
Damage DNA binding	
DNA binding	
Double stranded DNA binding	
Guanyl nucleotide binding	
Lipid binding	
Metal ion binding	
mRNA binding	
Nucleic acid binding	
Nucleotide binding	
Oxygen binding	
Protein binding	
Purine nucleotide binding	
Receptor binding	
RNA binding	
Single stranded DNA binding	
Transcription factor binding	
Transition metal ions binding	

SUB-CELLULAR LOCALIZATION CATEGORIES

Category	Glyph
Actin cytoskeleton	
Axon or dendrite	
Biological membrane	
Cell periphery	
Cytoplasm	
Cytoplasmic vesicle	
Endoplasmic reticulum	
Endosome	
Extracellular /cell surface	
Flagellum /cilium	
Golgi apparatus	
Lipid particle	
Microtubule cytoskeleton	
Mitochondrion	
Nuclear periphery	
Nucleolus	
Nucleus	
Peroxisome	
Plastid	
Protoplasm	
Ribonucleoprotein complex	
Site of polarized growth.	
Vacuole	
Virion provirus.	

Results - Cancer query – interactions only – restricted to BIND-metazoa + MGI divisions

Single-Line OntoGlyphs now appear on BIND search results…



854 Records Found.

Search took 0.16 seconds, as of Wednesday, 6 Oct 2004 16:45:10 EDT

Interaction 172037	Bat3	Amelx	Mus musculus
Interaction 172003	Msh3	Msh6	Mus musculus
Interaction 173251	Casc3	Stau1	Mus musculus
Interaction 173306	Gck	Brms1	Mus musculus
Interaction 171915	Crk	Bcar1	Mus musculus
Interaction 172191	Brca2	Rad51	Mus musculus
Interaction 171891	Bap1	BRCA1 …	Mus musculus, Homo Sapiens
Interaction 180101	Bad	Unknown	Mus musculus
Interaction 180102	Unknown	PKAc	Mus musculus
Interaction 180103	Unknown	Gk	Mus musculus
Interaction 180104	Unknown	WAVE-1	Mus musculus

MOUNT SINAI HOSPITAL

Blueprint

Any list of genes can be converted into 3 lines of ontoglyphs. BIND query, BLAST output, microarray data…

MOUNT
SINAI
HOSPITAL

Blueprint

Q6. Main Concerns to be Addressed

- The entire 300-year old publish-or-perish reward model.
  - Papers have to be stories, not observations (data)
  - No reward for novel, unique data from isolated experiments
  - No capture into print or into databases

Q12. Should or can journals enforce submission of data?

- BIND is working with 23 Journal titles currently including Science, NPG, Cell Press, NRC Press on pre and post-publication capture of interactions. We are working to extend these relationships with our IMEX partners.

- Yes, it can be done. The devil is in the details – it requires very focused and respectful consideration of the need of journal editors, and manuscript submitters. It needs trained curators.

- No scientist wants to have to make an XML document in addition to their paper. Training required is a large burden.
  - Tim Hughes - MIAME frustration – echoed by journal editors.

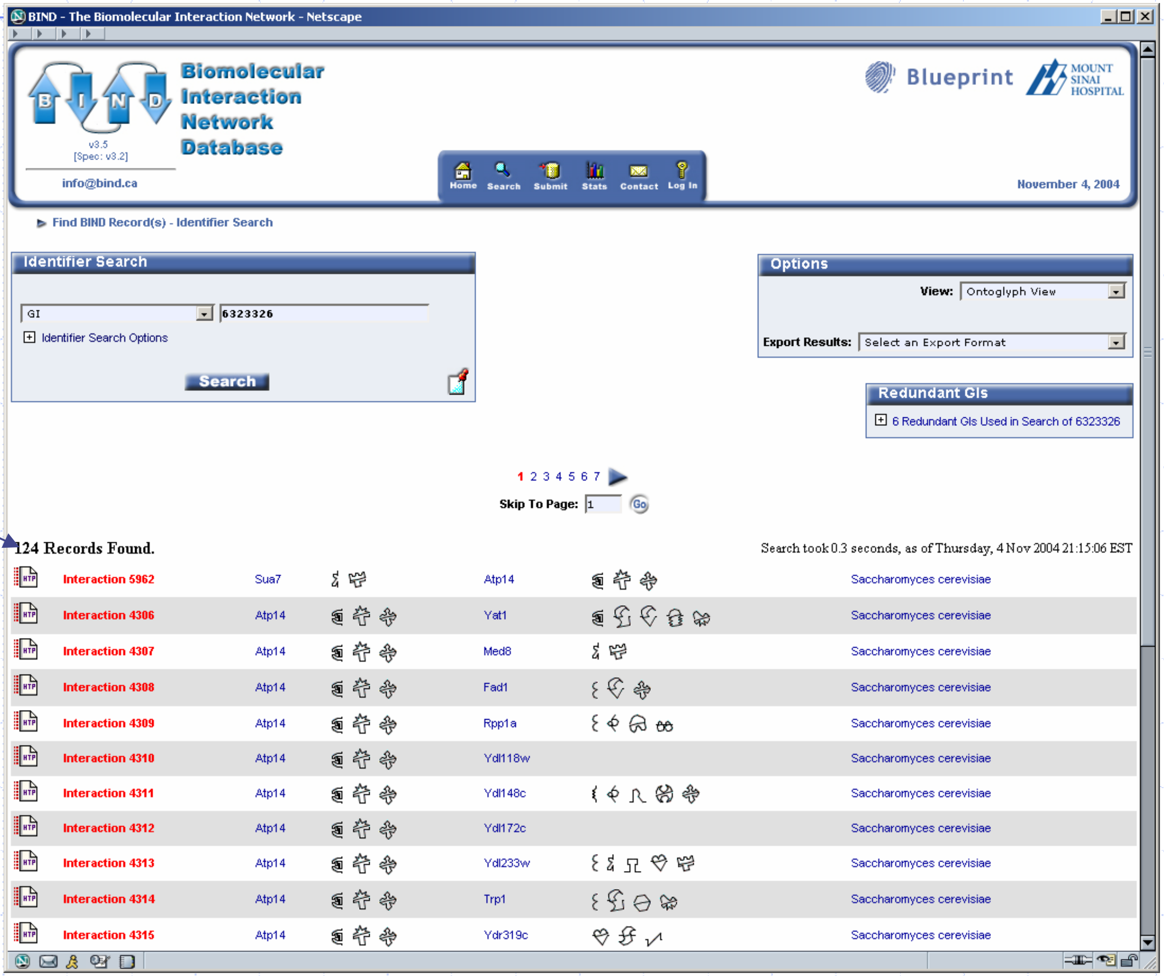
- BIND curators structure the data for the submitters, not just tag it with semantic terms. BIND data is intended to be computable in the long-term, justifying the effort.

Data Archiving vs Data Analysis

- Effort is not balanced
  - Data Generation >> Data Analysis
  - Data hoarding results…

- Archivists are expected to behave like
  - Librarians? (Archive the data)
  - Literary Critics? (Evaluate/Rank the data)

BIND Viewer Tool – atp14



BIND - The Biomolecular Interaction Network - Netscape

GI: 6323326

View: Ontoglyph View

Export Results: Select an Export Format

6 Redundant GIs Used in Search of 6323326

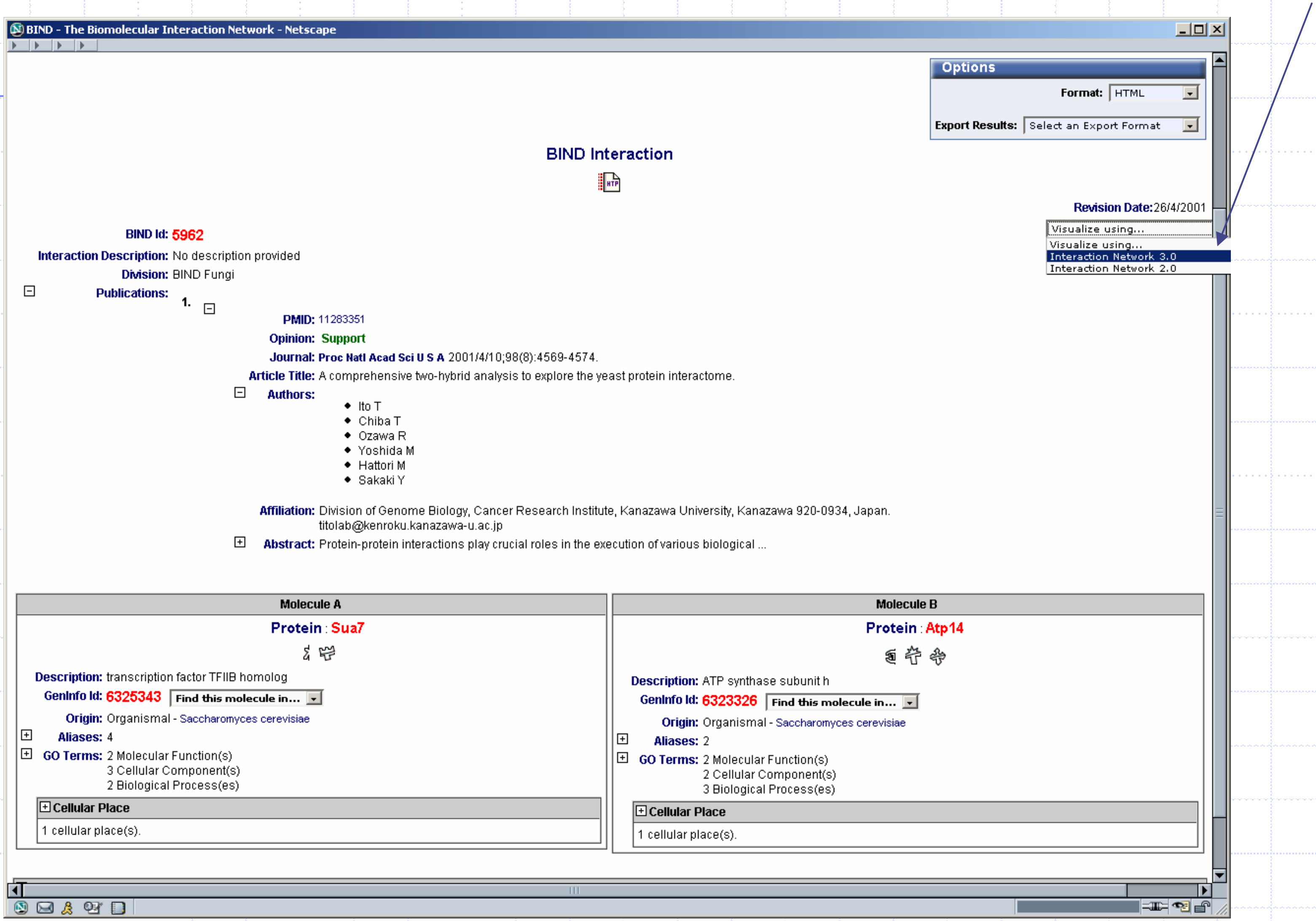
Skip To Page: 1

124 Records Found.

Search took 0.3 seconds, as of Thursday, 4 Nov 2004 21:15:06 EST

Interaction 5962	Sua7	Atp14	Saccharomyces cerevisiae
Interaction 4306	Atp14	Yat1	Saccharomyces cerevisiae
Interaction 4307	Atp14	Med8	Saccharomyces cerevisiae
Interaction 4308	Atp14	Fad1	Saccharomyces cerevisiae
Interaction 4309	Atp14	Rpp1a	Saccharomyces cerevisiae
Interaction 4310	Atp14	Ydl118w	Saccharomyces cerevisiae
Interaction 4311	Atp14	Ydl148c	Saccharomyces cerevisiae
Interaction 4312	Atp14	Ydl172c	Saccharomyces cerevisiae
Interaction 4313	Atp14	Ydl233w	Saccharomyces cerevisiae
Interaction 4314	Atp14	Trp1	Saccharomyces cerevisiae
Interaction 4315	Atp14	Ydr319c	Saccharomyces cerevisiae

Many hits From yeast-two-hybrid data



Format: HTML

Export Results: Select an Export Format

BIND Interaction

Revision Date: 26/4/2001

BIND Id: 5962

Interaction Description: No description provided

Division: BIND Fungi

Publications:

1.

PMID: 11283351

Opinion: Support

Journal: Proc Natl Acad Sci U S A 2001/4/10;98(8):4569-4574.

Article Title: A comprehensive two-hybrid analysis to explore the yeast protein interactome.

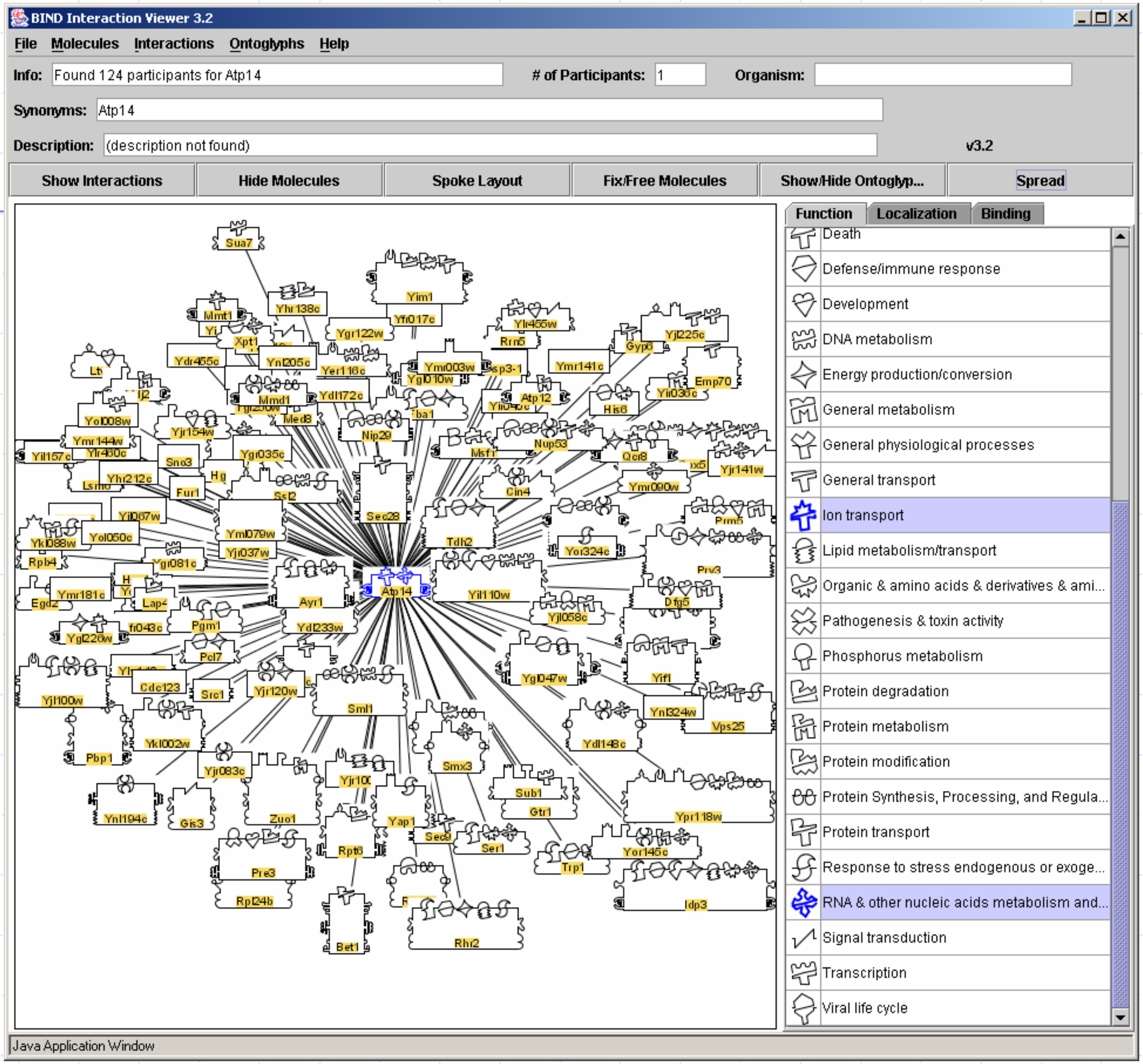
Authors:

- Ito T
- Chiba T
- Ozawa R
- Yoshida M
- Hattori M
- Sakaki Y

Affiliation: Division of Genome Biology, Cancer Research Institute, Kanazawa University, Kanazawa 920-0934, Japan. titolab@kenroku.kanazawa-u.ac.jp

Abstract: Protein-protein interactions play crucial roles in the execution of various biological ...

Molecule A	Molecule B
Protein : Sua7	**Protein** : Atp14
Description: transcription factor TFIIB homolog	**Description:** ATP synthase subunit h
GenInfo Id: 6325343	**GenInfo Id:** 6323326
Origin: Organismal - Saccharomyces cerevisiae	**Origin:** Organismal - Saccharomyces cerevisiae
Aliases: 4	**Aliases:** 2
GO Terms: 2 Molecular Function(s) 3 Cellular Component(s) 2 Biological Process(es)	**GO Terms:** 2 Molecular Function(s) 2 Cellular Component(s) 3 Biological Process(es)
1 cellular place(s).	1 cellular place(s).



BIND Interaction Viewer 3.2

Info: Found 124 participants for Atp14

# of Participants: 1

Organism:

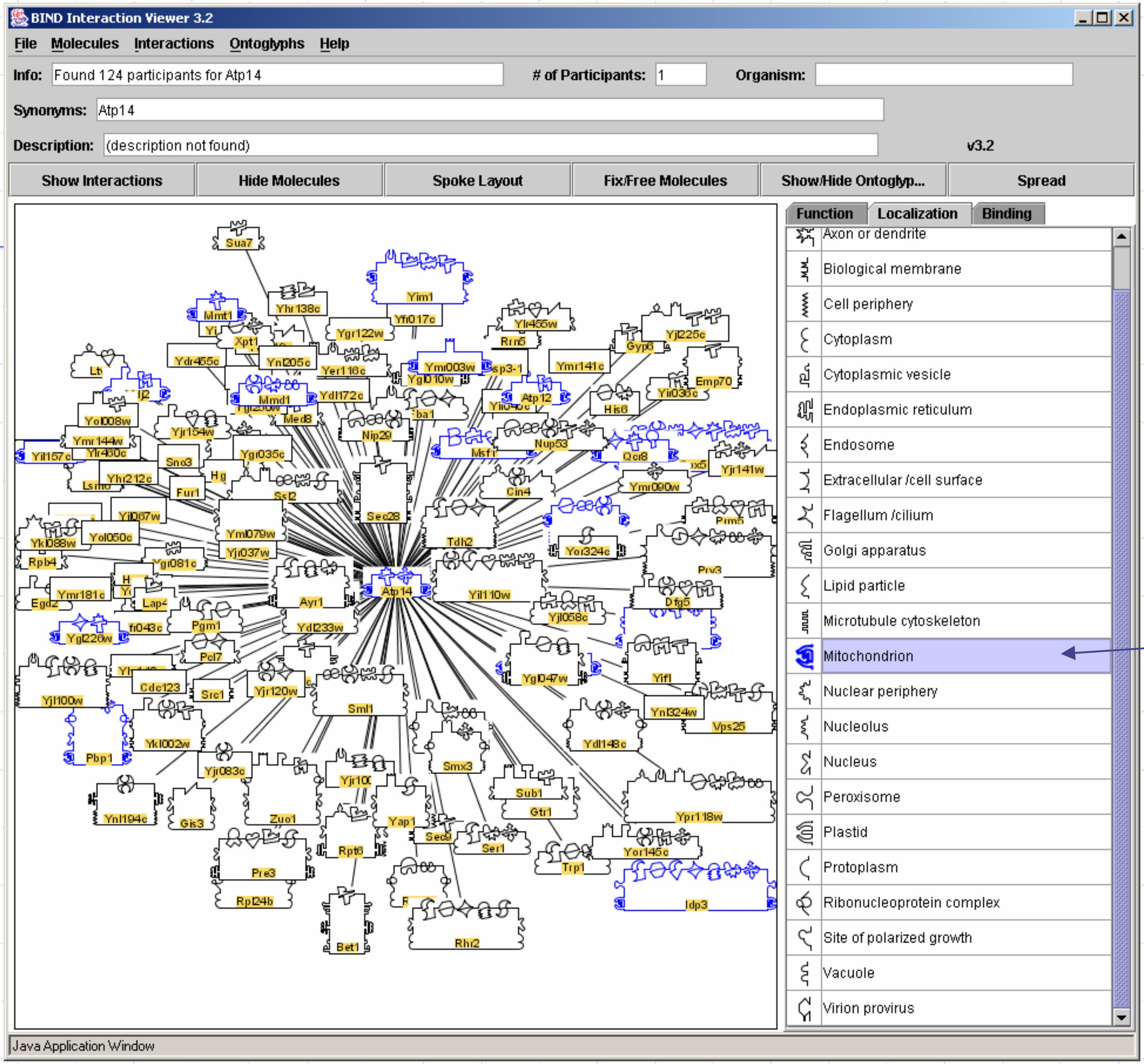
Synonyms: Atp14

Description: (description not found)

v3.2

Too many…

Which molecules are co-localized With the atp14?



Info: Found 124 participants for Atp14

# of Participants: 1

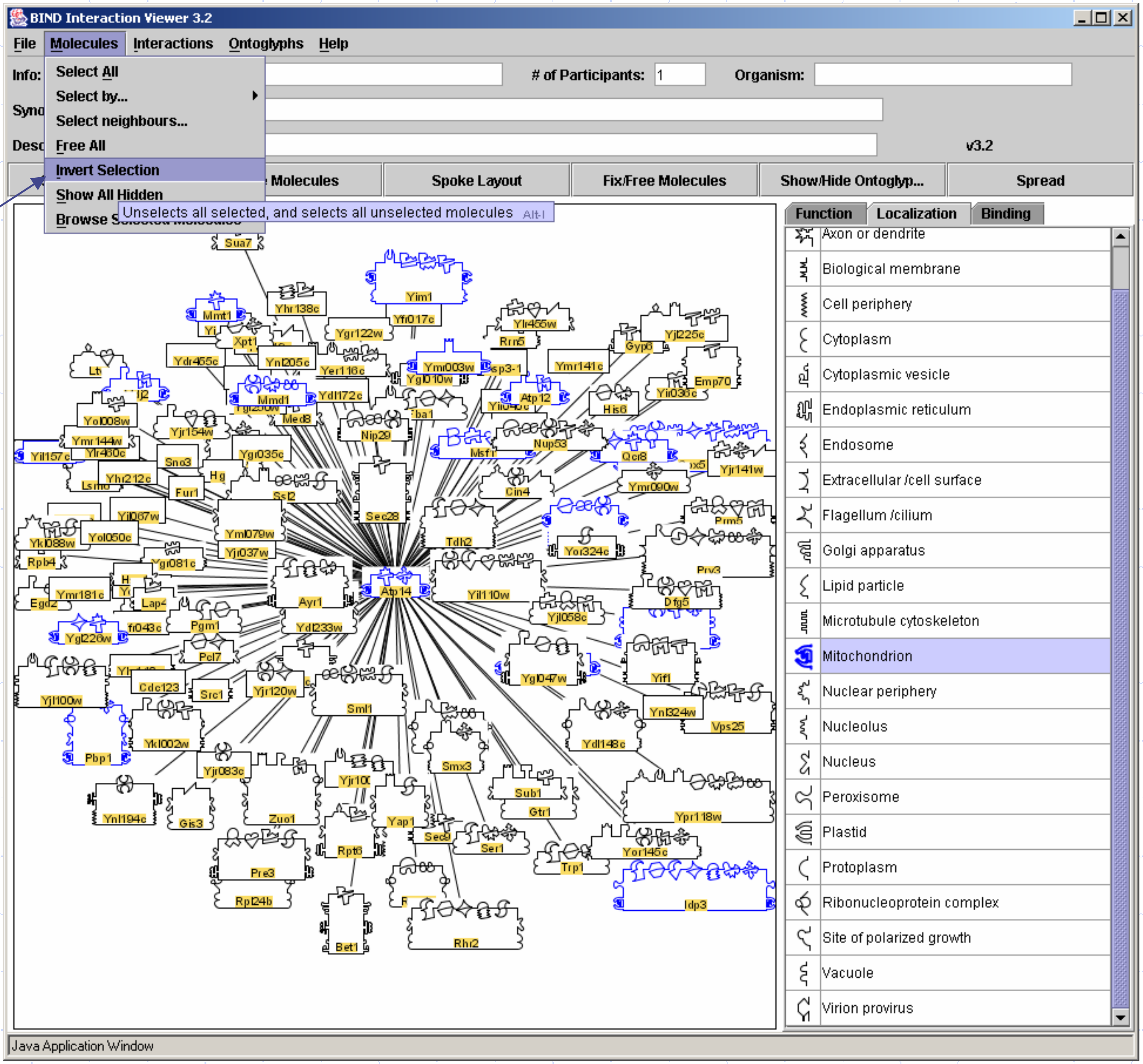
Organism:

Synonyms: Atp14

Description: (description not found)

Select

Invert



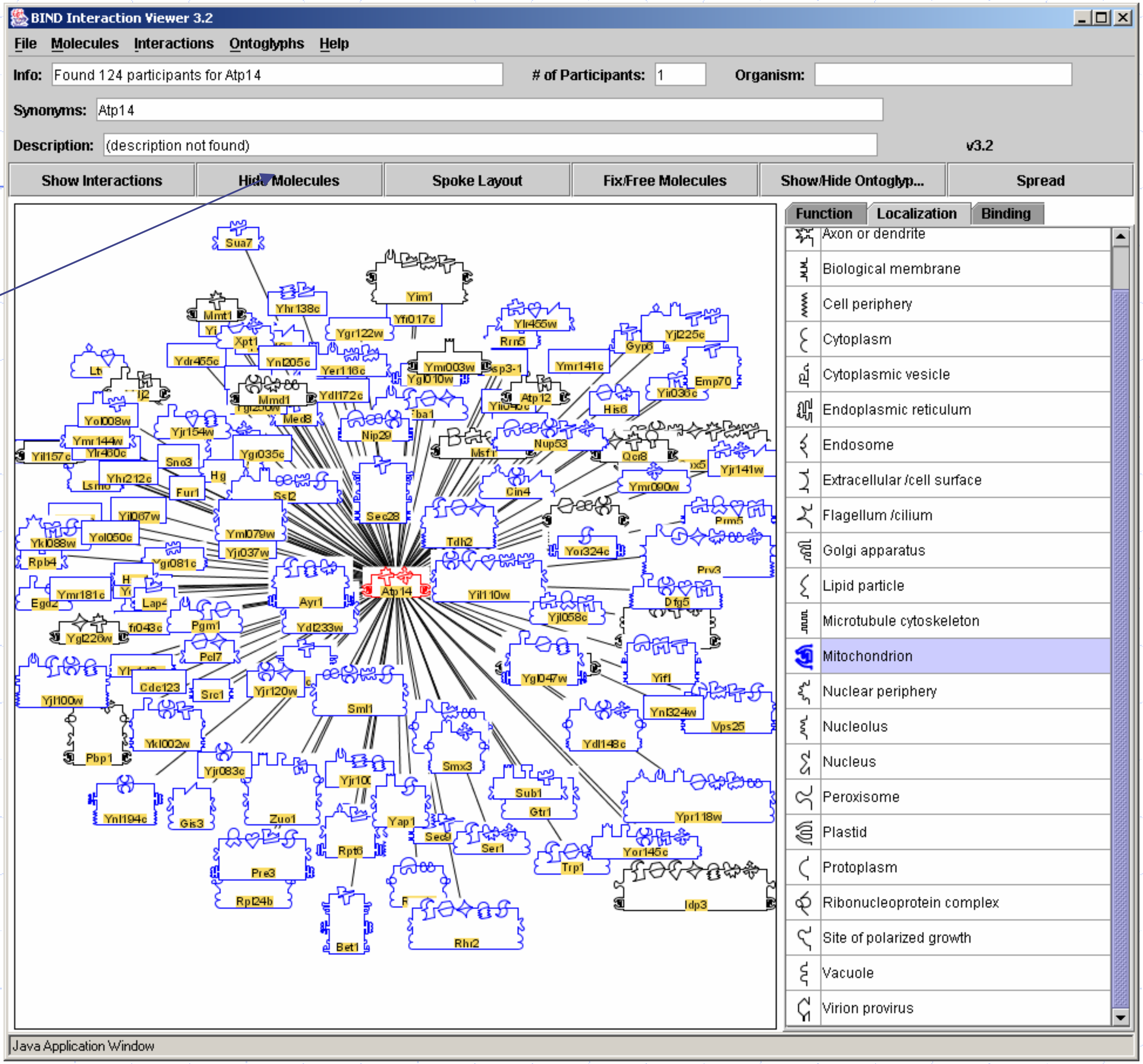
\# of Participants: 1

Organism:

v3.2

Unselects all selected, and selects all unselected molecules Alt-I

Hide Selected



Info: Found 124 participants for Atp14

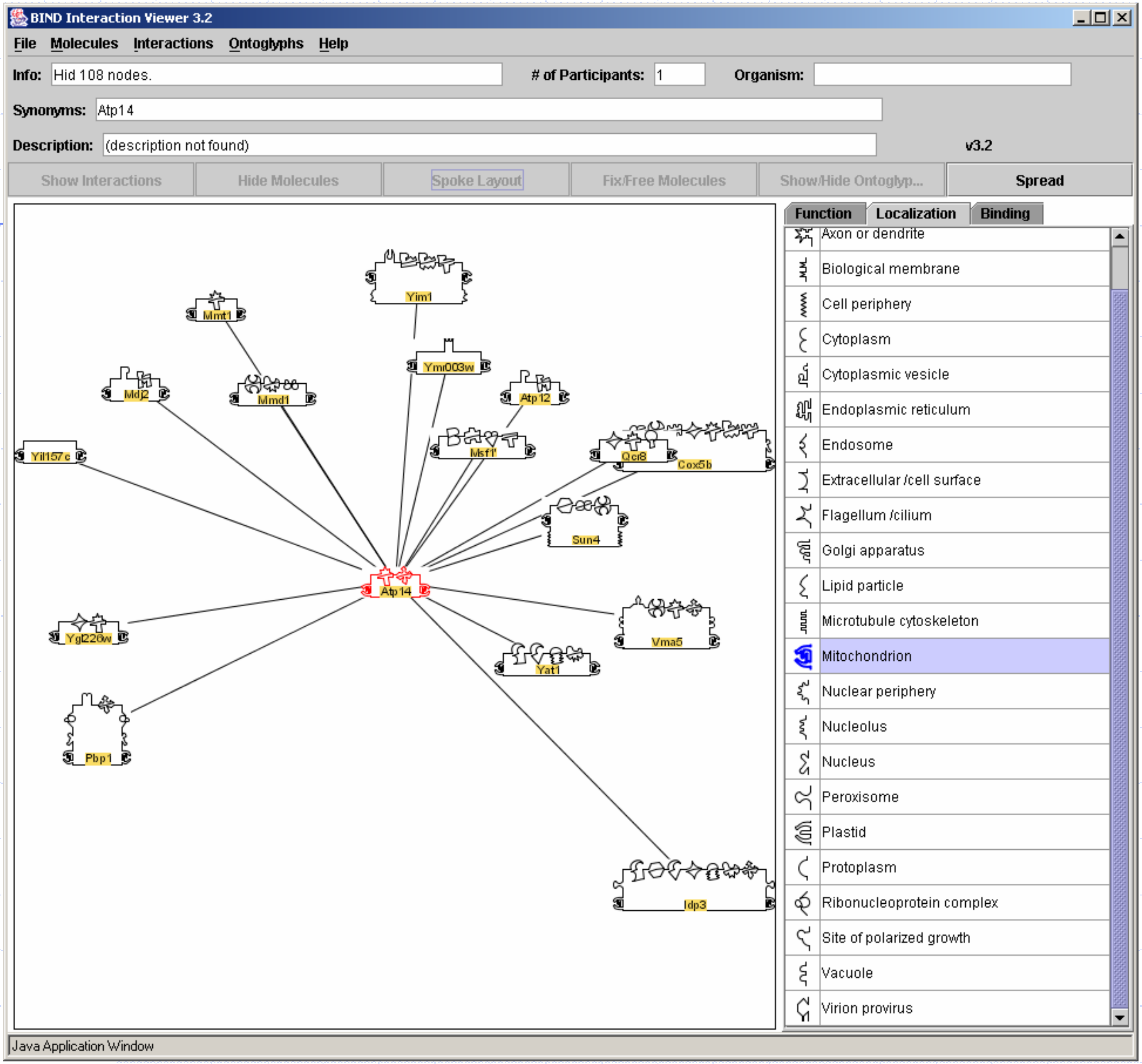
# of Participants: 1

Organism:

Synonyms: Atp14

Description: (description not found)

v3.2

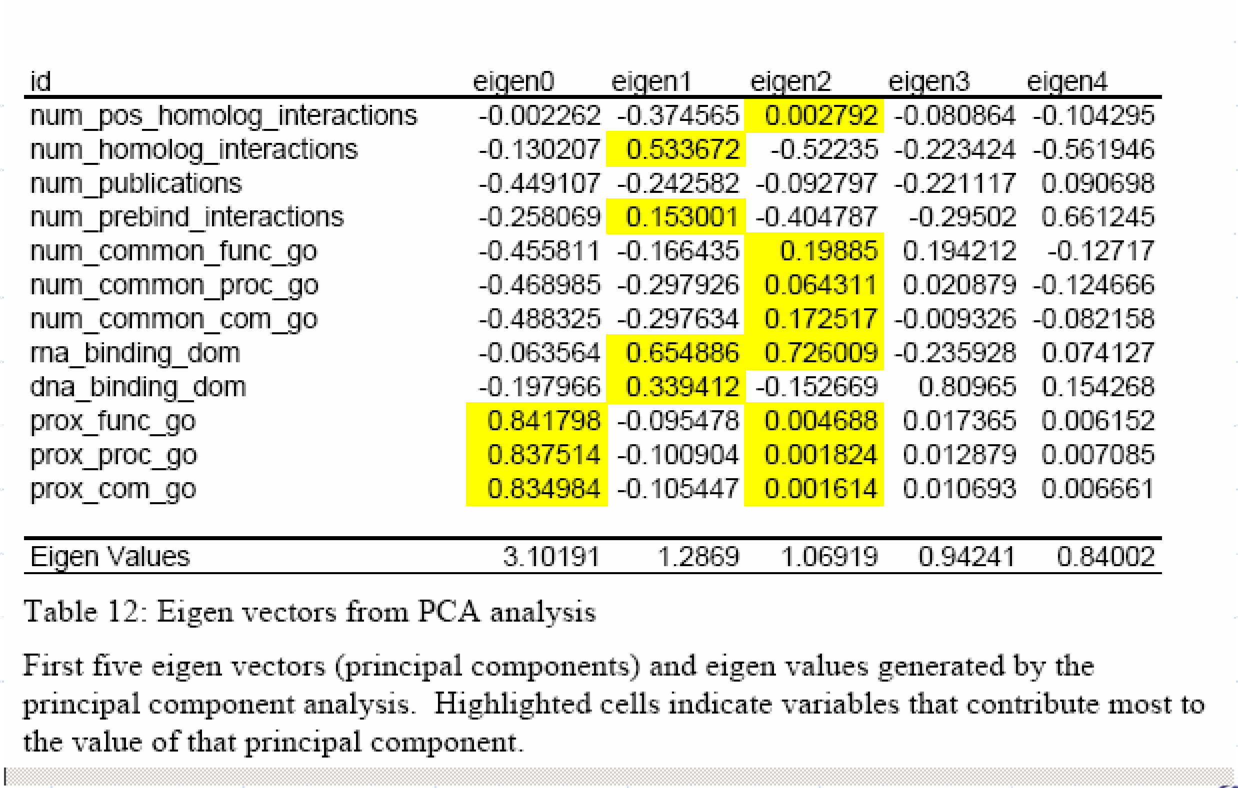


Info: Hid 108 nodes.
# of Participants: 1
Organism:
Synonyms: Atp14
Description: (description not found)
v3.2

Voila
Only
co-localized
Proteins!

Scoring High Throughput Interactions

- BIND PICKS – Protein Interaction Confidence Kernel Scores – SVM Classifier for Yeast.



id	eigen0	eigen1	eigen2	eigen3	eigen4
num_pos_homolog_interactions	-0.002262	-0.374565	0.002792	-0.080864	-0.104295
num_homolog_interactions	-0.130207	0.533672	-0.52235	-0.223424	-0.561946
num_publications	-0.449107	-0.242582	-0.092797	-0.221117	0.090698
num_prebind_interactions	-0.258069	0.153001	-0.404787	-0.29502	0.661245
num_common_func_go	-0.455811	-0.166435	0.19885	0.194212	-0.12717
num_common_proc_go	-0.468985	-0.297926	0.064311	0.020879	-0.124666
num_common_com_go	-0.488325	-0.297634	0.172517	-0.009326	-0.082158
rna_binding_dom	-0.063564	0.654886	0.726009	-0.235928	0.074127
dna_binding_dom	-0.197966	0.339412	-0.152669	0.80965	0.154268
prox_func_go	0.841798	-0.095478	0.004688	0.017365	0.006152
prox_proc_go	0.837514	-0.100904	0.001824	0.012879	0.007085
prox_com_go	0.834984	-0.105447	0.001614	0.010693	0.006661
Eigen Values	3.10191	1.2869	1.06919	0.94241	0.84002

Table 12: Eigen vectors from PCA analysis

First five eigen vectors (principal components) and eigen values generated by the principal component analysis. Highlighted cells indicate variables that contribute most to the value of that principal component.

MOUNT SINAI HOSPITAL

Blueprint

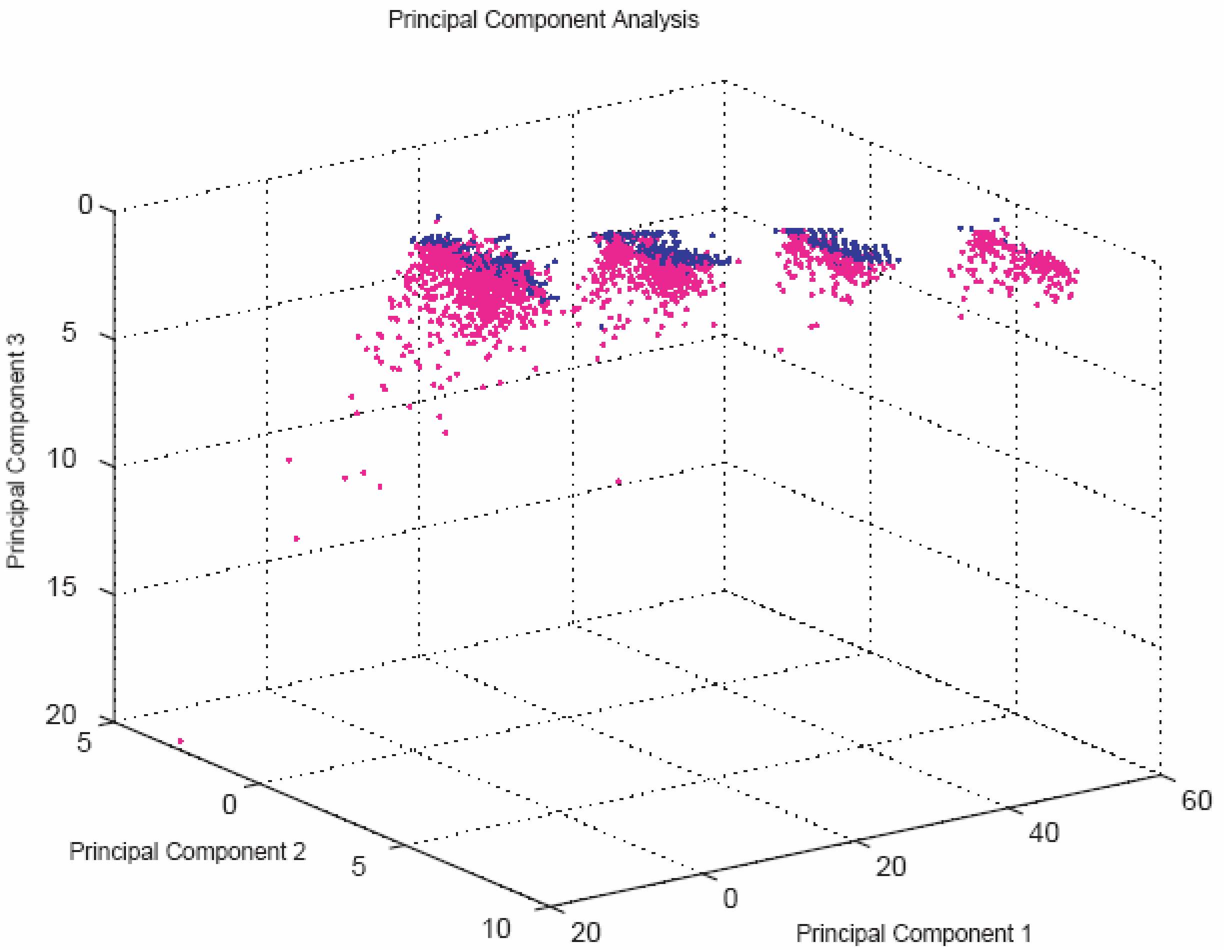
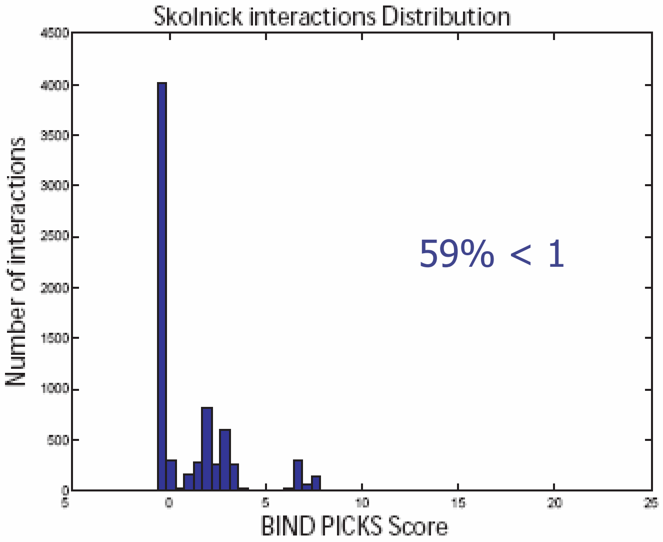


Figure 15: PCA

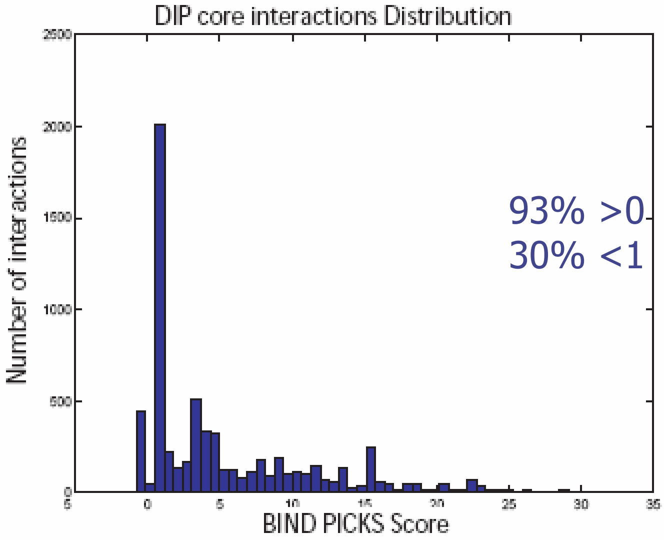
Positive subset (magenta) and negative subset (blue) were mapped onto to the first three principal components and graphed.

Skolnick Dataset



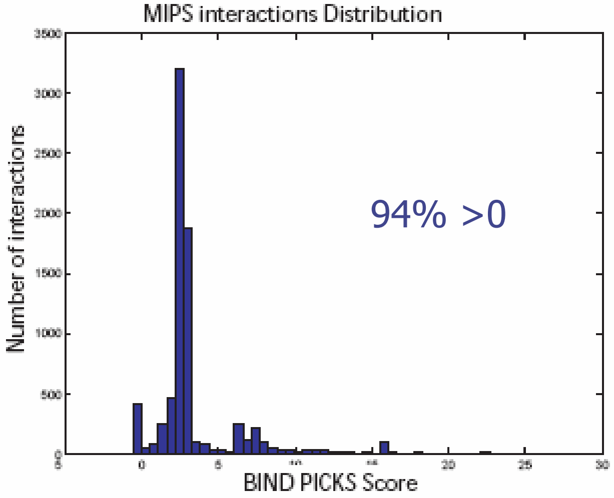
59% < 1

DIP core Dataset



93% >0
30% <1

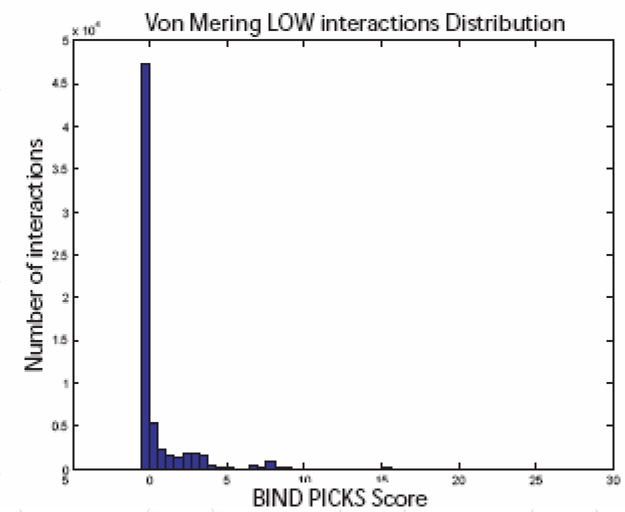
MIPS Dataset

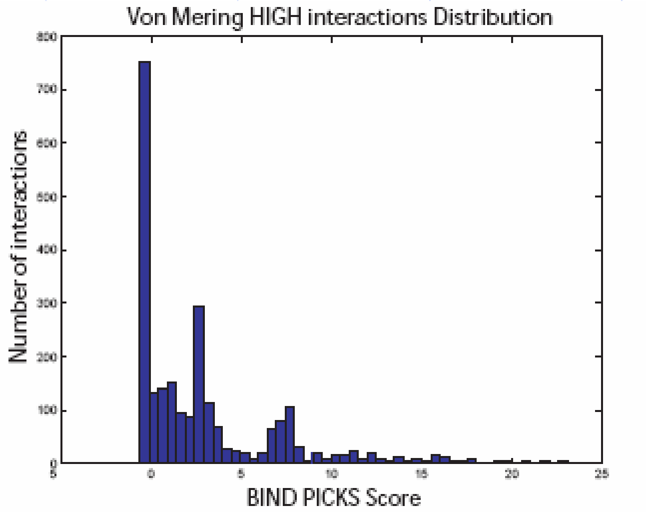


94% >0

Von Mering Scores vs. BIND PICKS

- High – supported by more than one method.
- Low – Seen only once in one method.



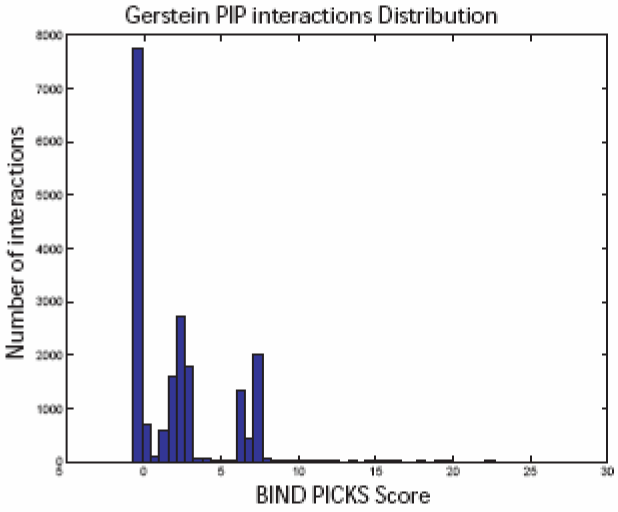


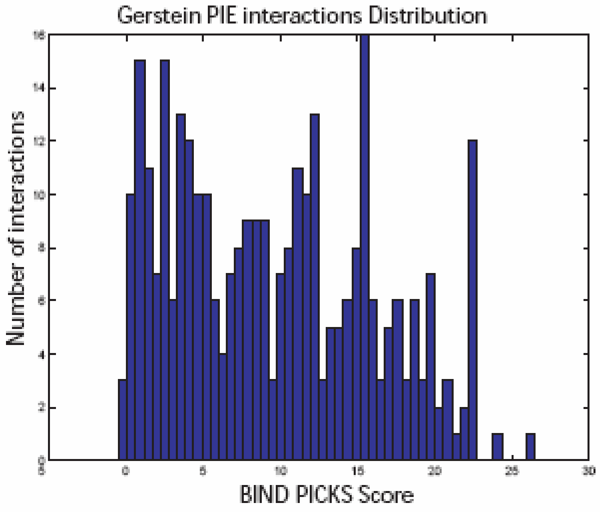
Total = 65,731 (Matrix)
Score > 0 = 24%

Total = 2445
Score > 0 = 68%

Medium – 47%

Gerstein Scores vs. BIND PICKS


Gerstein PIP interactions Distribution


Gerstein PIE interactions Distribution

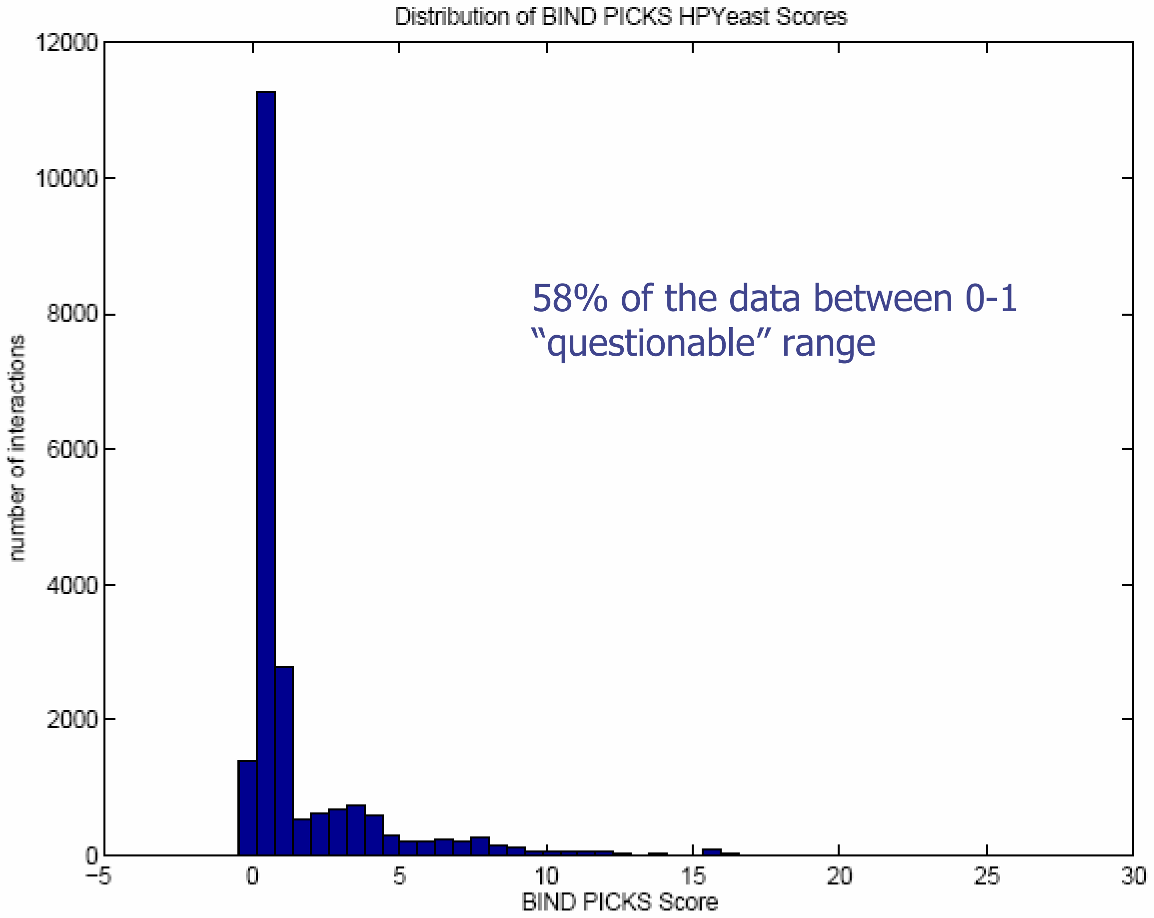
Total = 9897
Score > 0 = 57%

Total = 163
Score > 0 = 99%

MOUNT
SINAI
HOSPITAL

Blueprint

Reliable “spoke” HTP data? 36% >1

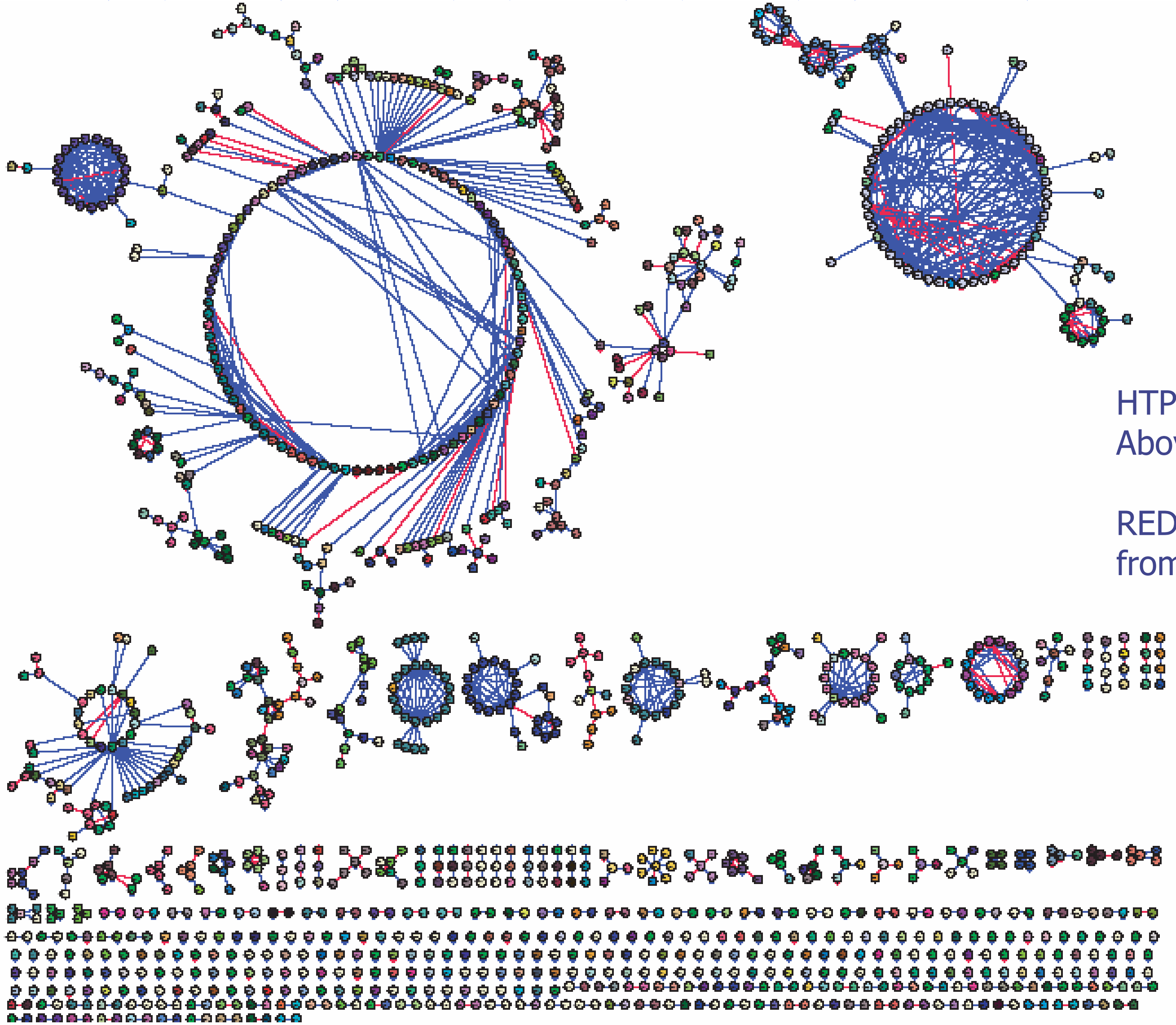


58% of the data between 0-1 “questionable” range

Figure 18: Distribution of BIND PICKS scores for HTP yeast interaction subset. The percentage of interactions that were identified as true interactions (score >1) was 36%.

MOUNT SINAI HOSPITAL

Blueprint



HTP spoke interactions
Above score = 5

RED – positives
from SVM training set

BIND 4.0 – New Record Types Supporting Pathway/Proteomics Data

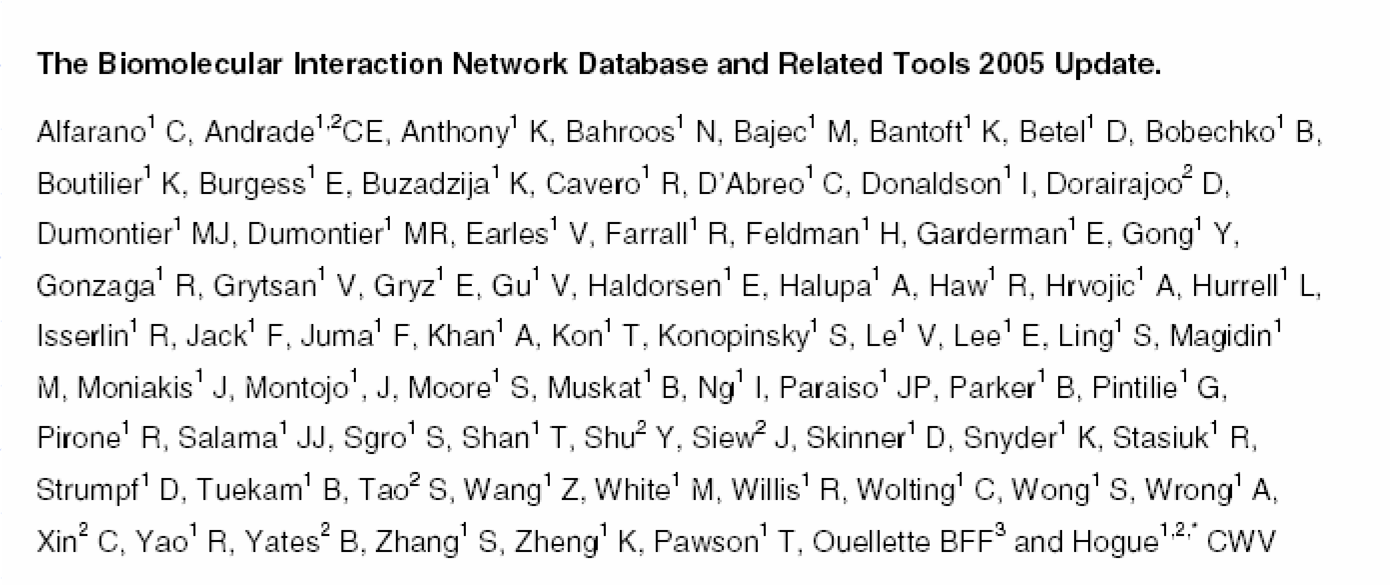
- 3 Types in use:
  - Interaction, Complex, Pathway
- Improvements to Complex, Pathway
  - EcoCyc, aMaze, PID, STKE
- 2 high-level types being added in v4.0
  - Reactant List
  - Assembly List

New BIND record types

- Reactant List: In-vitro reagents required to elicit activity (e.g. Mg, ATP, etc).
  - Arises from an active collaboration with pathway dbs, data analysis and mapping, and efforts to integrate their data into the BIND query system.

- Assembly List: Observed List of Proteins/Genes (non-ratiometric!)
  - Currently unarchived data.
  - Phosphoproteomics, localization (e.g. human nucleolar proteins), transcription factor target proteins, concentration/copy number experiments.
  - May be suitable for “GeneString” conversion.

Staff and Contributors



The Biomolecular Interaction Network Database and Related Tools 2005 Update.

Alfarano[1] C, Andrade[1,2]CE, Anthony[1] K, Bahroos[1] N, Bajec[1] M, Bantoft[1] K, Betel[1] D, Bobechko[1] B, Boutilier[1] K, Burgess[1] E, Buzadzija[1] K, Cavero[1] R, D'Abreo[1] C, Donaldson[1] I, Dorairajoo[2] D, Dumontier[1] MJ, Dumontier[1] MR, Earles[1] V, Farrall[1] R, Feldman[1] H, Garderman[1] E, Gong[1] Y, Gonzaga[1] R, Grytsan[1] V, Gryz[1] E, Gu[1] V, Haldorsen[1] E, Halupa[1] A, Haw[1] R, Hrvojic[1] A, Hurrell[1] L, Isserlin[1] R, Jack[1] F, Juma[1] F, Khan[1] A, Kon[1] T, Konopinsky[1] S, Le[1] V, Lee[1] E, Ling[1] S, Magidin[1] M, Moniakis[1] J, Montojo[1], J, Moore[1] S, Muskat[1] B, Ng[1] I, Paraiso[1] JP, Parker[1] B, Pintilie[1] G, Pirone[1] R, Salama[1] JJ, Sgro[1] S, Shan[1] T, Shu[2] Y, Siew[2] J, Skinner[1] D, Snyder[1] K, Stasiuk[1] R, Strumpf[1] D, Tuekam[1] B, Tao[2] S, Wang[1] Z, White[1] M, Willis[1] R, Wolting[1] C, Wong[1] S, Wrong[1] A, Xin[2] C, Yao[1] R, Yates[2] B, Zhang[1] S, Zheng[1] K, Pawson[1] T, Ouellette BFF[3] and Hogue[1,2,*] CWV