

# Elucidating Biological Pathways by Integrating Gene Annotations and Gene Expression Data

**Arthur Fridman**

Jeff Sachs

CHI Beyond Genome: Bioinformatics

June 17, 2003

San Diego, CA

# Acknowledgements

## Kui Shao

### 3T3L1 Data

- Bei Zhang
- David Gerhold
- Jian Xu
- Richard Raubertas
- Rosetta 3T3L1 team

### Annotation Method Support

- Shaun Deignan
- Rick Blevins
- Alex Elbrecht
- David Haynor
- Roland Stoughton

# Outline

- Objectives and executive summary
- Background
  - Things that don't work on typical Merck data
  - Phenotype-Function-nets: combining TP & annotation data
- The data
- PF-Nets
  - Data setup
  - Prediction examples

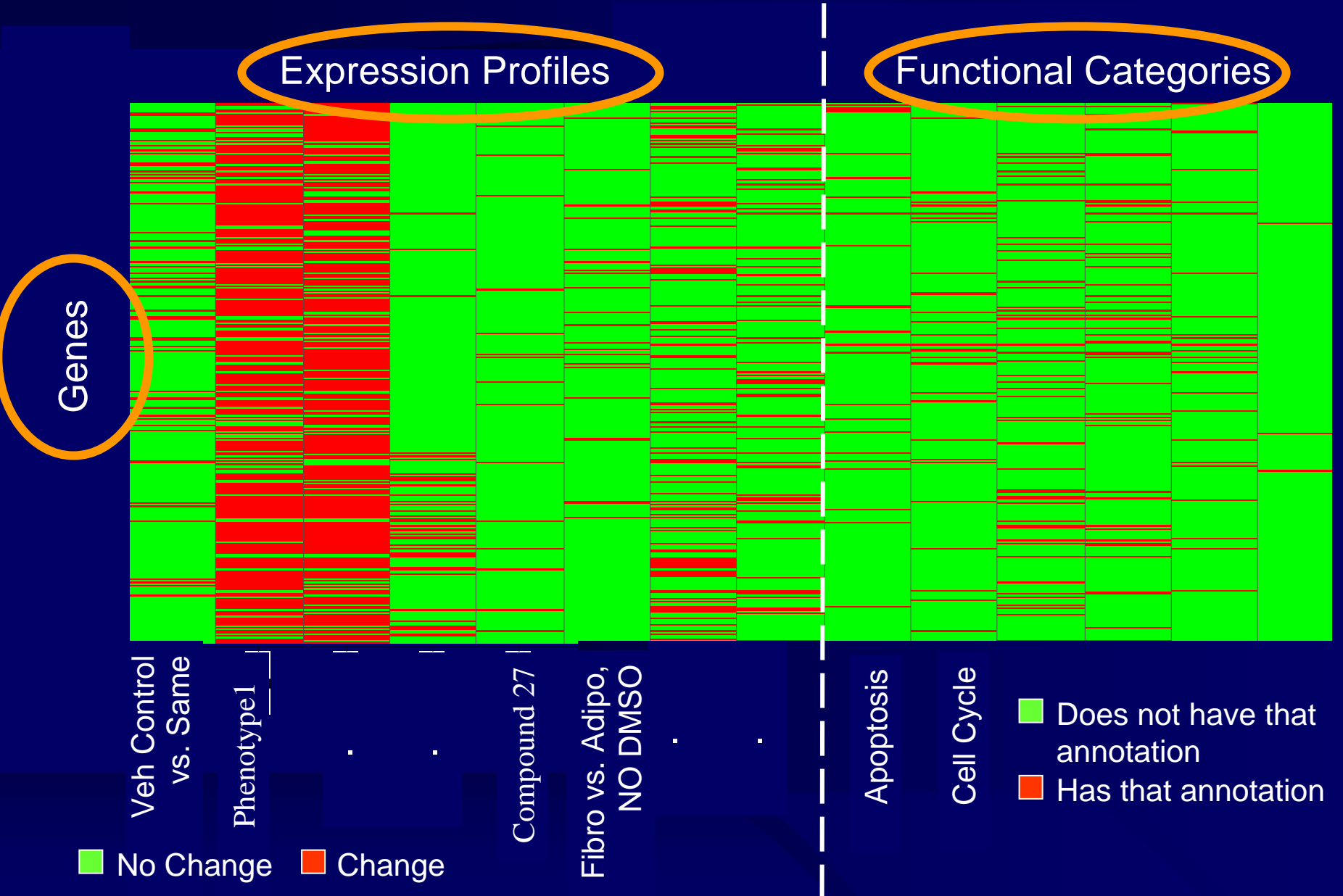
# Objectives

- Design, develop, and apply software for inferring putative biological mechanisms from Merck and other data.
- Develop graphical models that help explain, visualize, and predict aspects of the complex interactions among a substantial number of variables.
  - Must be supported by sufficient data that are currently available and of a kind anticipated to be frequently available in the future
- Elucidate significant and interesting aspects of biology

# Executive Summary – 2 tech slides follow

- New data integration concept + new algorithmic technology
- Integration: Allows us to find statistically significant relationships between
  - Gene lists
  - Phenotypes (wt, ko, ob/ob, ...)
  - treatments (compound, time, dose)
  - Compound structure class
  - Gene/protein Annotation (biochem, location, GO, etc)
- Should allow us to help position gene sets and pathways with respect to one another as a part of ranking biomarkers
  - Similarly for compound classes
- \* Not networks of genes/proteins

# Phenotype-Function Network: Example Input



# Phenotype-Function Network: Output

- Rules for gene regulation and lists of relevant genes
  - Example: “genes that are regulated by compound x and y are 90% likely to be upregulated by compound z”
  - Rules are statistically ranked and probabilistic, and imply statistically supported relationships between phenotypes and inter-phenotype transitions
- Also rules including relationships: gene functions  $\leftrightarrow$  phenotypes
  - identify gene functional categories associated with responses particular to a subset of compounds.

Example: The label “genes associated with function abc” is strongly associated with compounds of class C known to be less specific than other compounds in program XYZ



Is abc related to off-target effects for Class C compounds?

# Approaches Considered

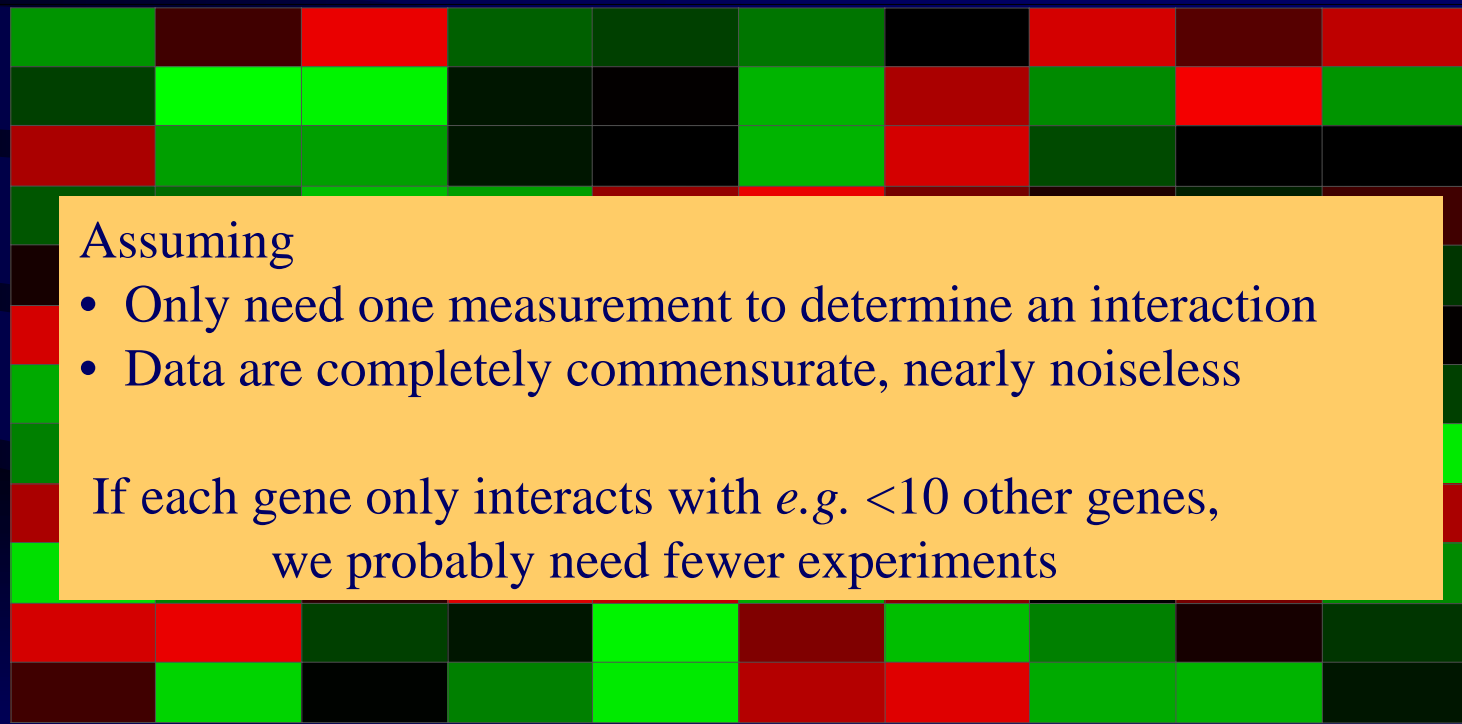
- *De Novo* Gene/Protein Network
  - **Question:** Can we determine a statistically supportable network of gene interactions of interest using only available profile data
- Expansion of Known Network
  - **Question:** Using Merck data, can we place a new gene in a previously known network?
- *De Novo* Phenotype-Function Network
  - **Question:** Can we determine a statistically supportable network of interactions of gene functions and phenotypes of interest using only available data?

# De Novo Gene/Protein Network: Input

Estimate strength of  $\sim 5000^2/2$  interactions

Experiment 1  $\longrightarrow$  Experiment 2500

Gene/Protein 1



Experiment 1

Experiment 3

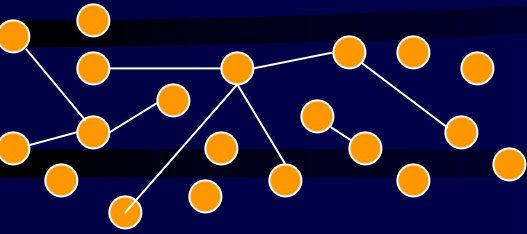
Experiment 5

Experiment 7

Experiment 9

Gene/Protein 5000

# Expansion of Known Network : Input



+

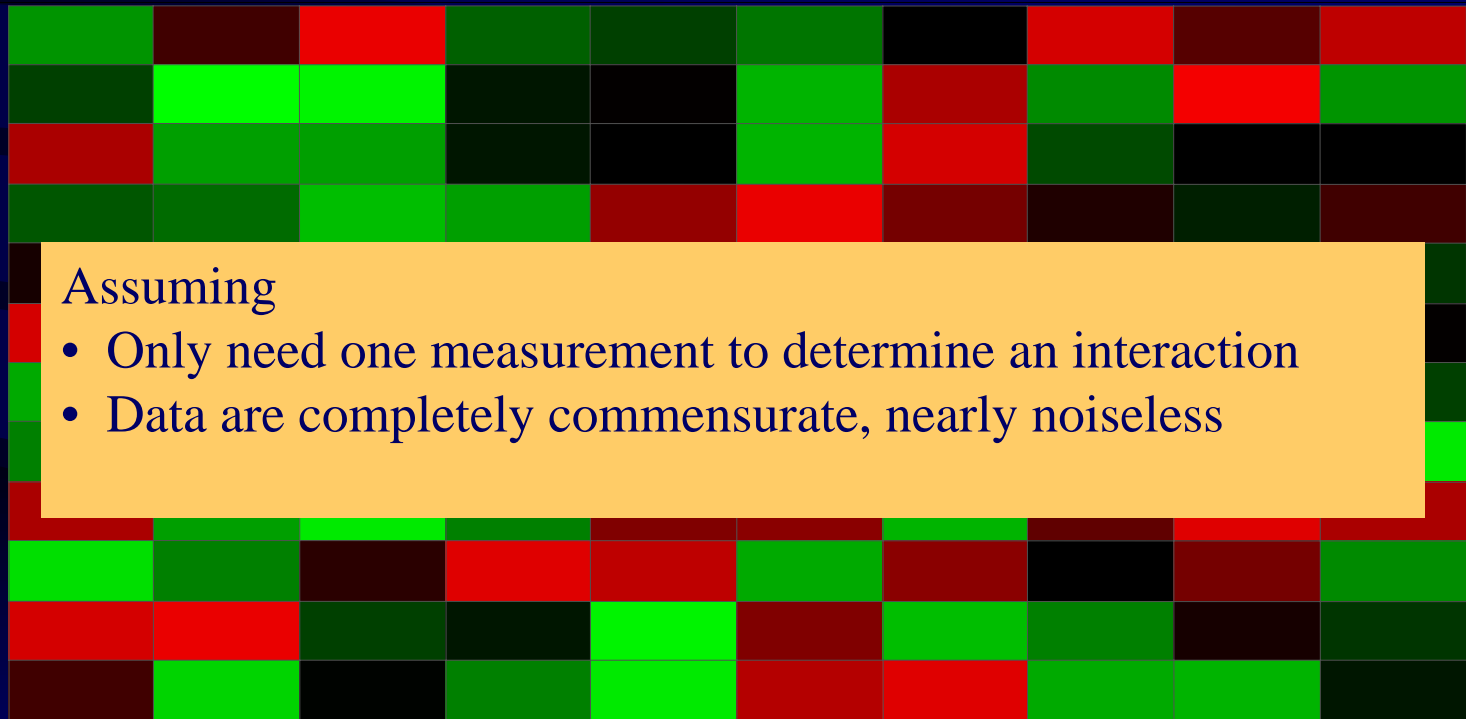
Estimate strength of  $\sim 20 \times 5020$  interactions

Experiment 1



Experiment 20

Gene/Protein 1



Assuming

- Only need one measurement to determine an interaction
- Data are completely commensurate, nearly noiseless

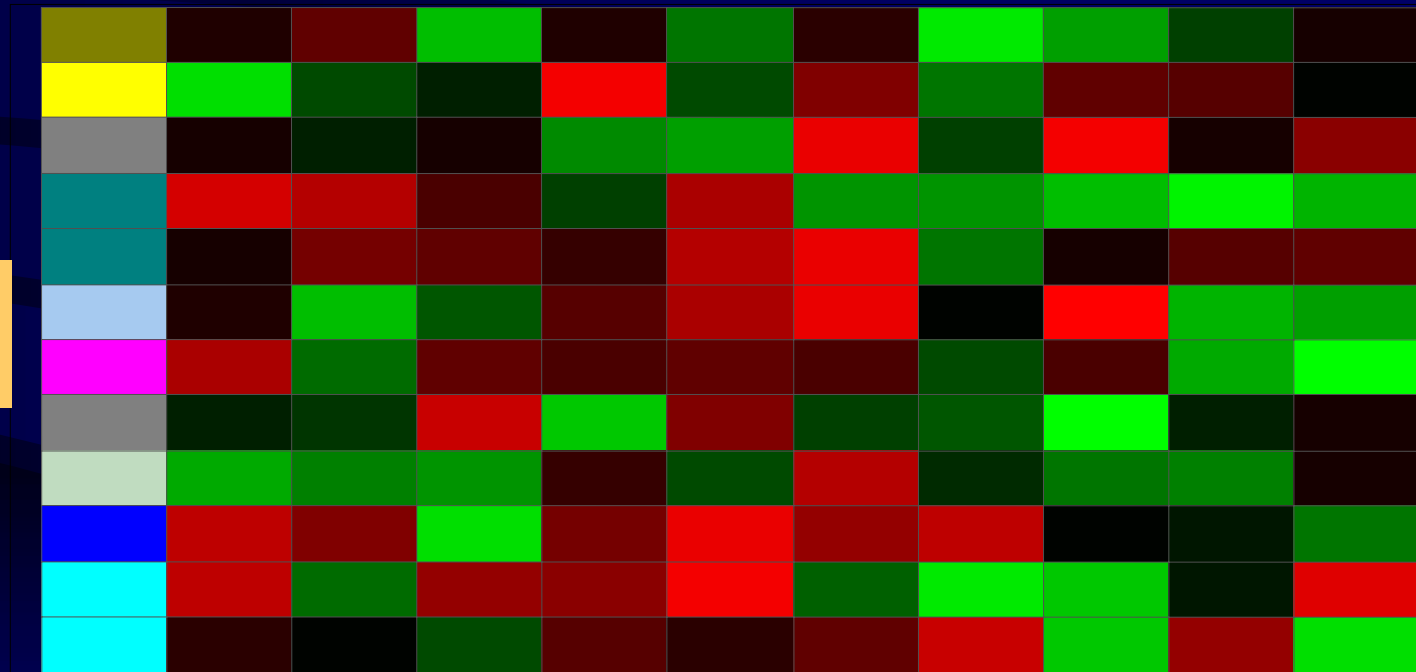
Experiment 1      Experiment 3      Experiment 5      Experiment 7      Experiment 9

Gene/Protein 5000

# De Novo Gene Function and Phenotype Network: Input

Category      Phenotype 1      Phenotype 10  
Experiment 1      Experiment 10

Gene/Protein 1



Determining  
~  $10^2/2$  interactions

Gene/Protein 50  
more is better

# Phenotype-Function Networks

- Combine gene expression and annotation data
  - for biological annotation, a challenge in itself
  - novel solution strategy designed and implemented
- Treat genes as observations, experimental conditions and annotations as variables
- Find unexpected relationships between variables
  - statistically significant patterns
- Represent them graphically
  - Bayes or dependency networks

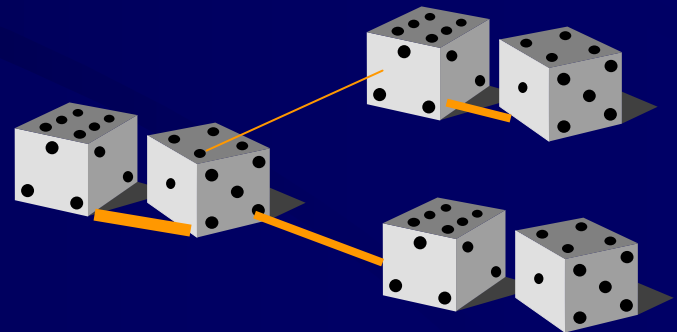
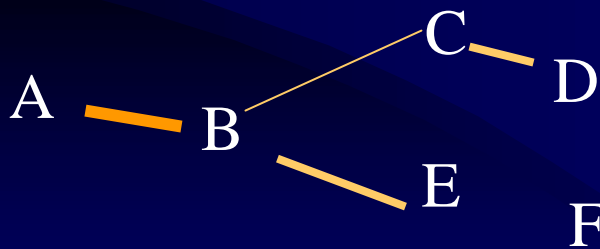
# Bayes Networks - An Intuitive Primer

- Imagine rolling a die 600 times to see if it's fair.
  - $\chi^2$  test on all 6 counts determines if fair, individual count deviations indicate likely “weighted faces”

Face up	1	2	3	4	5	6
Times seen	70	120	105	120	75	110

# Bayes Networks - An Intuitive Primer

- Imagine rolling 6 dice (A, B,...F) 600 times:
  - Are they interacting with each other?
  - Are the dice “connected” (magnets, strings...)?
  - Does knowing what some dice do help to predict others?
  - Are some configurations more likely than expected?
  - How do you represent the interactions?



# A sample application

- 3T3L1 cells which undergo differentiation
- PPAR (Peroxisome Proliferator-Activated Receptors)
  - Diabetes
- Fibroblasts → Preadipocytes → Adipocytes
- Treat with vehicle (DMSO), various compounds
- What can we learn about PPAR-related mechanisms in this experiment?

# Input Data

- **Expression** data: 3T3L1 Phase 3 experiment
  - 46x3 experiments x 19100 probes
  - another 4.5K probes were not in Unigene, so excluded
  - multiple experiments combined and thresholded
- **Annotation** data: 20 selected functional categories from Bioknowledge database
  - input from Bei Zhang
- Probes on chip annotated via Unigene

# Annotation Source: GO Database

Provides labels descriptive of the

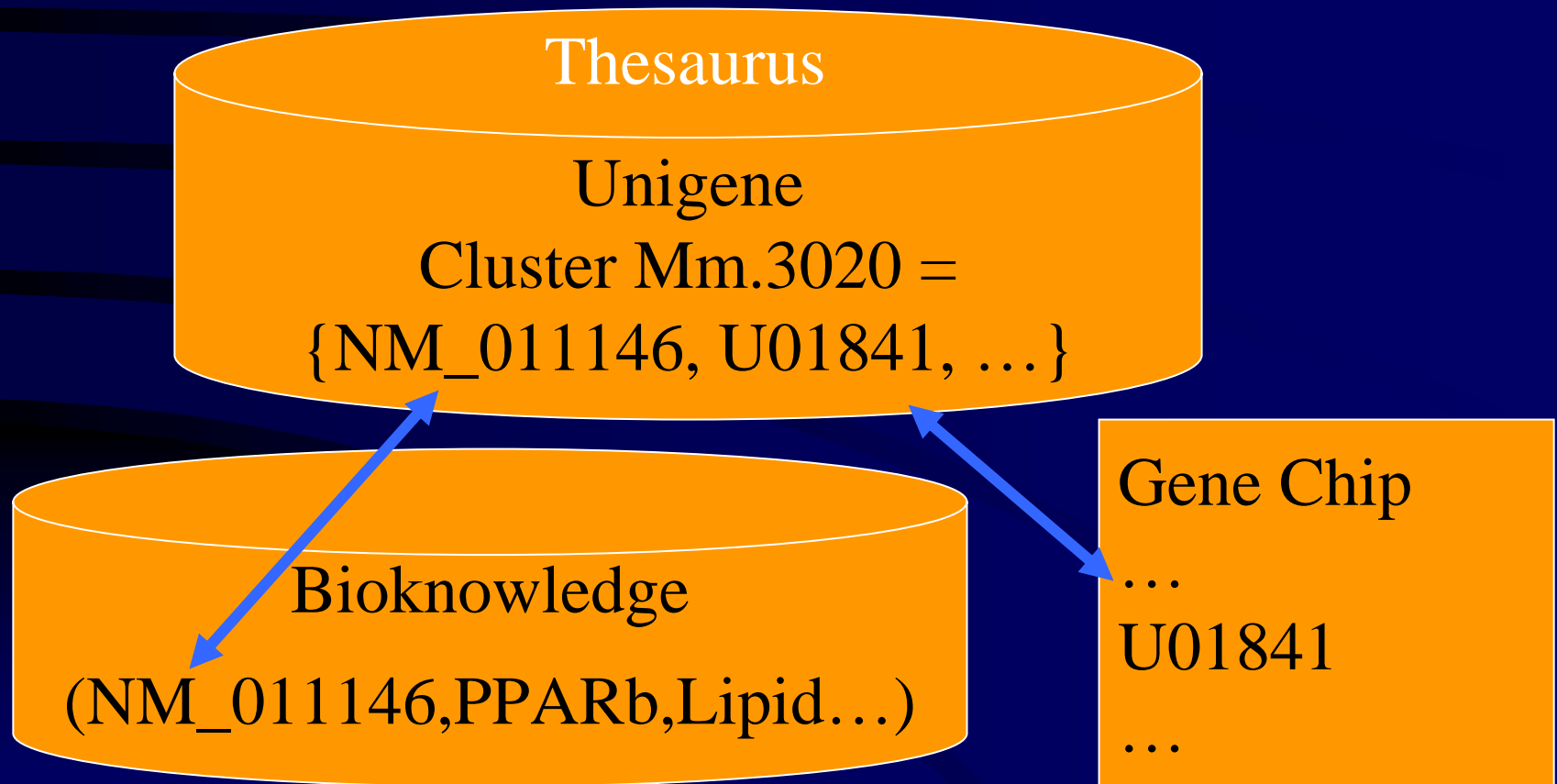
- molecular function,
- biological process, and
- cellular component

of gene products

- Organized into 3 “trees”

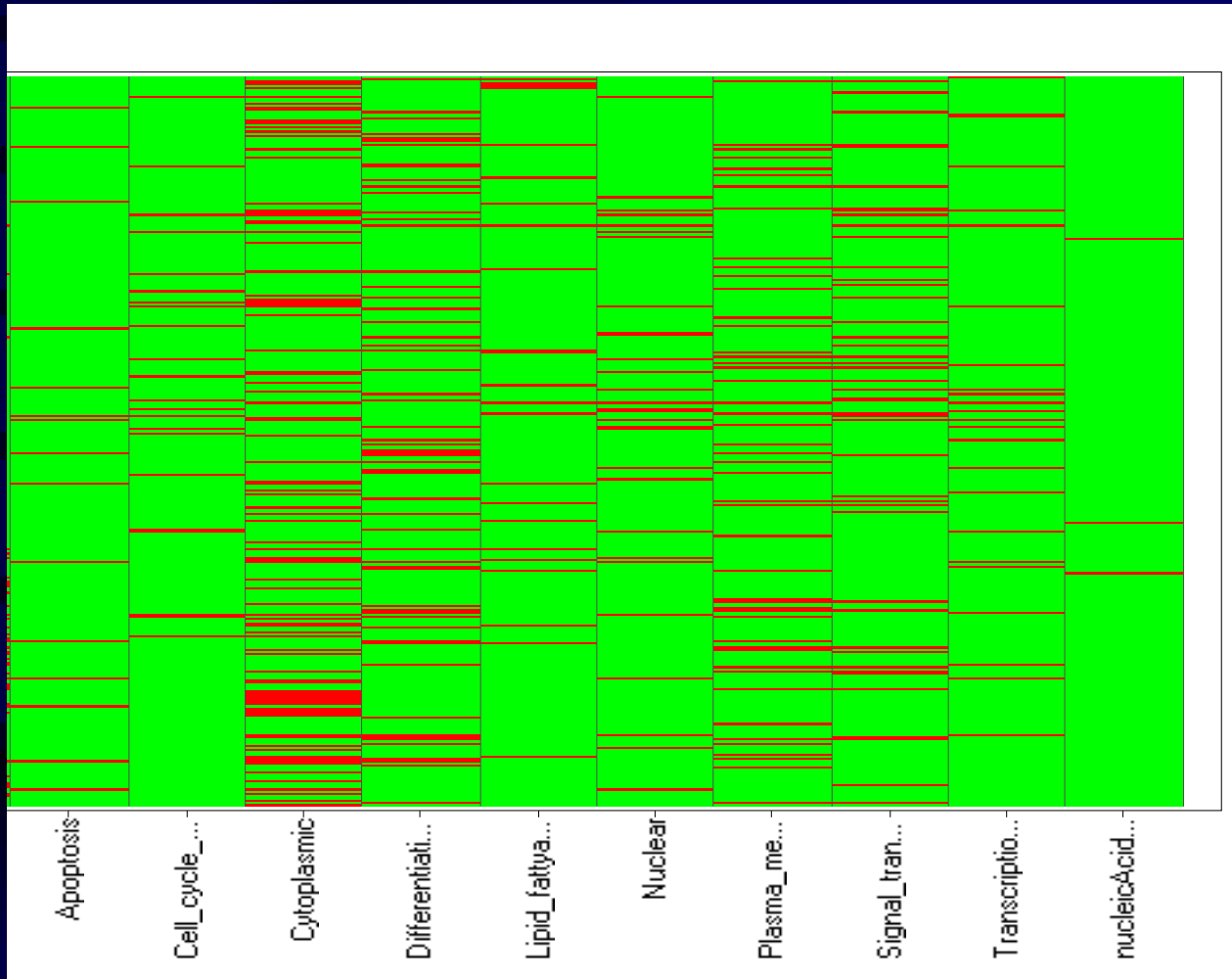
# Annotating Expression Data

How?: Use Unigene & Bioknowledge



# Prepared Annotation Table

17,774 gene clusters



20 Annotation columns

# Prepared Profile and Annotation Data

46

Expression Profiles

20

Functional Categories

19,100 Genes



Veh Control vs. Same

L-000000  
77.6M

Fibro vs. Adipo,  
NO DMSO

Apoptosis

Cell Cycle

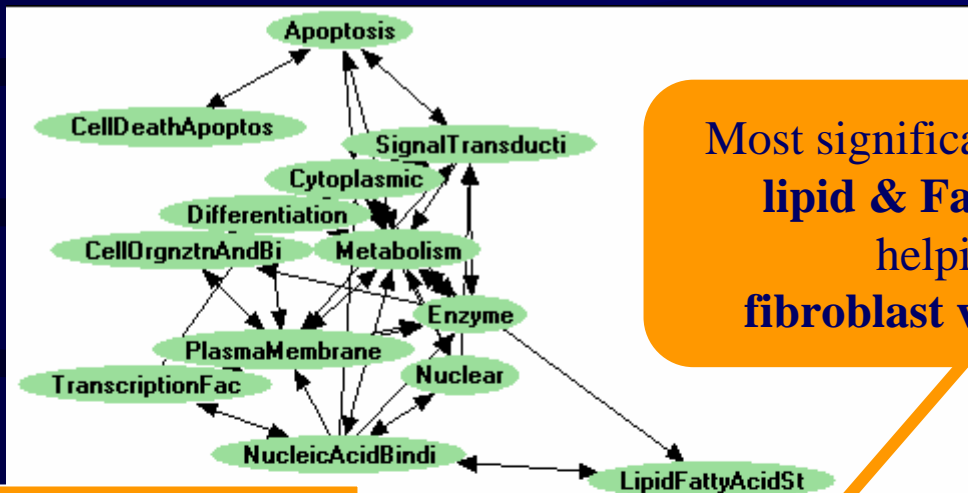
■ No Change   ■ Change

■ Does not have that annotation  
■ Has that annotation

# PF-nets: combine expression and annotation

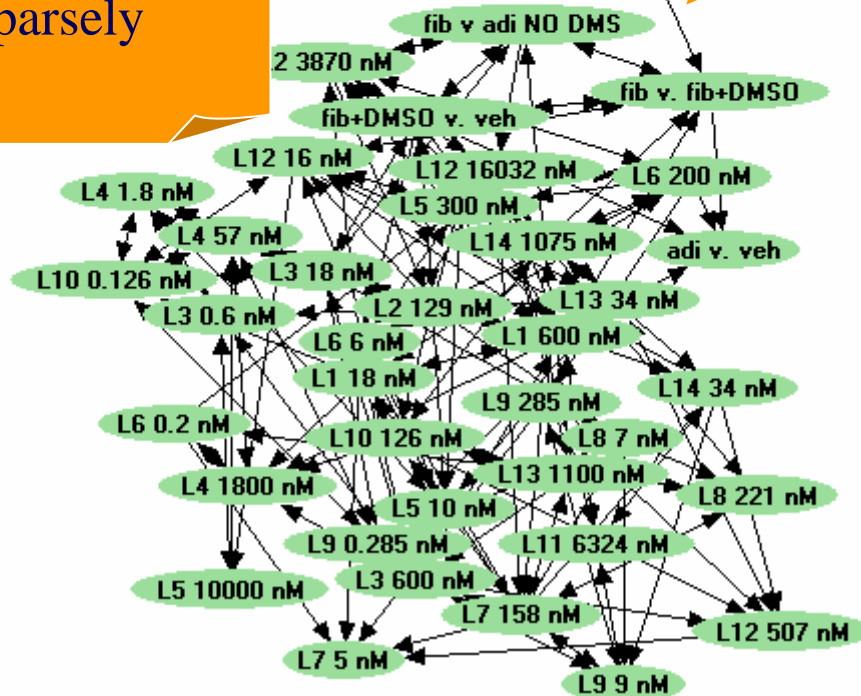
- Algorithm: fit a ML Dependency net or Bayes net
  - 46 expression variables (or more)
  - 20 annotation variables (or more)
- Predictions!
  - Expression  $\rightarrow$  Annotation
  - Annotation  $\rightarrow$  Expression
  - Expression, Annotation  $\rightarrow$  Expression
  - others

# A Complete PF-net

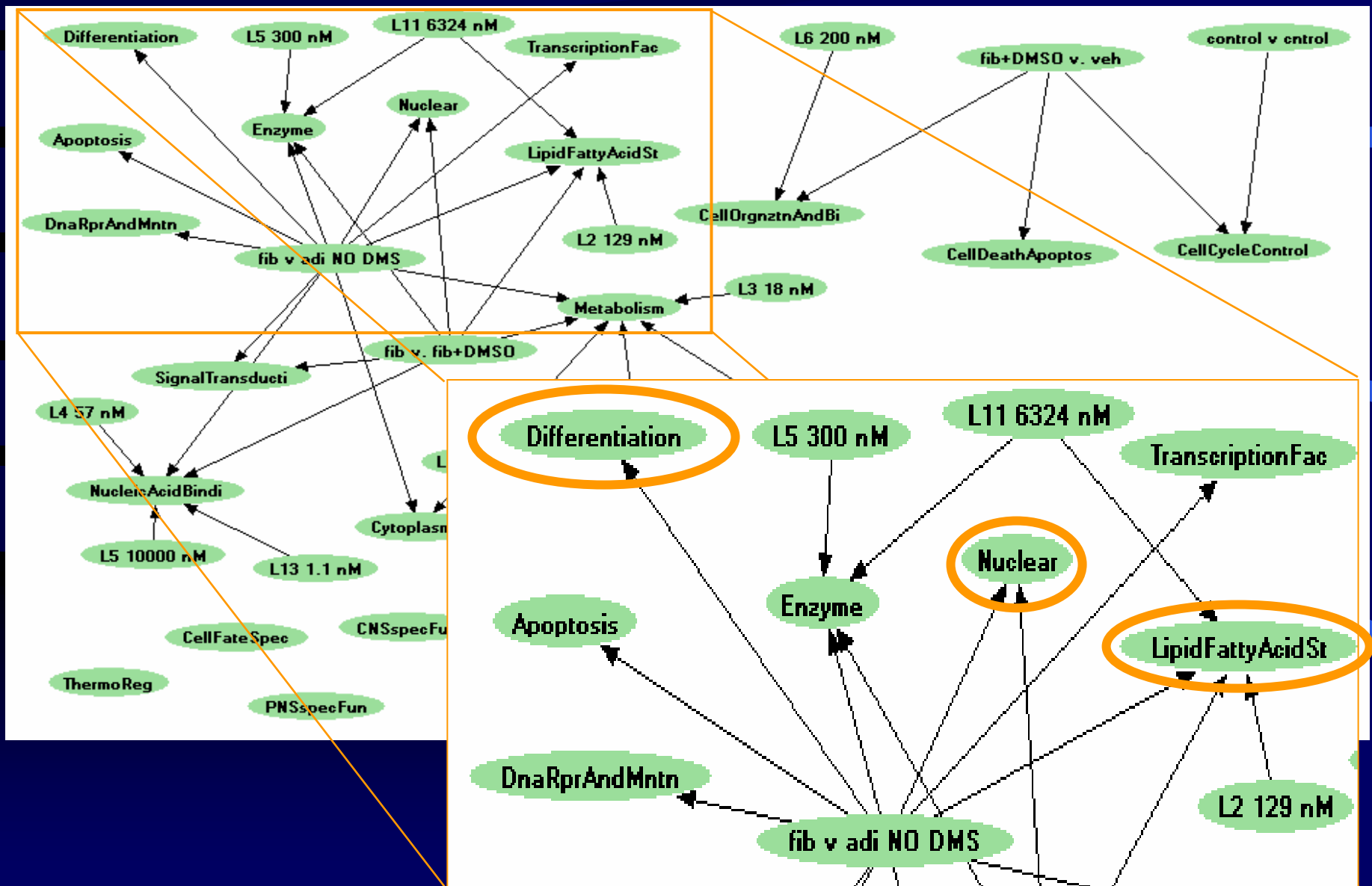


Most significant interconnection is  
**lipid & Fatty acid metabolism**  
 helping to predict  
**fibroblast vs fibroblast+DMSO**

Expression and annotation domains:  
 highly *intra*connected, sparsely  
*inter*connected



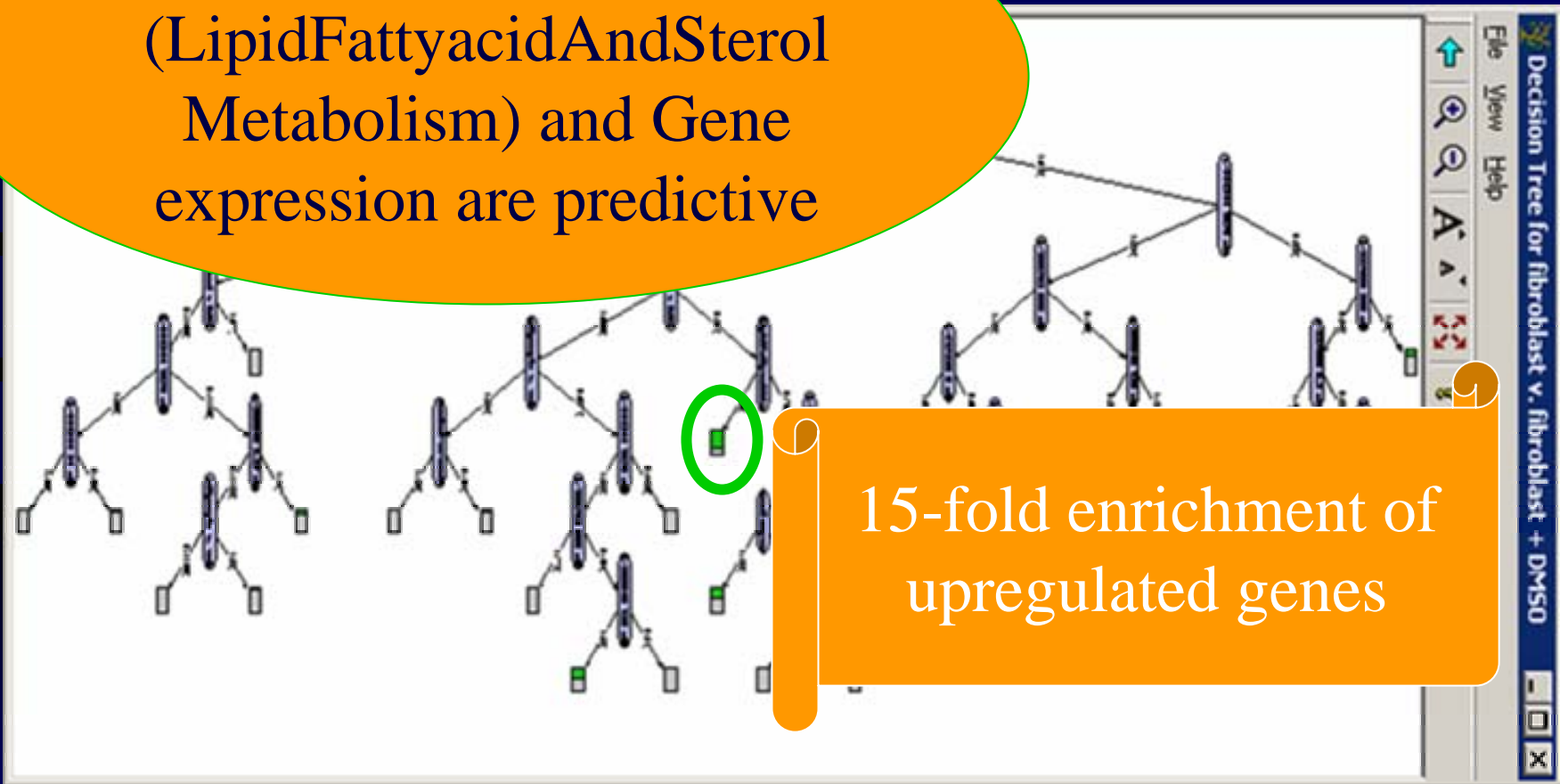
# Expression Predicts Annotation



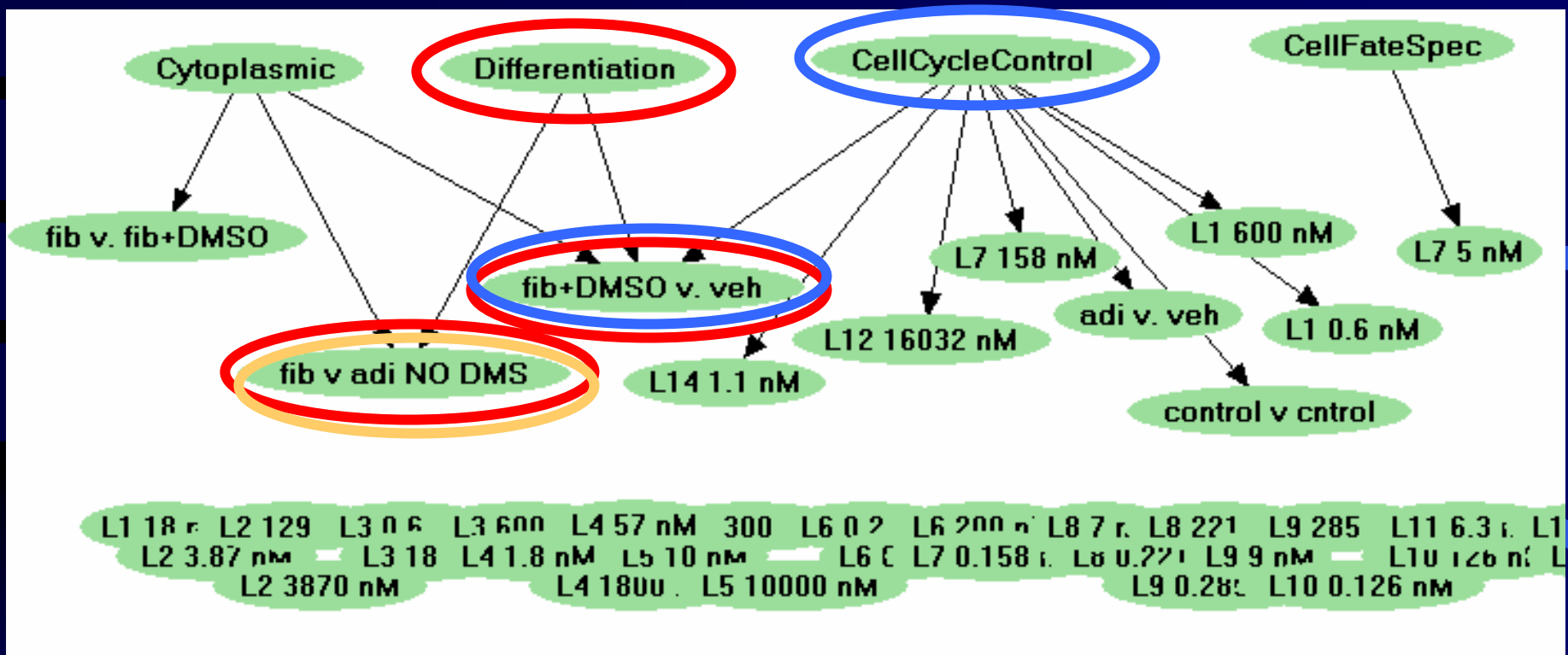
# PF Net for predicting Expression

(Using expression & annotation)

Both annotation  
(LipidFattyacidAndSterol  
Metabolism) and Gene  
expression are predictive



# Annotation Predicts Expression Changes!

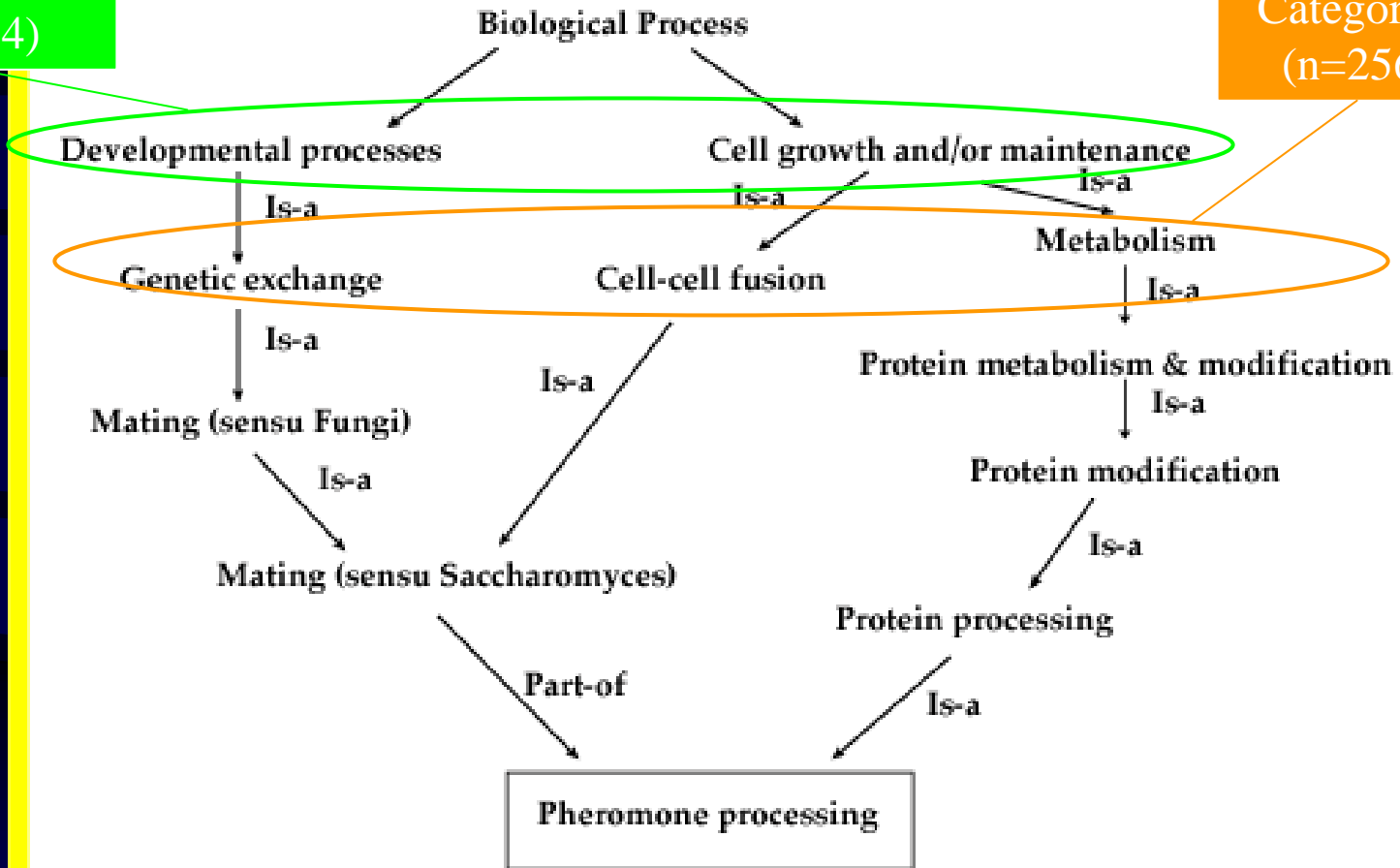


# PANs – Pure Annotation Networks

# GO Hierarchy

Level 1  
Categories  
(n=34)

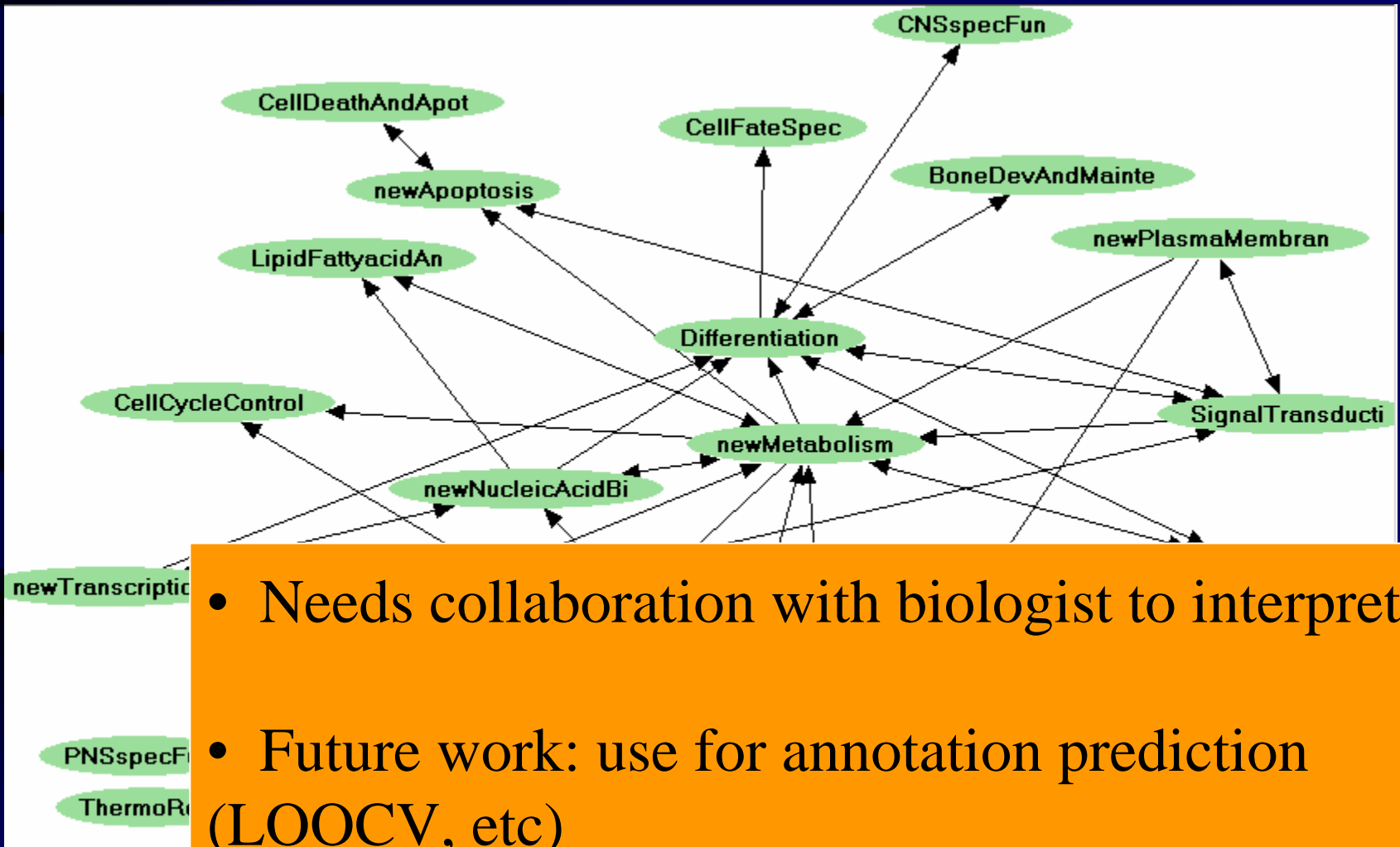
Level 2  
Categories  
(n=256)



# PAN – Pure Annotation Network

- Look for statistically significant correspondences between annotations
- Each Unigene cluster is treated as a single observation
- We used the 17,774 observations corresponding to the genes on the 3T3L1 chip (convenience)
  - 19K annotated unmasked probes
  - 2.2k probes with synonyms, multiple probes/gene
- Each observation has 0 or 1 under each annotation column
  - Columns used were for 3T3L1 (convenience)

# Unigene Partial PAN



*Fini*