

# High Performance Computing in Bioinformatics: A Focus on Next Generation Data Analysis Tools

---



ICCS 2013 – Computation at the Frontier of Science  
June 2013

Hesham H. Ali  
UNO Bioinformatics Core Facility  
College of Information Science and Technology

# Biomedical Research will never be the same

---

- IT advances promise to change Biosciences forever
- The availability of massive biological data shifted many branches in Biosciences from pure experimental disciplines to knowledge based disciplines
- Integrating Computational Sciences and Biosciences is critical but challenging
- Bioinformatics is the answer



# Few things they agreed on

---

... Health Care can only improve with the innovative application of Information Technology.

President Bush 2004

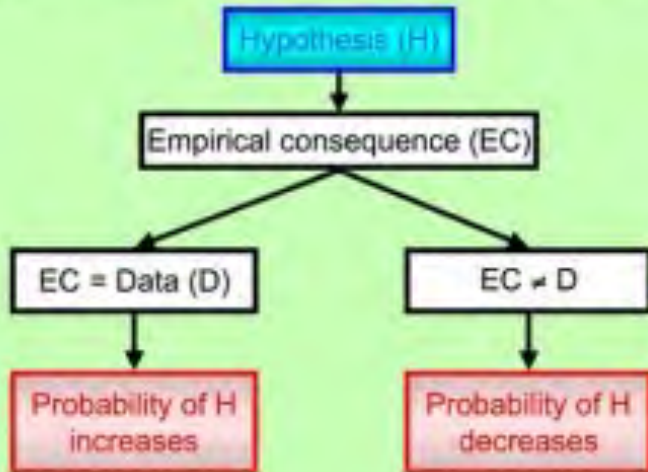
President Obama 2010



# A Potential Major Change

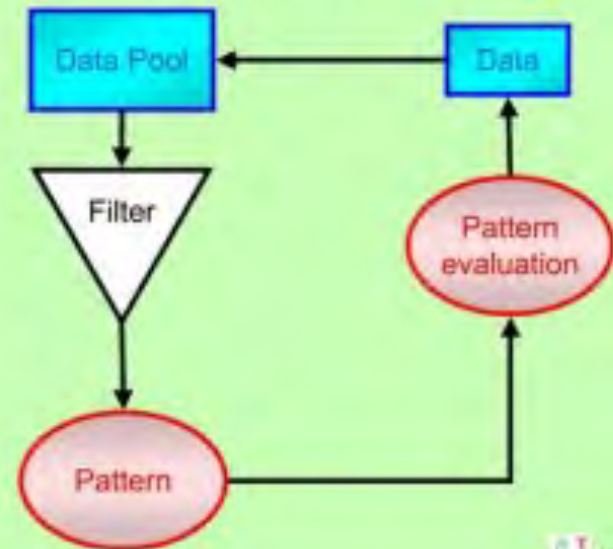
- Data driven research vs. Hypothesis driven research

## Hypothesis driven research - Concept



Leukippos Institute

## Data driven research - Concept



Leukippos Institute

# Bioinformatics technology: Where we are?

---

- High throughput data
- Next generation sequencing
- Personalized medicine
- Biomarkers
- Genome-wide association study
- Differentially expressed genes
- Single position variants and copy number variants
- ...



# Simple Question

---

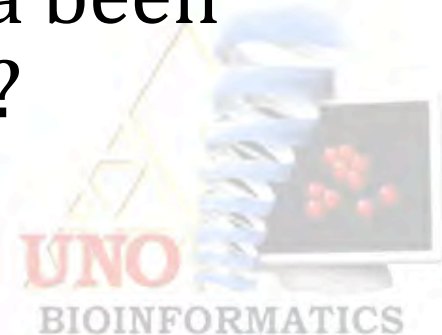
Where is the cure for cancer?

Why don't we have personalized medicine?

Why is AIDS still misunderstood?

*Can effectively be boiled down to:*

Why hasn't high-throughput data been effectively harnessed yet?



# Answer

---

It is not that easy:

- Complexity of the system
- Complexity of the organisms
- Size of the data (“big data”)
- Search space of inter-data relationships
- Heterogeneity of the data
- Lack of integration of data
- Computing power
- .....



## Robin Roberts: I will beat MDS Updated Mon June 11, 2012

"Good Morning America's" Robin Roberts is bravely facing a new health battle. The 51-year-old revealed Monday that five years after overcoming breast **cancer**, she's been diagnosed with a rare blood disorder that affects the bone marrow called...

<http://marquee.blogs.cnn.com/2012/06/11/robin-roberts-i-will-beat-mds/>

## Comedian Tommy Chong fighting prostate **cancer** Updated Sun June 10, 2012

Tommy Chong, one-half of the marijuana-loving "Cheech and Chong" comedy duo, is battling prostate **cancer**, he announced Saturday on CNN.

<http://www.cnn.com/2012/06/09/showbiz/chong-prostate-cancer/index.html>

## What can be done about the deepening polarization in America? Updated Wed June 6, 2012

By CNN's Jack Cafferty: The polarization of America is like a **cancer** that is slowly killing us. And like many forms of **cancer**, there appears to be no **cure**. We are more severely divided now than at any time in the last 25 years according to a new pew...

<http://caffertyfile.blogs.cnn.com/2012/06/06/what-can-be-done-about-the-deepening-polarization-in-america/>

## Experimental drug offers new way to battle certain breast **cancer** Updated Sun June 3, 2012

Doctors who treat breast **cancer** patients are very excited about an experimental drug that presents a whole new way of knocking out **cancer** cells.

<http://www.cnn.com/2012/06/03/health/breast-cancer-drug/index.html>

### Cancer Treatments

Sponsored Links

[www.hope4cancer.com/Treatments](http://www.hope4cancer.com/Treatments) - Natural Alternative **Treatments**. Call Us For A Free Consultation!

### New Hope for Cancer

[www.newhopemedicalcenter.com/](http://www.newhopemedicalcenter.com/) - Noninvasive alternative **treatments** to rebuild the immune system.

### Natural Cancer Treatment

[www.immunologyfoundation.org/](http://www.immunologyfoundation.org/) - New therapy removes TNF inhibitors, unblocking your immune response.



# But....

---

- Her2+ BC
  - 20-25% of breast cancers
  - Normal treatment: Herceptin
    - Eventually stops working if cancer comes back
- T-DM1 Drug
  - Trojan horse drug
  - 3 extra months of improvement
  - Lack of usual side effects



# Personalized Medicine

Resource

Cell

## Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes

Rui Chen,<sup>1,11</sup> George I. Mias,<sup>1,11</sup> Jennifer Li-Pook-Than,<sup>1,11</sup> Lihua Jiang,<sup>1,11</sup> Hugo Y.K. Lam,<sup>1,12</sup> Rong Chen,<sup>2,12</sup> Elana Miriami,<sup>1</sup> Konrad J. Karczewski,<sup>1</sup> Manoj Hariharan,<sup>1</sup> Frederick E. Dewey,<sup>3</sup> Yong Cheng,<sup>1</sup> Michael J. Clark,<sup>1</sup> Hogune Im,<sup>1</sup> Lukas Habegger,<sup>6,7</sup> Suganthi Balasubramanian,<sup>6,7</sup> Maeve O'Huallachain,<sup>1</sup> Joel T. Dudley,<sup>2</sup> Sara Hillenmeyer,<sup>1</sup> Rajini Haraksingh,<sup>1</sup> Donald Sharon,<sup>1</sup> Ghia Euskirchen,<sup>1</sup> Phil Lacroute,<sup>1</sup> Keith Bettinger,<sup>1</sup> Alan P. Boyle,<sup>1</sup> Maya Kasowski,<sup>1</sup> Fabian Grubert,<sup>1</sup> Scott Seki,<sup>2</sup> Marco Garcia,<sup>2</sup> Michelle Whirl-Carrillo,<sup>1</sup> Mercedes Gallardo,<sup>9,10</sup> Maria A. Blasco,<sup>9</sup> Peter L. Greenberg,<sup>4</sup> Phyllis Snyder,<sup>1</sup> Teri E. Klein,<sup>1</sup> Russ B. Altman,<sup>1,5</sup> Atul J. Butte,<sup>2</sup> Euan A. Ashley,<sup>3</sup> Mark Gerstein,<sup>6,7,8</sup> Kari C. Nadeau,<sup>2</sup> Hua Tang,<sup>1</sup> and Michael Snyder<sup>1,\*</sup>

<sup>1</sup>Department of Genetics, Stanford University School of Medicine

<sup>2</sup>Division of Systems Medicine and Division of Immunology and Allergy, Department of Pediatrics

<sup>3</sup>Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicine

<sup>4</sup>Division of Hematology, Department of Medicine

<sup>5</sup>Department of Bioengineering

Stanford University, Stanford, CA 94305, USA



# Methods

---

- 54yr old male volunteer
- Plasma and serum used for testing
- 14 month time course
- Complete medical exams and labs at each meeting (20 time points total)
- Extensive sampling at 2 periods of viral infection:
  - HRV (human rhinovirus) - common cold
  - RSV (respiratory syntical) - bronchitis



# Time-course summary

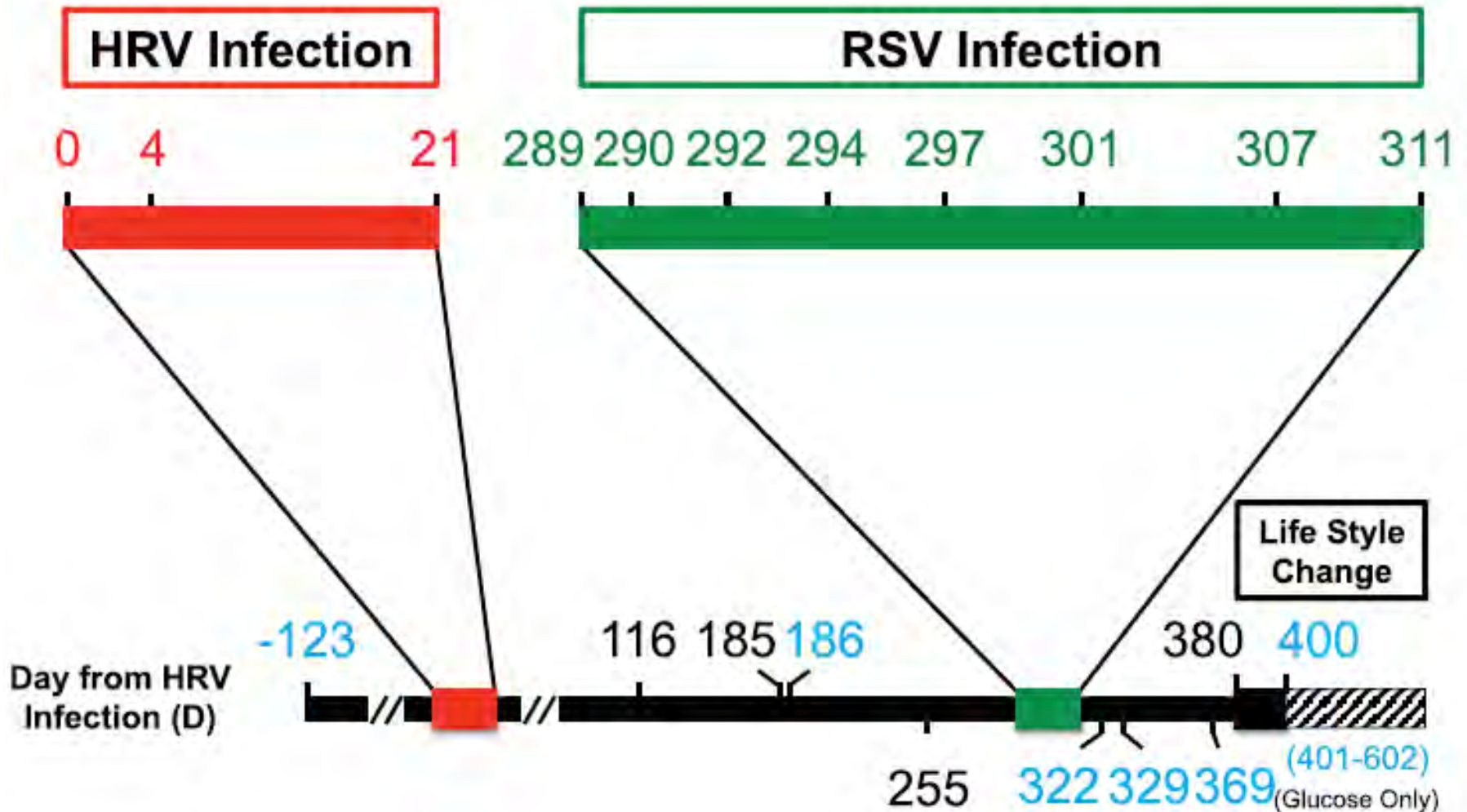
726 days total

HRV - red

RSV - green

Fasting - blue

Lifestyle change: ↑exercise, took ibuprofen daily, ↓sugar intake

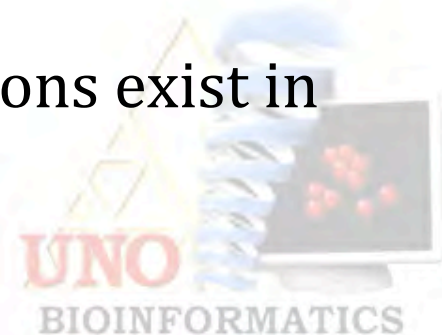


# Methods

---

- Whole-genome sequencing
  - Complete Genomics Deep WGS and Illumina
    - CG 35nt paired end
    - Illumina 100 nt paired end
    - 150 and 120-fold coverage
- 91% mapped to human RefSeq
- Remaining 9%:
  - 1,425 contigs mapped to non-RefSeq data
  - Remaining were unique
    - 2919 exons expressed using RNA-Seq

“A large number of undocumented genetic regions exist in the personal human genome”



# WBS Based Disease Risk Eval

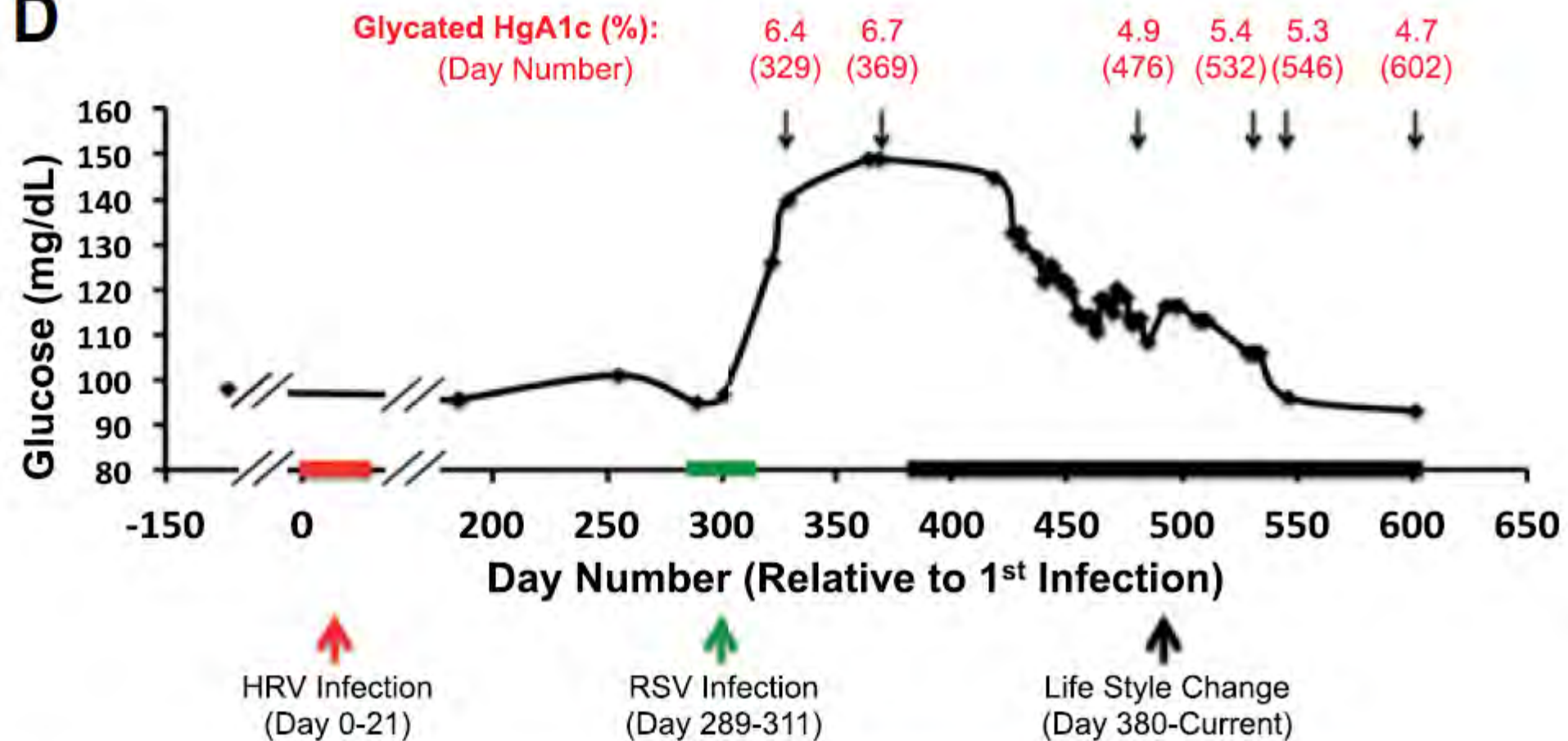
- 51 SNVs and 4 indels found in OMIM
  - LOF mutations
  - Validated by Sanger sequencing
- High interest genes:
  - Serpina1
  - TERT (acquired aplastic anemia)
  - Diabetes and hypertension: GKCR, KCNJ11, TCF7
- BASED ON THESE MUTATIONS:
  - Medical phenotype monitoring
    - Monitor blood glucose levels

**Table 2. Summary of Disease-Related Rare Variants**

Category	Count
Total high confidence rare SNVs	289,989
Coding	2,546
Missense	1,320
Synonymous	1,214
Nonsense	11
Nonstop	1
Damaging or possibly damaging	233
Putative loss-of-function SNVs <sup>a</sup>	51
Total high confidence rare indels	51,248
Coding indels	61
Frameshift indels	27
miRNA indels	3
miRNA target sequence indels	5
Putative loss-of-function indels <sup>a</sup>	4

<sup>a</sup>In curated Mendelian disease genes.



**D**

# Dynamic Omics Analysis

---

- Transcriptome: RNA-Seq of 20 time points
  - 2.67 billion paired end reads
  - 19,714 isoforms for 12,659 genes tracked
- Proteome: Quantification of 6,280 proteins
  - 14 time points via TMT and LC/MS
- Metabolome: 1,020 metabolites tracked during viral infections
  - miRNA analysis also during HRV infection





# Techniques Used

- Summary of techniques used:
  - Sample collection
  - HRV and RSV detection
  - Whole-genome sequencing
  - Whole-exome sequencing
  - Sanger-DNA sequencing
  - Whole-transcriptome sequencing: mRNA-Seq
  - Small RNA sequencing: microRNA-Seq
  - Serum Shotgun Proteome Profiling
  - Serum Metabolome Profiling
  - Serum Cytokine Profiling
  - Autoantibodyome Profiling
  - Telomere Length Assay
  - Genome Phasing
  - Omics Data Analysis



# Team Count: 41

Resource

Cell

## Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes

Rui Chen,<sup>1,11</sup> George I. Mias,<sup>1,11</sup> Jennifer Li-Pook-Tham,<sup>1,11</sup> Lihua Jiang,<sup>1,11</sup> Hugo Y.K. Lam,<sup>1,12</sup> Rong Chen,<sup>2,12</sup> Elana Miriami,<sup>1</sup> Konrad J. Karczewski,<sup>1</sup> Manoj Hariharan,<sup>1</sup> Frederick E. Dewey,<sup>3</sup> Yong Cheng,<sup>1</sup> Michael J. Clark,<sup>1</sup> Hogune Im,<sup>1</sup> Lukas Habegger,<sup>6,7</sup> Suganthi Balasubramanian,<sup>6,7</sup> Maeve O'Huallachain,<sup>1</sup> Joel T. Dudley,<sup>2</sup> Sara Hillenmeyer,<sup>1</sup> Rajini Haraksingh,<sup>1</sup> Donald Sharon,<sup>1</sup> Ghia Euskirchen,<sup>1</sup> Phil Lacroute,<sup>1</sup> Keith Bettinger,<sup>1</sup> Alan P. Boyle,<sup>1</sup> Maya Kasowski,<sup>1</sup> Fabian Grubert,<sup>1</sup> Scott Seki,<sup>2</sup> Marco Garcia,<sup>2</sup> Michelle Whirl-Carrillo,<sup>1</sup> Mercedes Gallardo,<sup>9,10</sup> Maria A. Blasco,<sup>9</sup> Peter L. Greenberg,<sup>4</sup> Phyllis Snyder,<sup>1</sup> Teri E. Klein,<sup>1</sup> Russ B. Altman,<sup>1,5</sup> Atul J. Butte,<sup>2</sup> Euan A. Ashley,<sup>3</sup> Mark Gerstein,<sup>6,7,8</sup> Kari C. Nadeau,<sup>2</sup> Hua Tang,<sup>1</sup> and Michael Snyder<sup>1,\*</sup>

<sup>1</sup>Department of Genetics, Stanford University School of Medicine

<sup>2</sup>Division of Systems Medicine and Division of Immunology and Allergy, Department of Pediatrics

<sup>3</sup>Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicine

<sup>4</sup>Division of Hematology, Department of Medicine

<sup>5</sup>Department of Bioengineering

Stanford University, Stanford, CA 94305, USA



# State of the Field - Bioinformatics

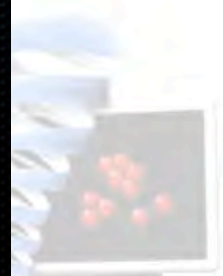
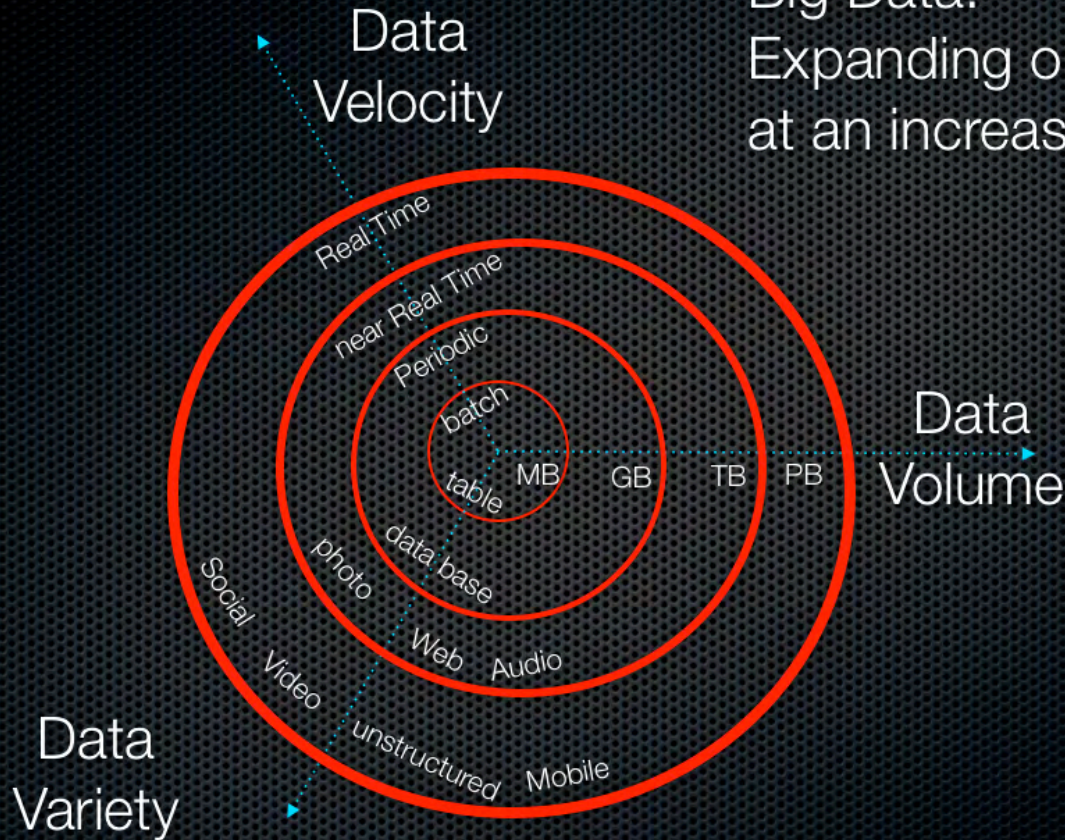
---

- Availability of many large useful database systems; private and public
- Availability of numerous helpful software packages
- Lack of data integration and trendiness of the discipline
- Fragmented efforts by computational scientists and bioscientists
- Advances in new technologies as high throughput next generation sequencing
- Increasing interest among researchers and educators



# Big Biological Data

Big Data:  
Expanding on 3 fronts  
at an increasing rate.



# Issues: Current Biological Databases

---

- The large degree of heterogeneity of the available data in terms of quality, completeness and format
- The available data are mostly in raw format and significant amount of processing is needed to take advantage of it
- Archival data used for research - mostly available in semi flat files – hence the lack of structure that support advanced searching and data mining



# Data versus Knowledge

---

- With high throughput data collection, Biology needs ways not only to store data but also to store knowledge (Smart data)
- Data: Things that are measured
- Information: Processed data
- Knowledge: Processed data plus meaningful relationships between measured entities
- Decision support systems



# Data Generation vs. Data Analysis/ Integration

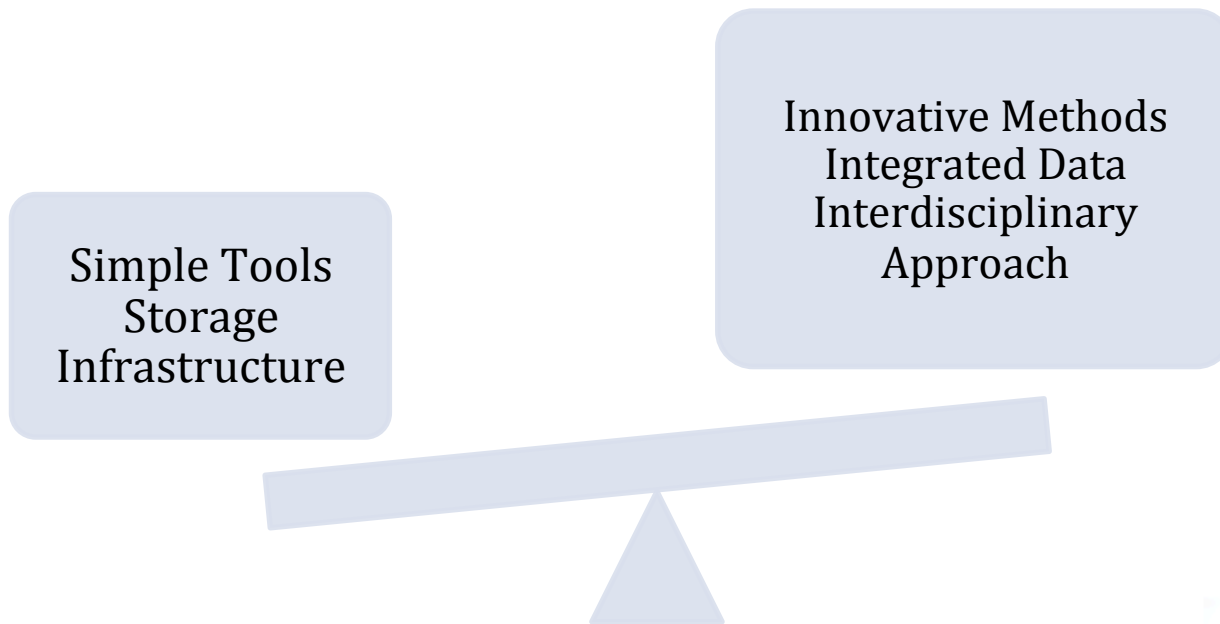
---

- New technologies lead to new data:
  - Competition to have the latest technology
  - Focus on storage needs to store yet more data
- Bioinformatics community needs to move from a total focus on data generation to a blended focus of measured data generation (to take advantage of new technologies) and data analysis/interpretation/visualization
- How do we leverage data? Integratable? Scalable?
- From Data to Information to Knowledge to Decision making



# Tipping the Balance

---





# So what do we really need?

---

- Advanced Tools – a new model:
  - Beyond surface-level adaptation of previous algorithms
- Systems approach
  - Take into consideration relationships and interactions among the various biological processes
- Genuine integration of computational methods and Bioinformatics data

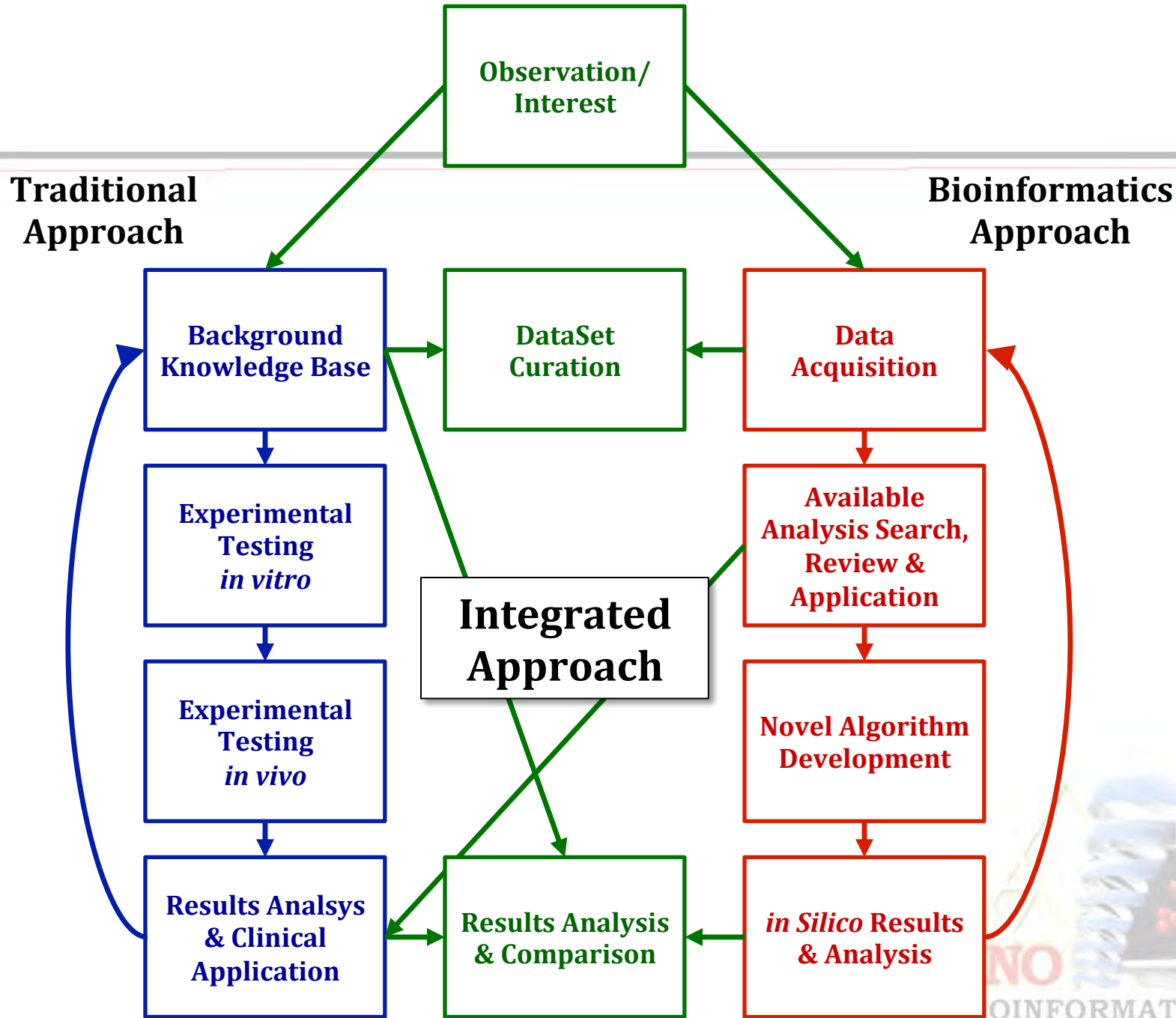


# First Generation Bioinformatics Tools

---

- Filled an important gap
- Mostly data independent
- Based on standard computational techniques
- Has little room for incorporating biological knowledge
- Developed in isolation
- Focus on trendy technologies
- Lack of data integration
- Lack of embedded assessment





# Next Generation Tools

---

- Dynamic: Custom built and domain dependent
- Collaborative: Incorporate biological knowledge and expertise
- Intelligent: based on a learning model that gets better with additional data/information

Intelligent Collaborative Dynamic (ICD) Tools



# Case Study: The Sequence Identification Problem

---

- Identification of organisms using obtained sequences is a very important problem
- Relying on wet lab methods only is not enough
- Employing identification algorithms using signature motifs to complement the experimental approaches
- Currently, no robust software tool is available for aiding researchers and clinicians in the identification process
- Such a tool would have to utilize biological knowledge and databases to identify sequences
- Issues related to size of data and quality of data are suspect and would need to be dealt with



# The Computational Approach

---

- Sequence similarity and graph clustering are employed to identify unknown sequences
- Earlier results were not conclusive
- Local similarity in specific regions rather than global similarity is used, in particular, test validity of identifying *Mycobacterium* based on ITS region and 16S region
- Graph Clustering based on region similarity produced very good results, particularly when using ITS region
- Grammar based description of selected regions is used for identification



# The Mycobacterium Case Study

---

- 30 species associated with variety of human and animal diseases such as tuberculosis
- Certain pathogenic species specific to humans. Some only affect animals
- Certain pathogenic species are drug-resistant
- Laboratory identification slow, tedious, and error-prone
- Sequencing provides an alternative to laboratory methods
- Researchers wanted to test validity of identifying *Mycobacterium* based on ITS region and 16S region



# How to Define Region Preferences

---

- Simple Definition
  - Letters (ACGT)
  - Wild Card (N)
  - Limits (wild cards, mismatches, Region Size)
- Grammar Based Definition
  - Employs regular expression for flexible region definitions
  - Powerful and Robust but a bit more complex





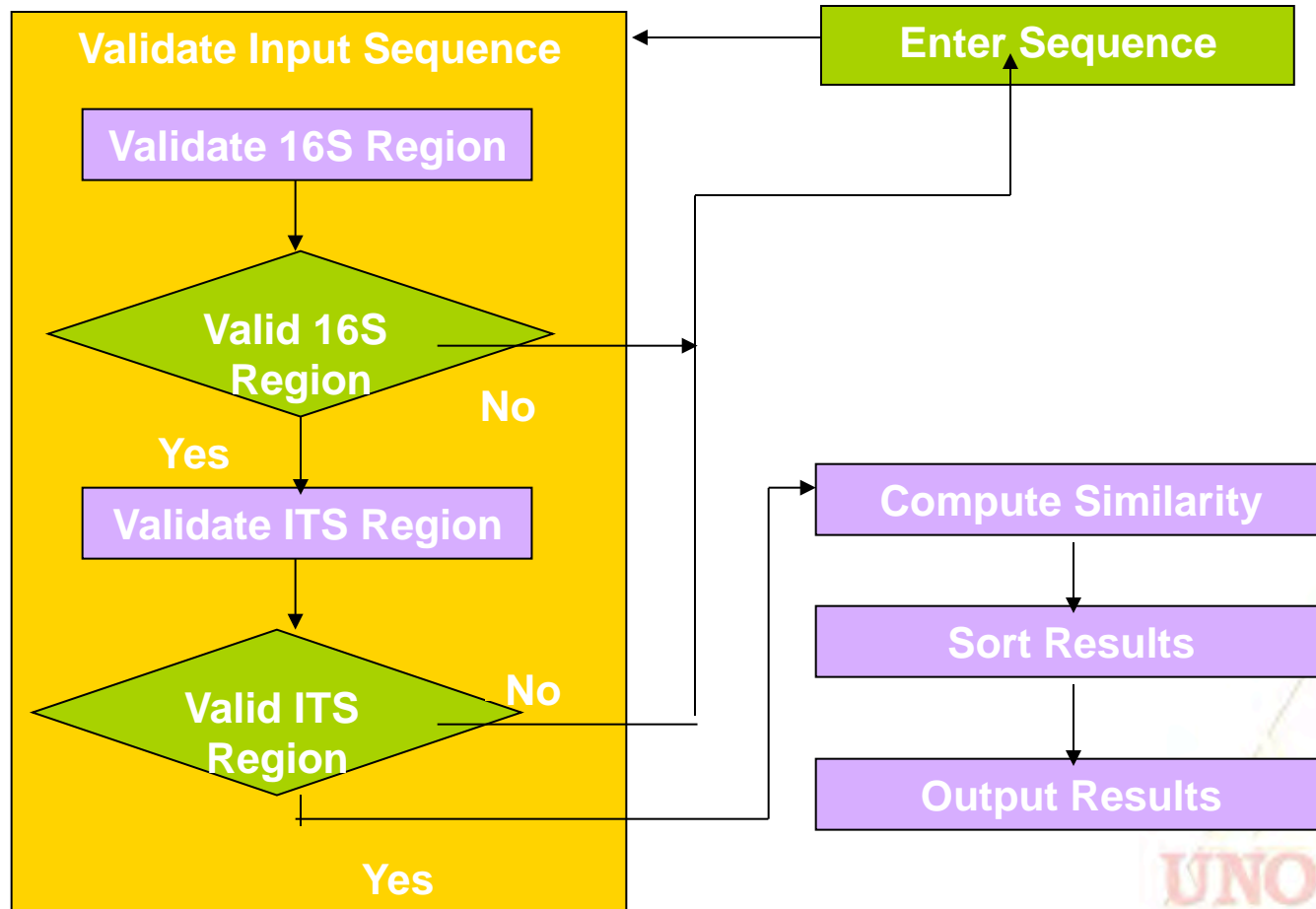
# Regular Expressions

Expression	Expression Name	Function
<u>Metacharacter Matches</u>		
. [...] [^...] <u>\char</u>	<u>wildcard match</u> <u>character class</u> <u>negated character class</u> <u>escaped character</u>	Match any one character Match any character inside braces Match any character not listed Matches to a terminal (e.g. \! becomes !)
<u>Counting Modulators</u>		
? + * <u>{min, max}</u>	<u>question</u> <u>plus</u> <u>star</u> <u>specified range</u>	One subsequence allowed, optional One required, more optional Any number allowed, but optional <u>min</u> required, <u>max</u> allowed
<u>Position Matchers</u>		
^ \$ \ \ <u>&gt;</u>	<u>caret</u> <u>dollar</u> <u>word boundary</u> <u>word boundary</u>	Matches position at start of word Matches position at end of word Matches position at beginning of word Matches position at end of word
<u>Clarification and Flexibility Operators: Topology Line Operators</u>		
 (,) <u>\1, \2, ...</u>	<u>alternation</u> <u>parentheses</u> <u>back-reference</u>	Matches subsequence before or after pipe Logically groups subsequences Matched pattern must occur again

Table 1: RegEx functionality

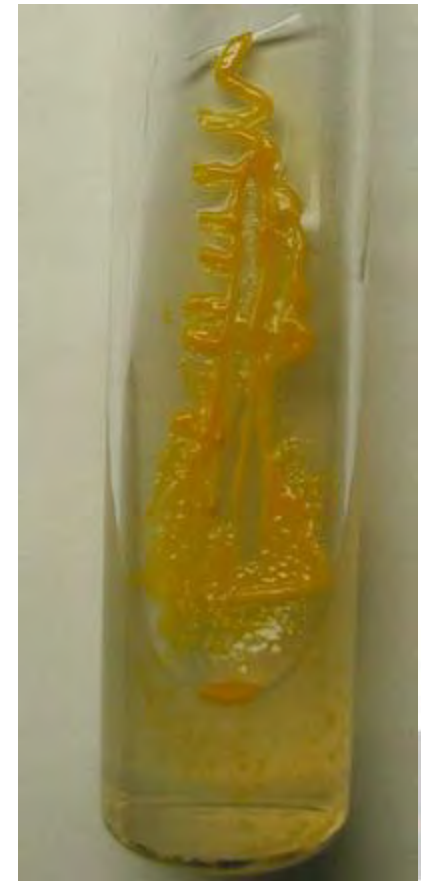


# Case Study: *Mycobacterium*



# Nebraska gets its very own organism

- ✚ While trying to pinpoint the cause of a lung infection in local cancer patients, they discovered a previously unknown micro-organism. And they've named it "mycobacterium nebraskense," after the Cornhusker state.
- ✚ It was discovered few weeks ago using Mycoalign: A Bioinformatics program developed at PKI



Source: Omaha World Herald,

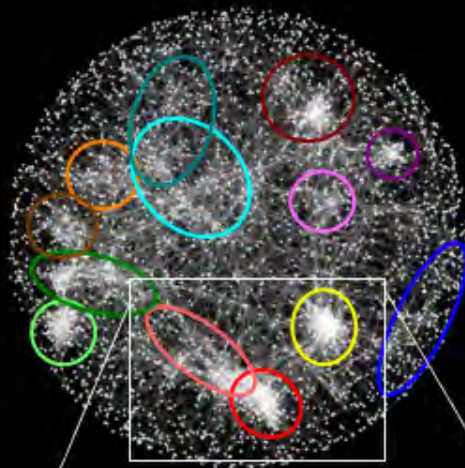
# The “Systems Biology” Approach

---

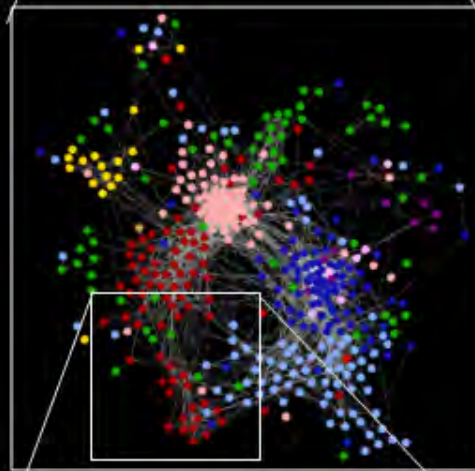
- Integrated Approach:
  - Networks model relationships, not just elements
  - Discover groups of relationships between genes
- Discovery
  - Examine changes in systems
    - Normal vs. diseased
    - Young vs. old
    - Stage I v. State II v. Stage III v. Stave IV



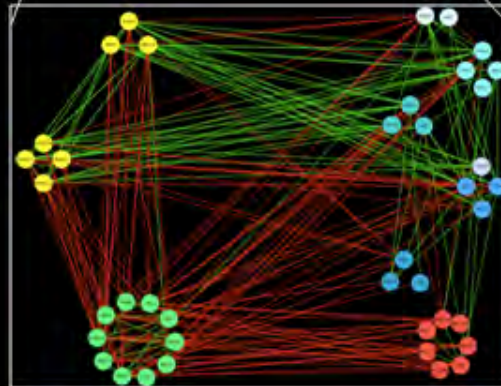
- Network
- System
  - Inter
  - Use
  - Pers



Global level



Process level

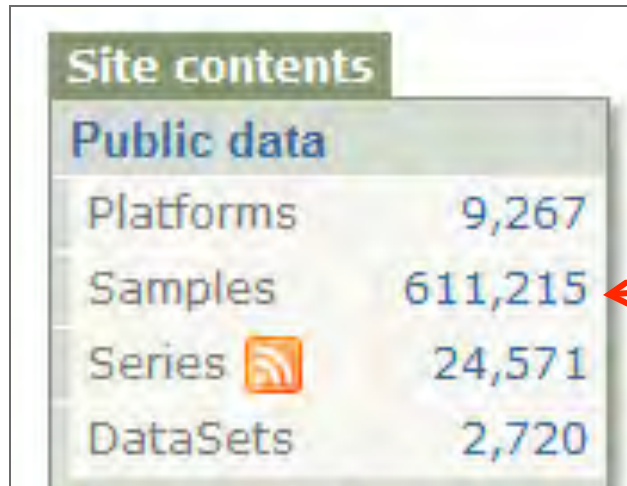



Pathway/complex level



# Why Networks?

- Explosion of biological data



Site contents	
Public data	
Platforms	9,267
Samples	611,215
Series 	24,571
DataSets	2,720

**Each sample can have over 40,000 genes**

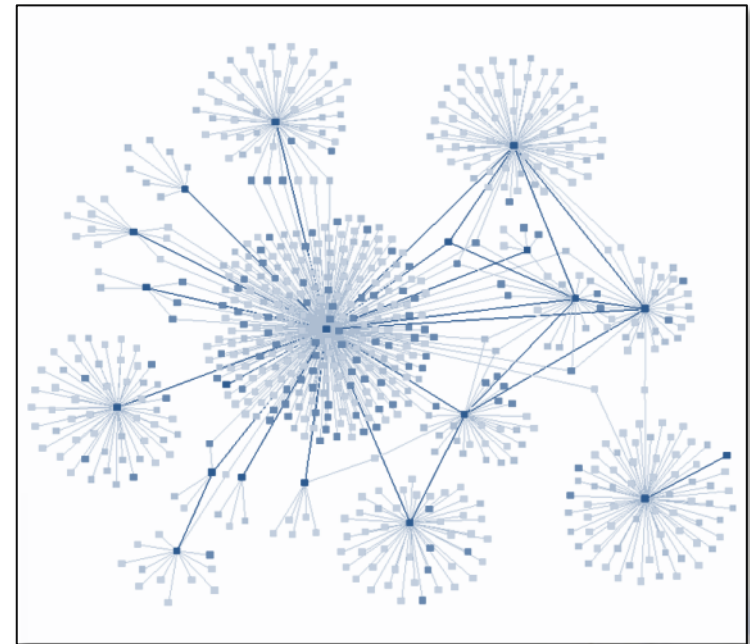
- Average microarray experiment: 1200 pages of data\*
- How can we extract information from data?



\* <http://www.mc.vanderbilt.edu/peerreview/fall026.html>

# Biological Networks

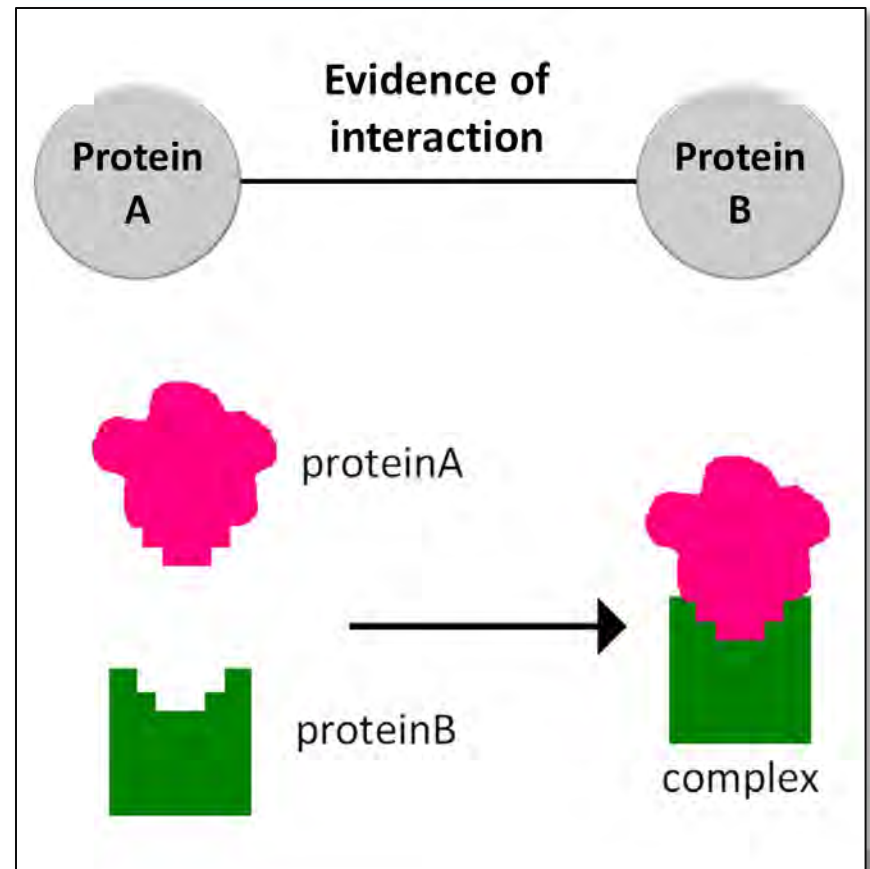
- A biological network represents elements and their interactions
- Nodes → elements
- Edges → interactions
- Can represent multiple types of elements and interactions



# Types of Biological Networks

- **Protein-protein interaction network**

- Synthetic lethality
- Metabolome
- Signal transduction
- Correlation/co-expression network



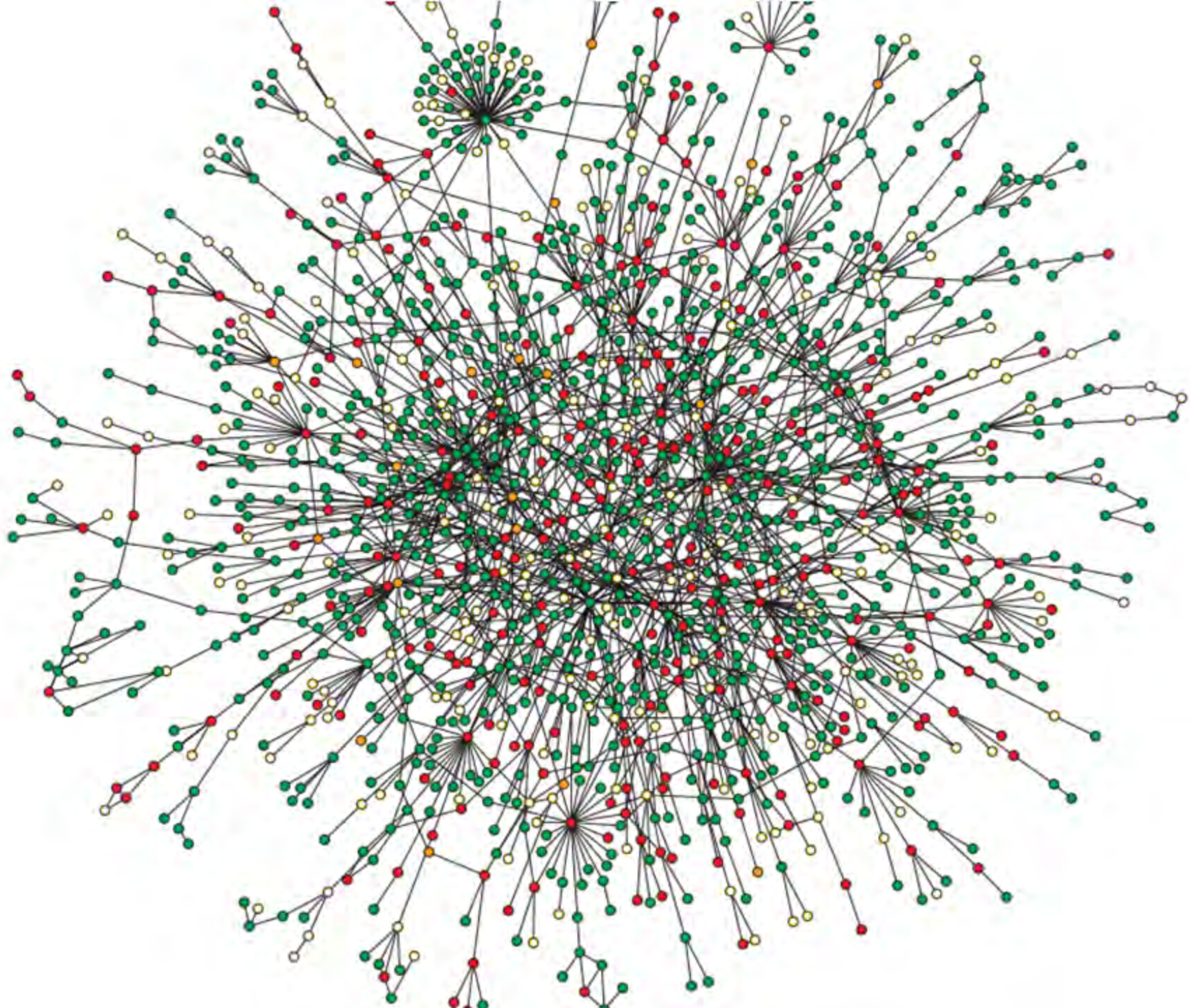


# Protein-Protein Interaction Networks

---

- “Hub” proteins in biological networks began from the study of PPI’s
- Study done by Jeong (2001)
  - 1870 proteins (nodes)
  - 2240 interactions (edges)
- Forms a scale-free network (*incomplete*)





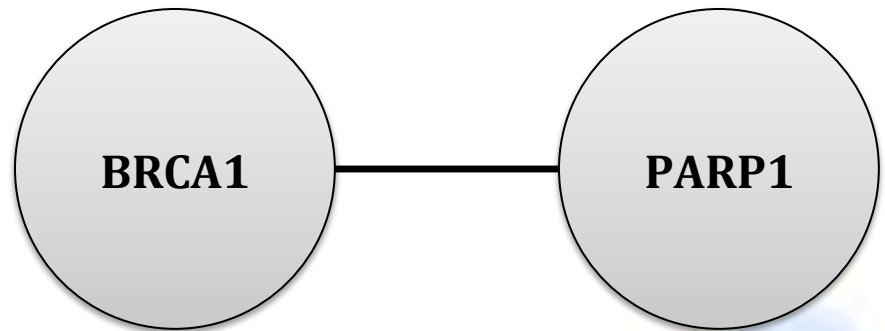
# Genetic Interaction Networks

BRCA1 + PARP1 → synthetic lethal interactors  
*Tumor cells only*

**Normal**



**Diseased**



PARP1 only  
expressed in  
tumor cells

# Genetic Interaction Networks Applications

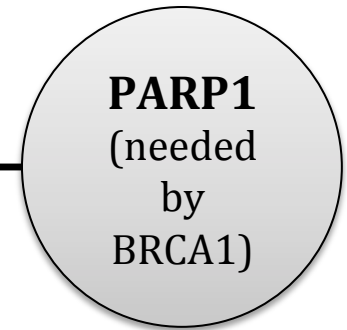
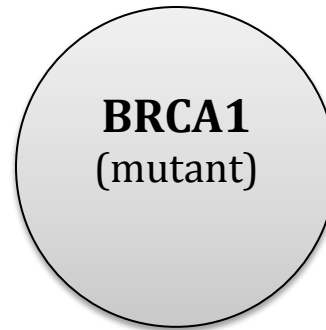
## Normal



Cell death by apoptosis

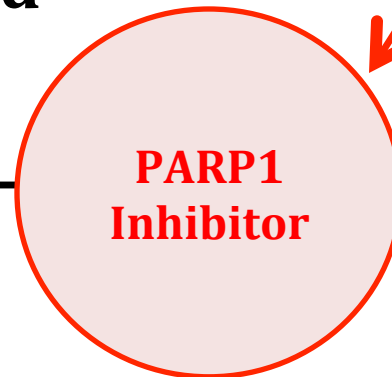
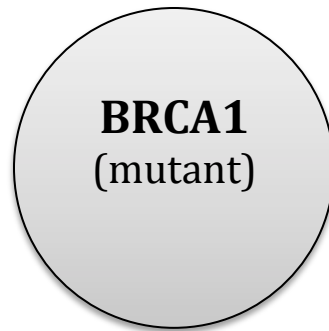


## Diseased



Uncontrolled cell growth

## Treated



Cell death by apoptosis

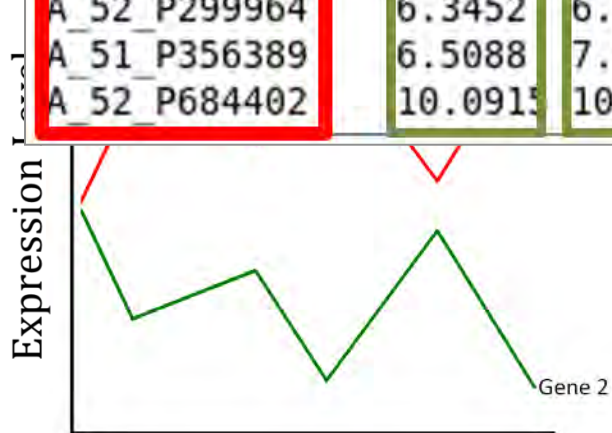


# Correlation Networks

Gene ID	Sample1	Sample2	Sample3	Sample4	Sample5
A_52_P616356	5.9813	6.0079	5.9525	7.2753	6.2
A_52_P580582	7.7845	7.7512	8.0943	8.3608	8.1
A_52_P403405	5.9301	6.5153	6.0526	7.1707	6.1
A_52_P819156	6.5748	6.8645	6.981	7.4937	6.9
A_51_P331831	7.1732	7.8754	7.7632	8.1875	7.6
A_51_P430630	6.0661	6.4009	5.9525	7.1208	6.6
A_52_P502357	5.936	6.3206	5.9525	7.1819	6.2
A_52_P299964	6.3452	6.8025	6.6457	7.3445	6.7
A_51_P356389	6.5088	7.0545	7.2346	7.631	7.0
A_52_P684402	10.0915	10.7124	10.2245	10.377	10.0

- 10,000-45,000+ probes
- UNO Blackforest cluster
- HCC Firefly

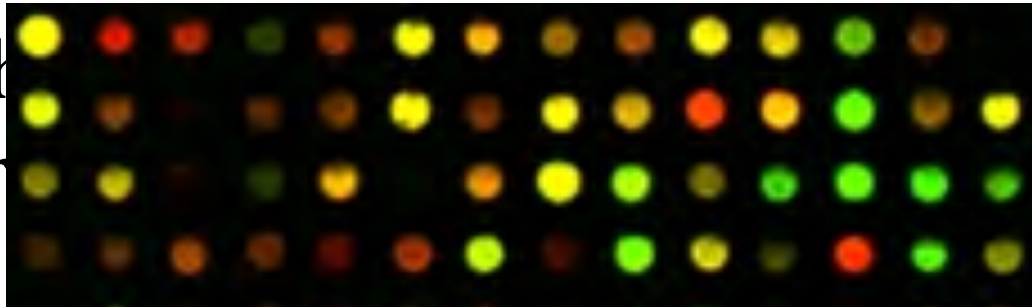
**Correlation = 1**



Sample

# Correlation Networks

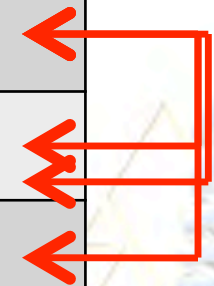
A graph  
corr



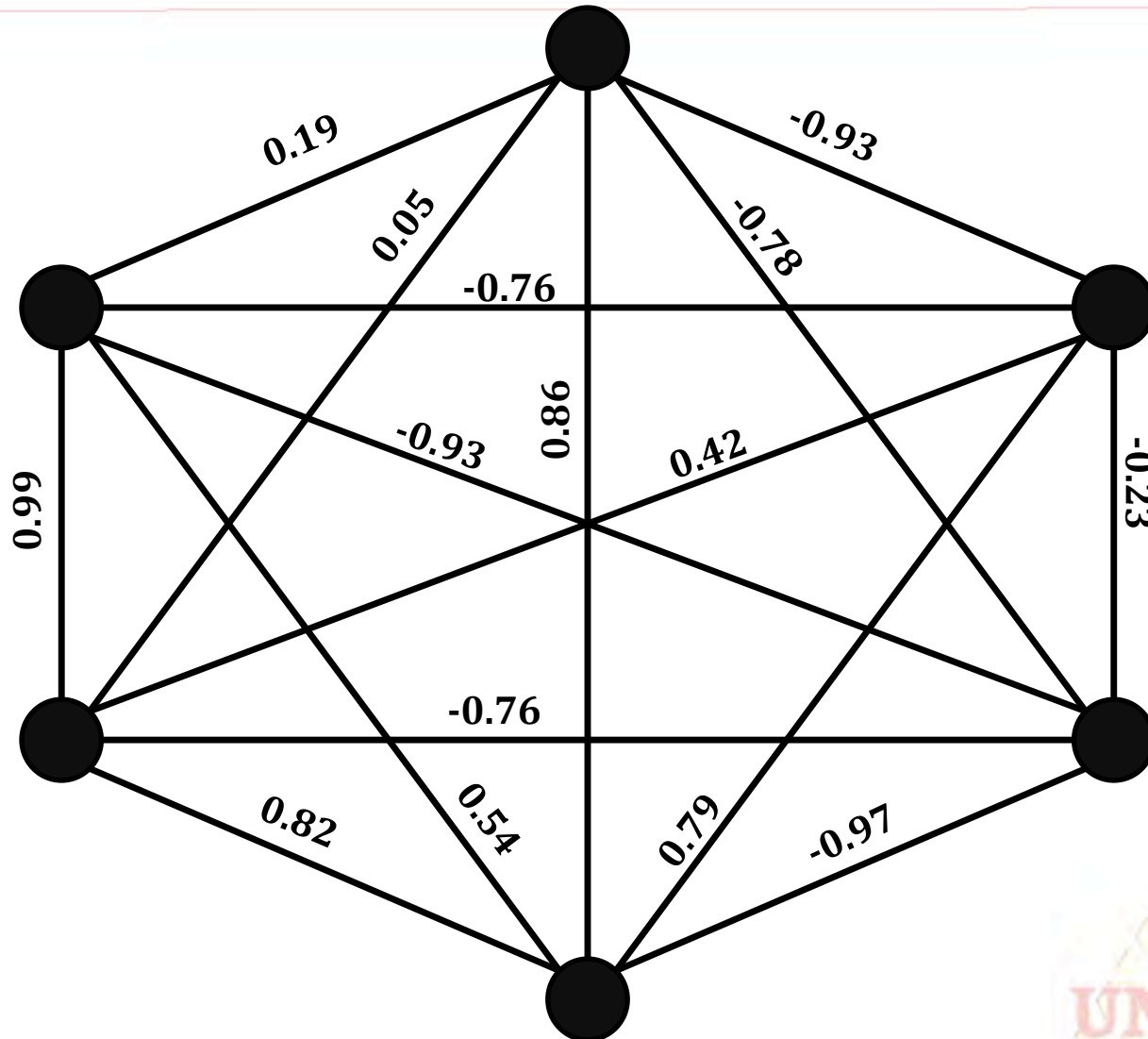
degree of  
entity



	Sample 1	Sample 2	Sample 3
Gene 1	10.5	11.0	12.1
Gene 2	3.2	3.3	2.9
Gene 3	1.4	1.5	0.9
Gene 4	7.8	7.1	8.2

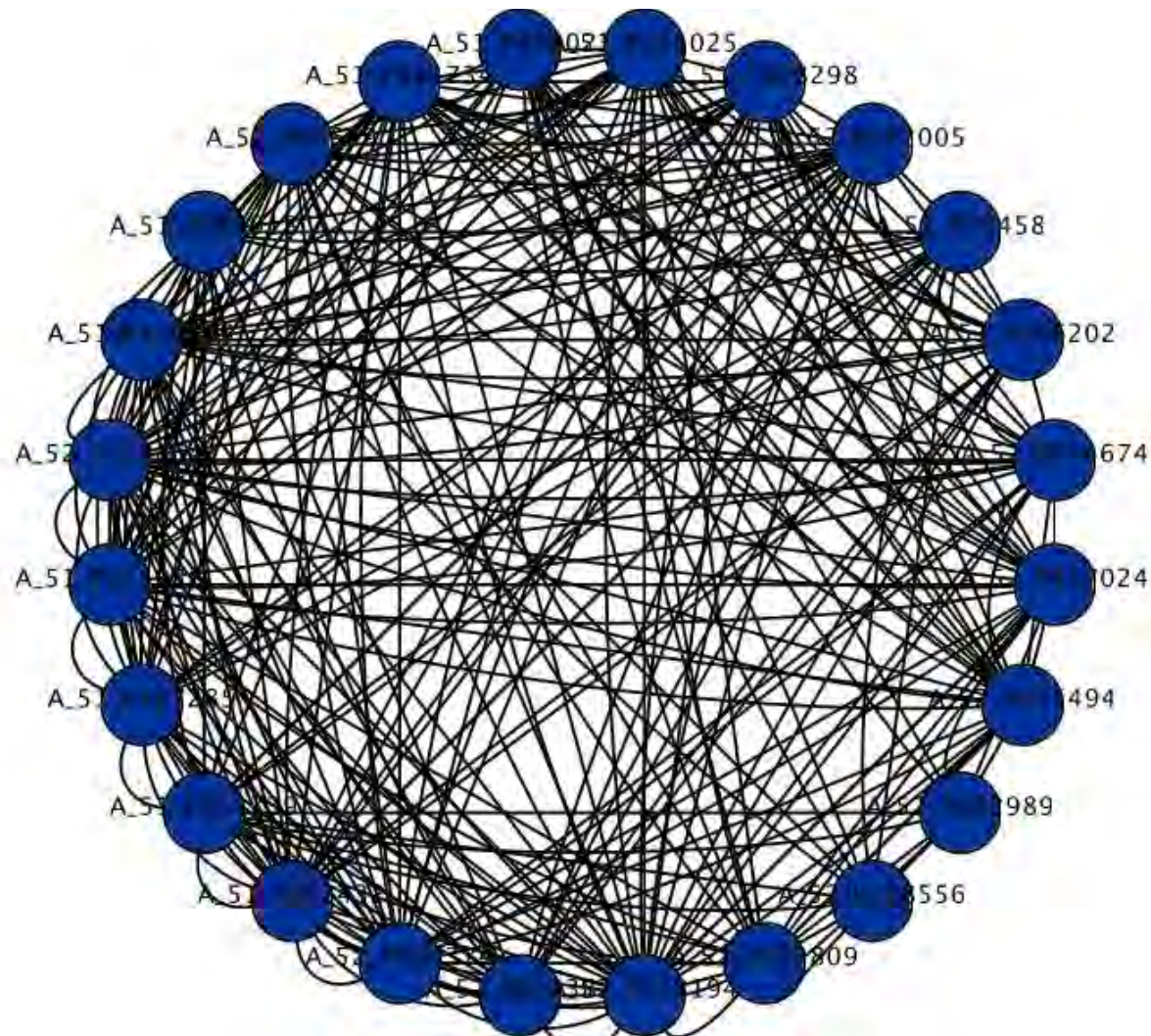


# Correlation Networks



# Correlation Networks

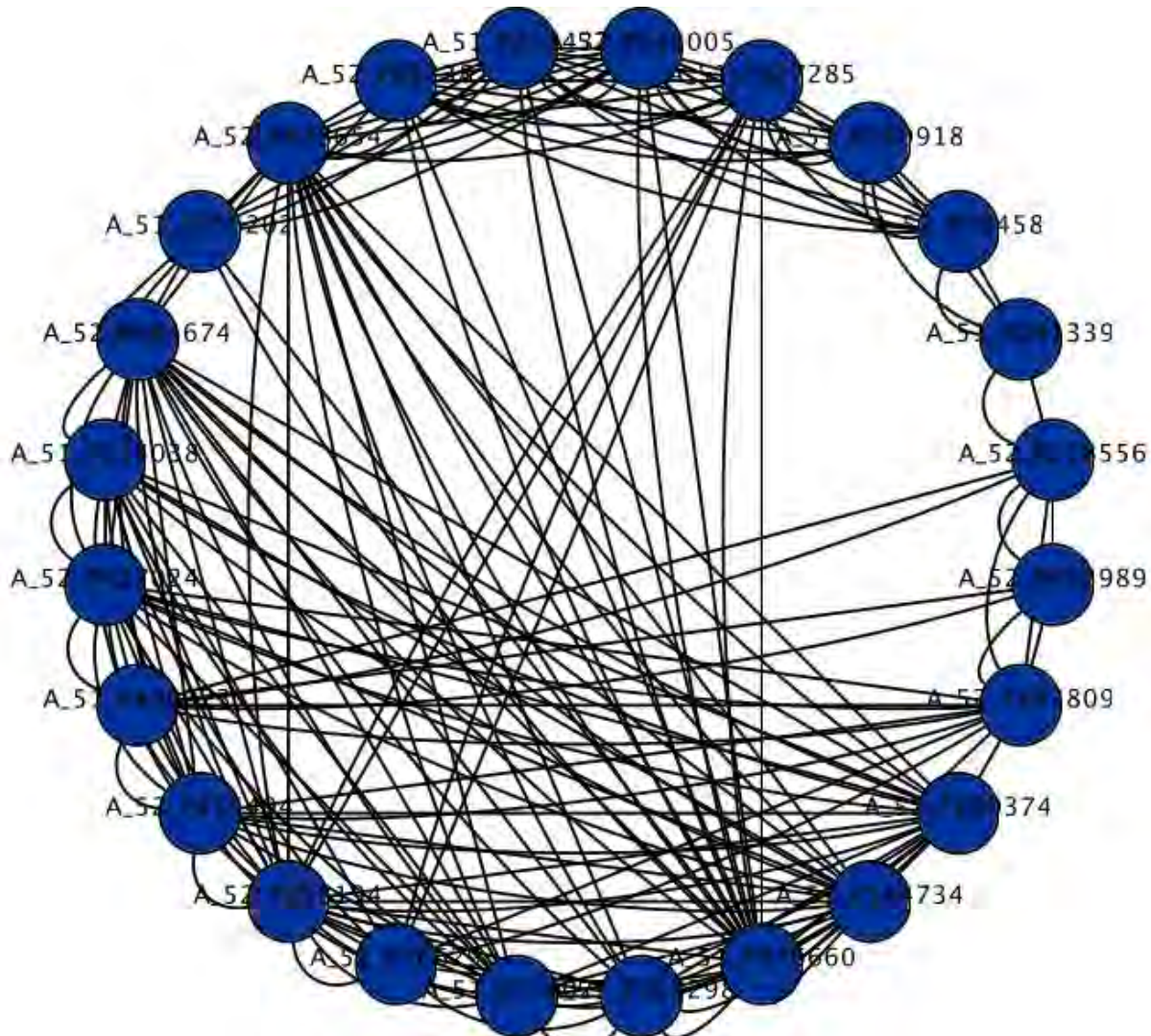
24 node sample  
Threshold: 0.00-1.00





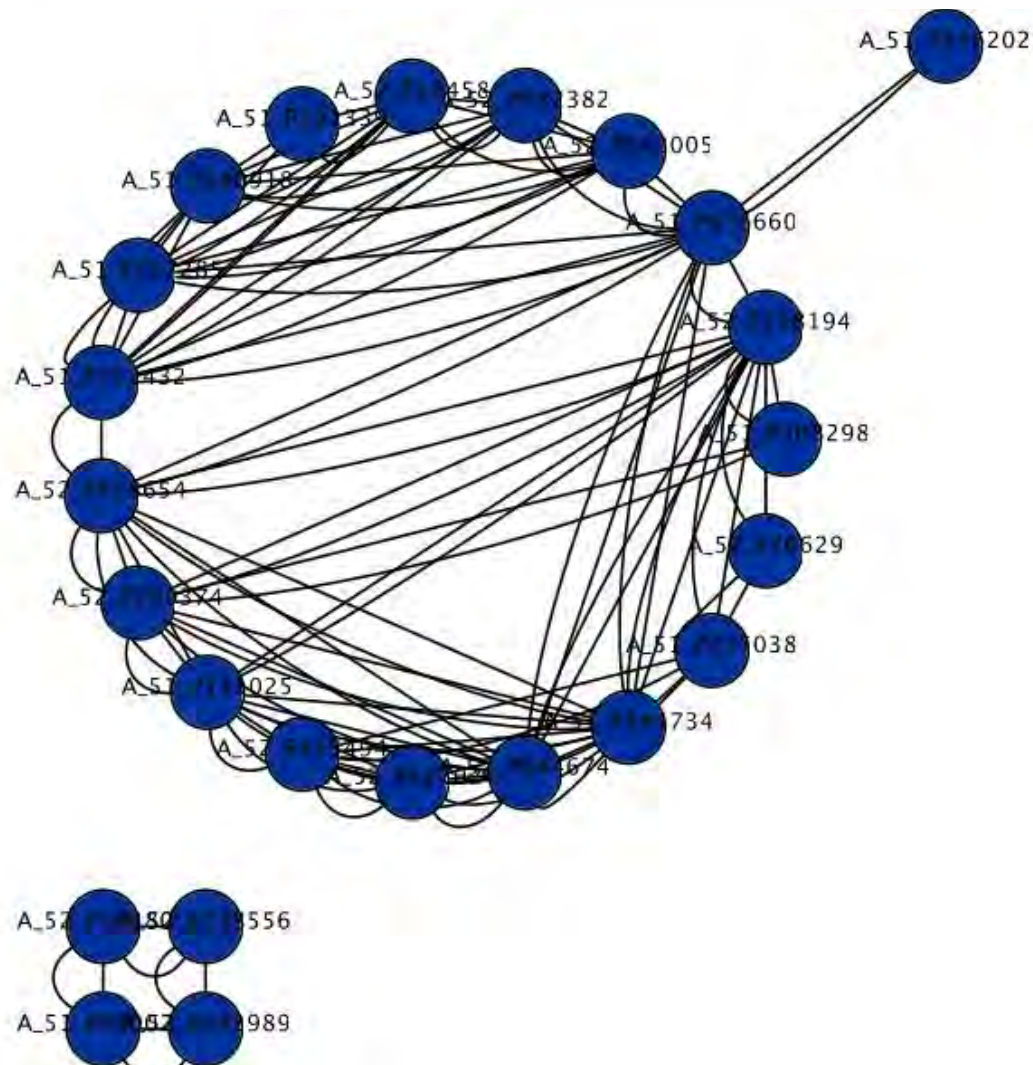
# Correlation Networks

24 node sample  
Threshold: 0.30-1.00



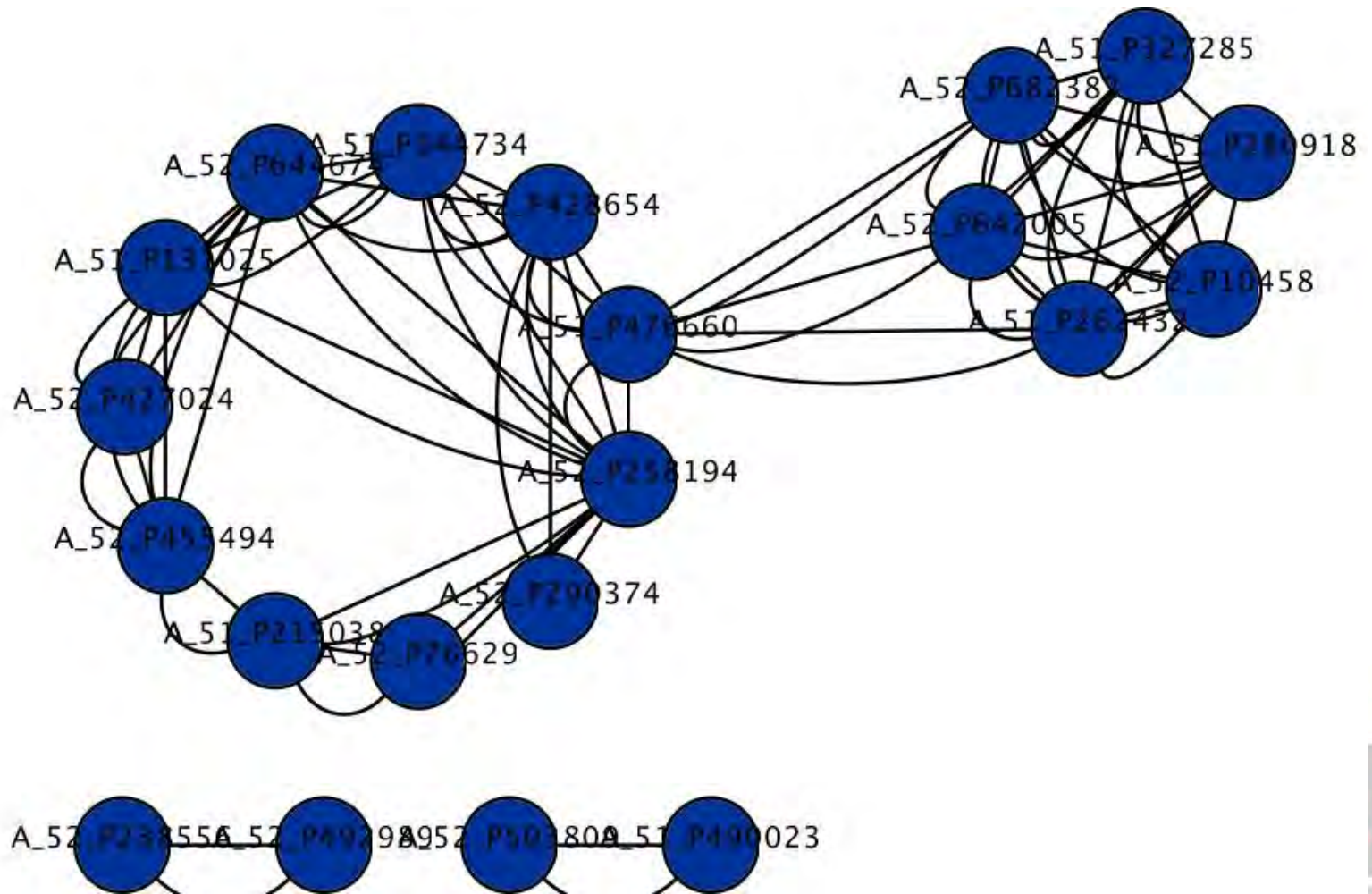
# Correlation Networks

24 node sample  
Threshold: 0.50-1.00



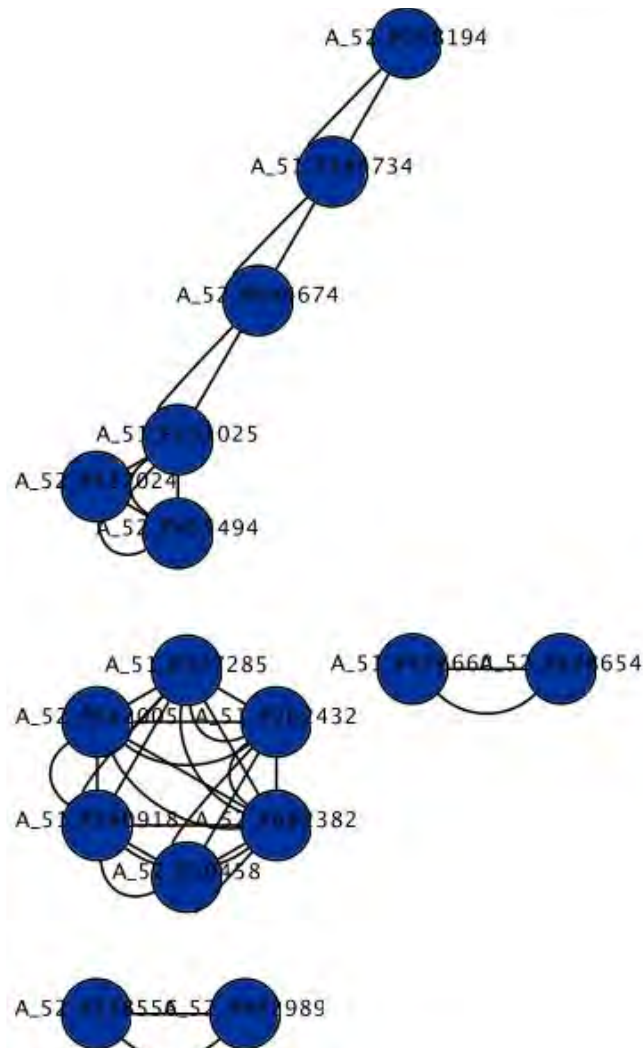
# Correlation Networks

24 node sample  
Threshold: 0.60-1.00



# Correlation Networks

24 node sample  
Threshold: 0.80-1.00



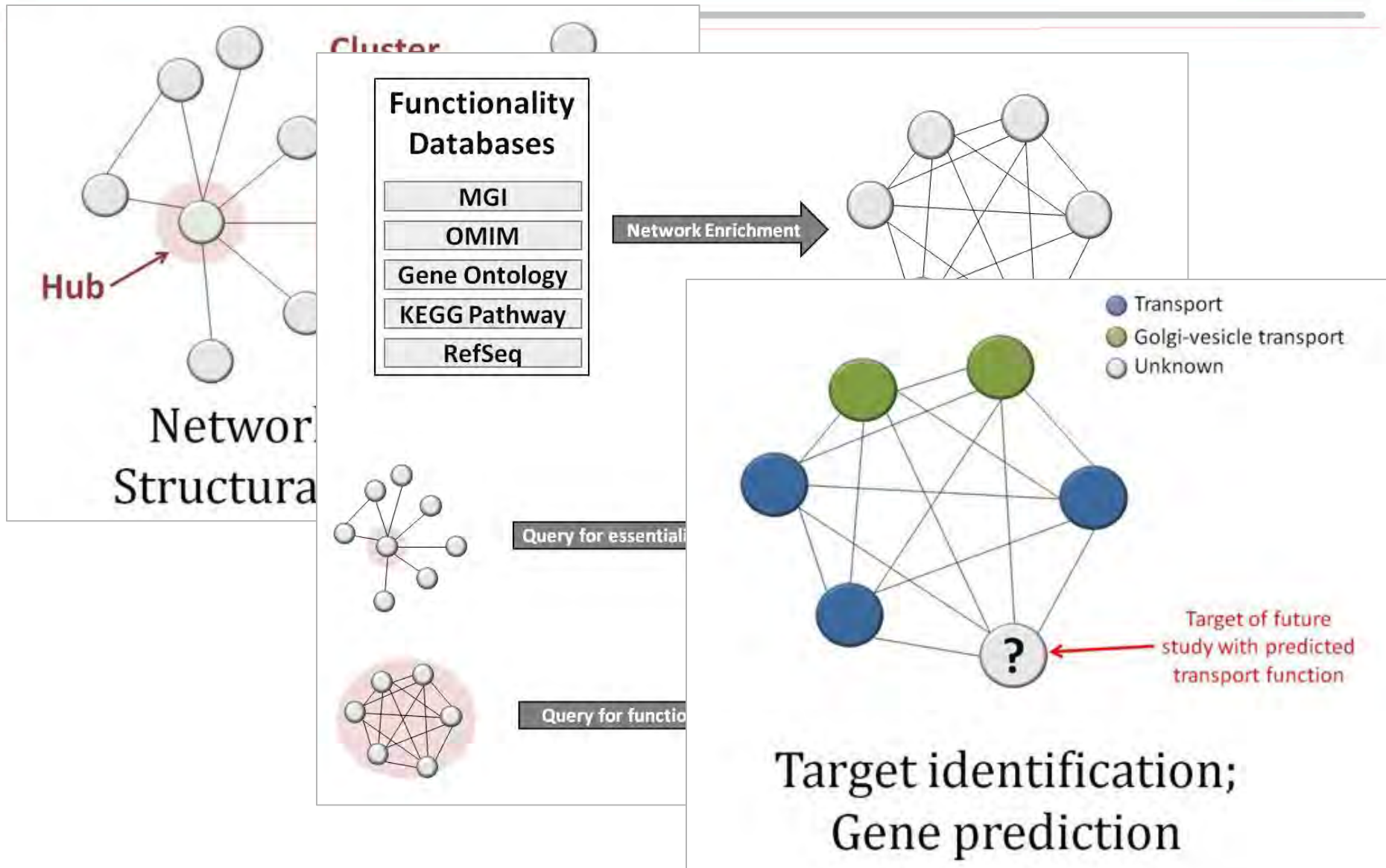
# Correlation Network Applications

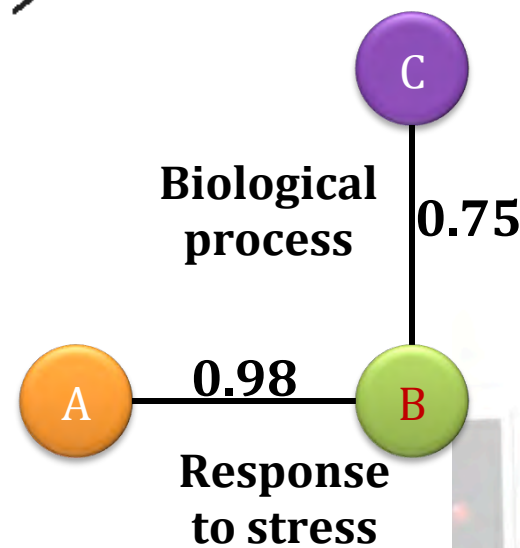
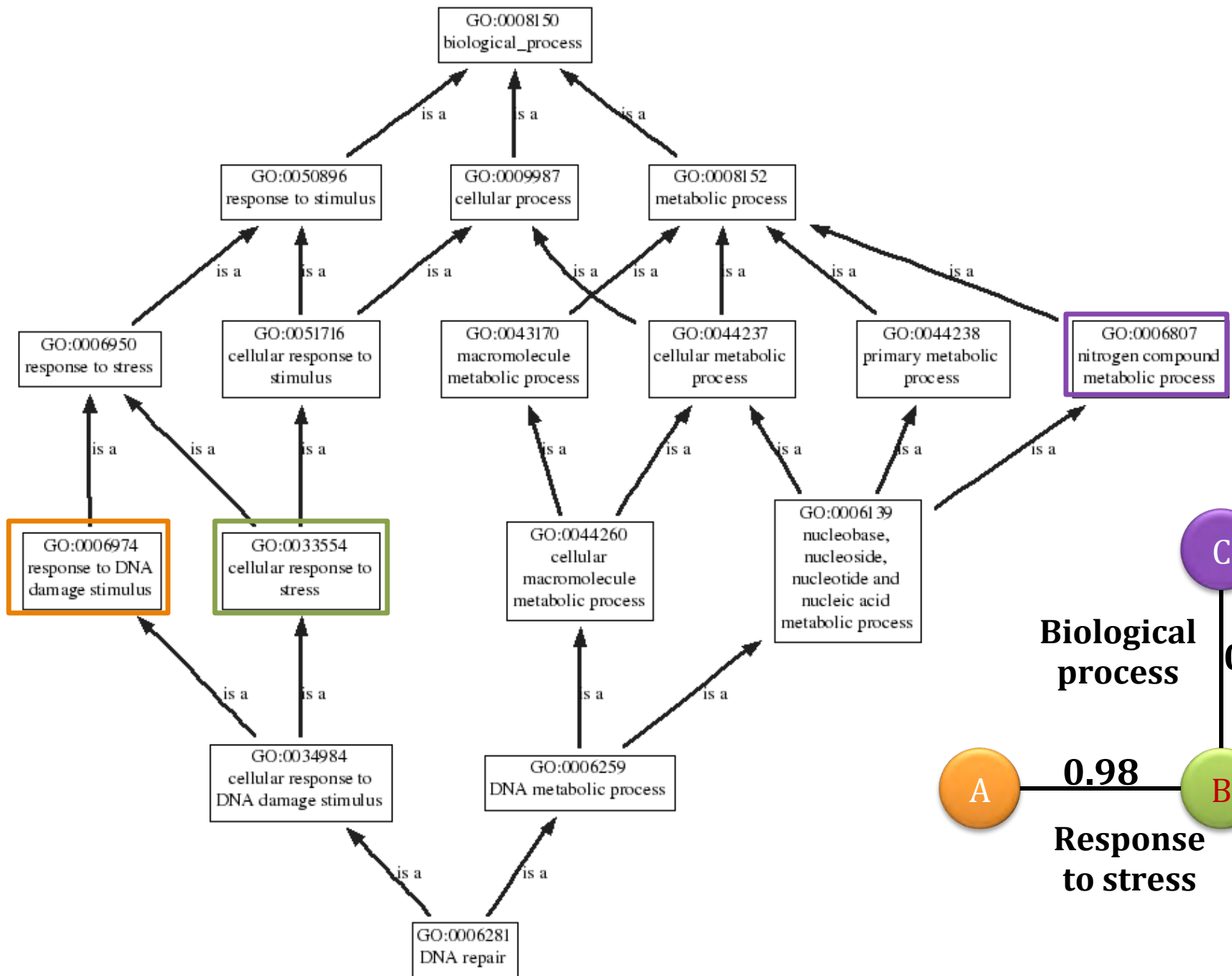
---

- “Versus” analysis
  - Normal vs. disease
  - Times/environments
- Model for high-throughput data
  - Especially useful in microarrays
- Identification of groups of causative genes
  - Ability to rank based on graph structure
  - Identify sets of co-regulated, co-expressed genes



# Integrated Data Model





# Case Study in Aging

---

- With aging, certain behaviors decrease
  - Eating, drinking, activity levels
- Observed gene expression changes in the hypothalamus
  - Can we capture these expression changes?
  - Can we correlate these changes to behavioral decreases?
- Goal: Identify temporal biological relationships
  - Progression of disease
  - Effect of pharmaceuticals on systems of the body
  - Aging





# Case Study in Aging

- 5 sets of temporal gene expression data

Strain	Gender	Tissue Type	Ages
BalbC	Male	Hypothalamus	Young, mid-age, aged
CBA	Male	Hypothalamus	Young, mid-age, aged
C57_J20	Male	Hypothalamus	Young, aged
BalbC	Female	Hypothalamus	Young, aged
BalbC	Female	Frontal cortex	Young, aged

# Hubs and Drivers

---

- **Hub:** a high-degree node in a network
- Node degree in filtered correlation networks follows power-law relationship
- Few nodes with high degree

Albert et al 2005

- High degree nodes → highly essential

Bergmann et al 2004

Carlson et al 2006



# Hub Lethality

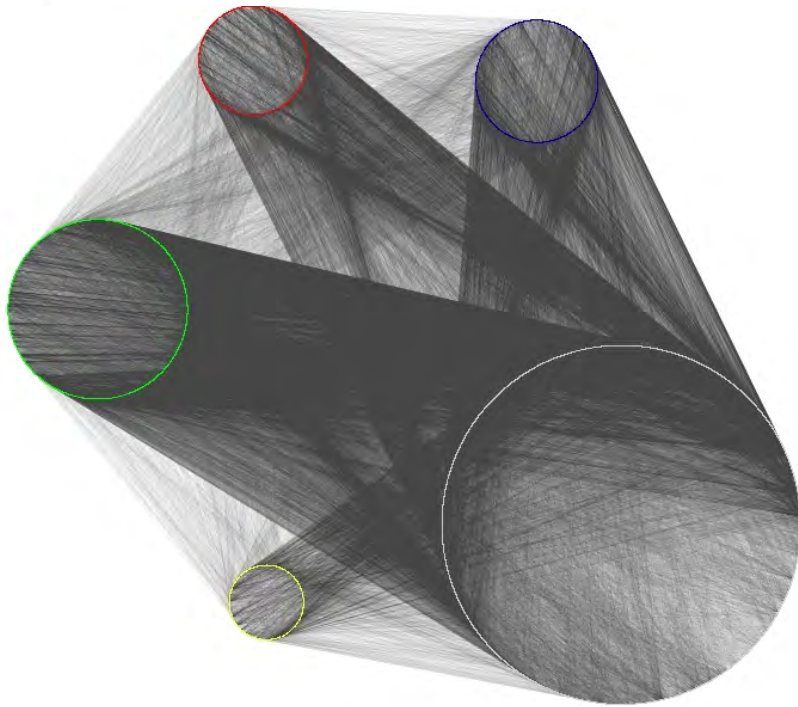
---

- Young Male BalbC Mouse
  - 12/20 hubs tested for *in vivo* knockout
    - 8/12 lethal phenotype
    - 4/12 non-lethal but system-affecting
    - 0/12 no observed phenotype
  
- Aged Male BalbC Mouse
  - 11/20 hubs tested for *in vivo* knockout
    - 7/11 lethal phenotype pre-/peri-natally
    - 3/11 non-lethal but system-affecting
    - 1/11 no observed phenotype (Aldh3a1)

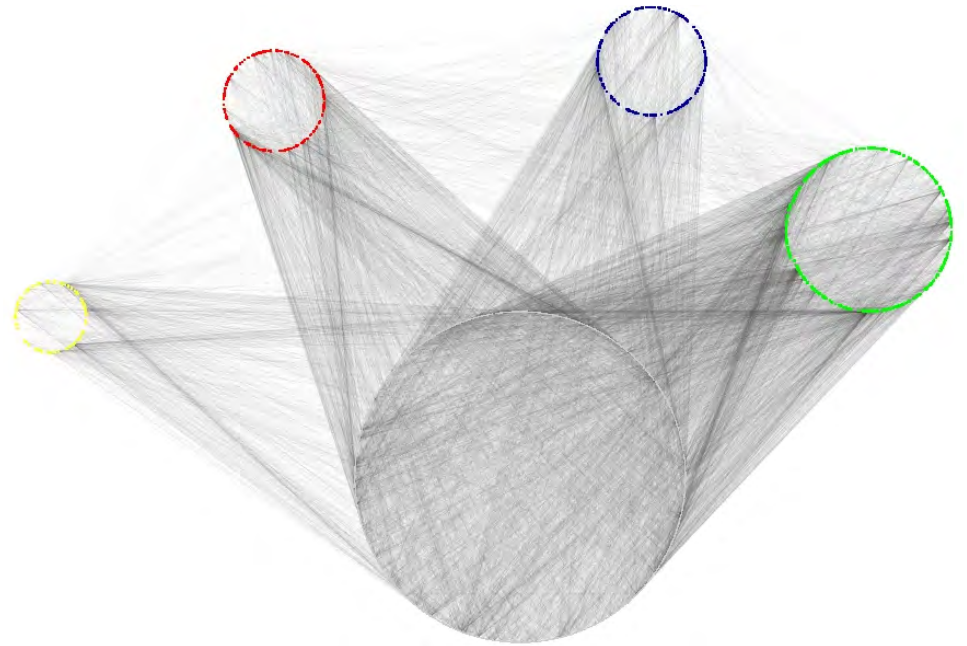


# Aging and Biological Networks

---

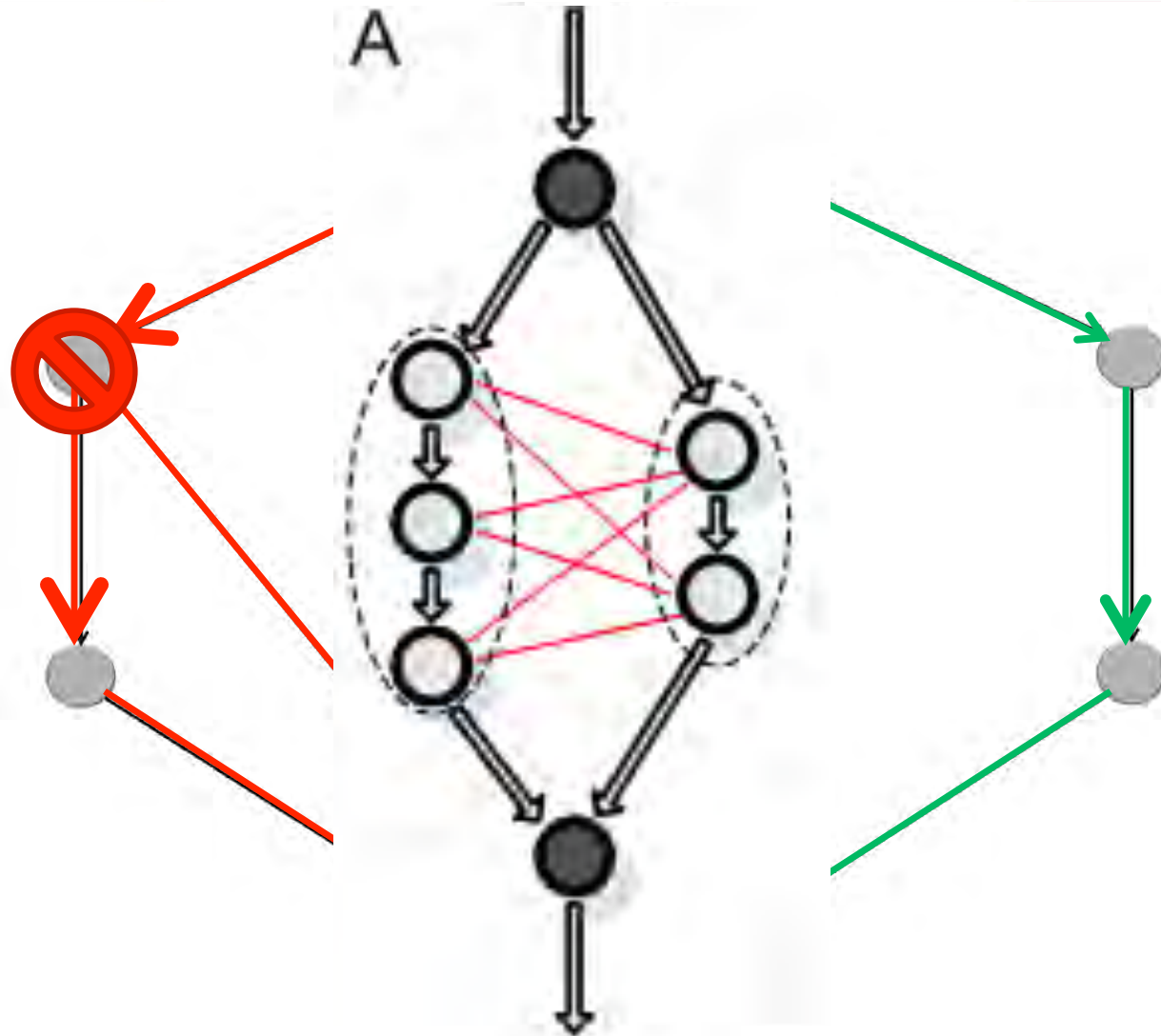


[young]



[aged]

# Compensatory Pathways



# Structures & their Functions

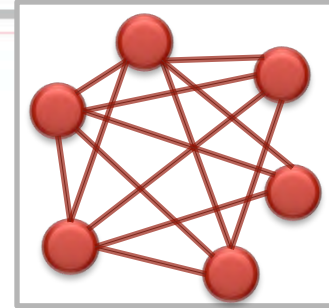
Key Hypothesis:

Network structures correspond to key cellular structures

# Local Network Structures

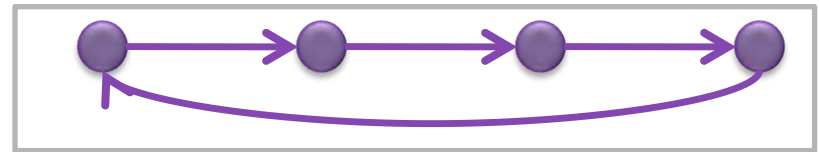
- **Cliques**

Protein complexes, regulatory modules



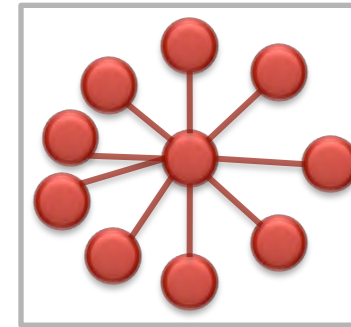
- **Pathways**

Signaling cascades

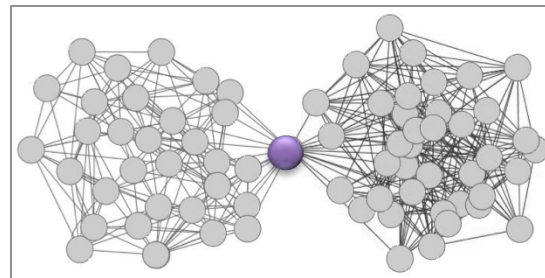


- **Hubs**

Regulators, TFs, active proteins



- **Articulation points**



# HPC and Big Data in Biological Networks

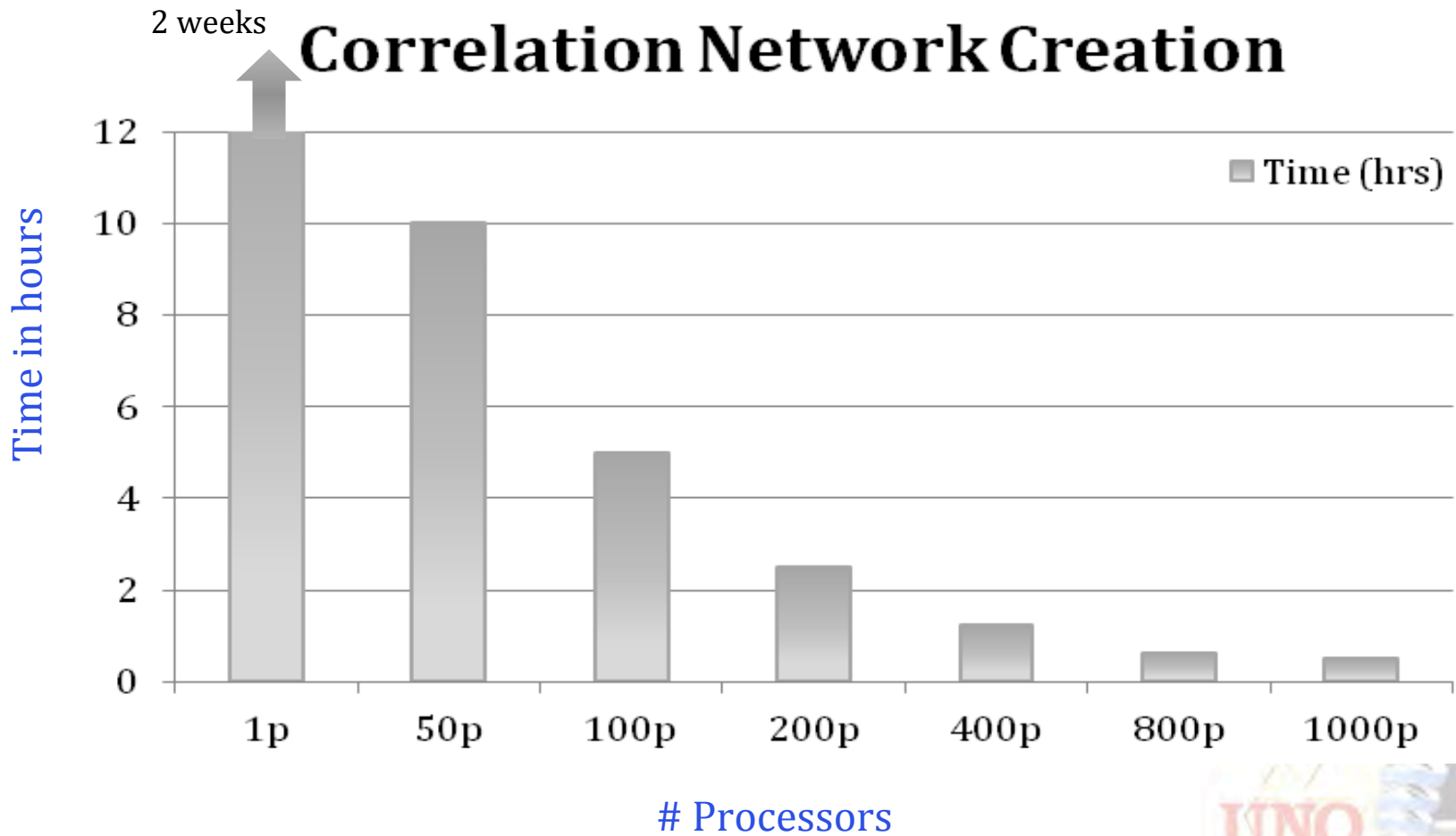
---

- Network creation: 2 weeks on PC
  - 10 hours in parallel, 50 nodes
  - 40,000 nodes = 800 million edges (pairwise)
  - 40,000 ! Potential relationships
  - Big data or big relationship domain
- Network analysis: Best in parallel
  - Only 3% of entire genome forms complexes
- Holland Computing Center: Firefly 1150 8-core cluster – from weeks to hours/minutes

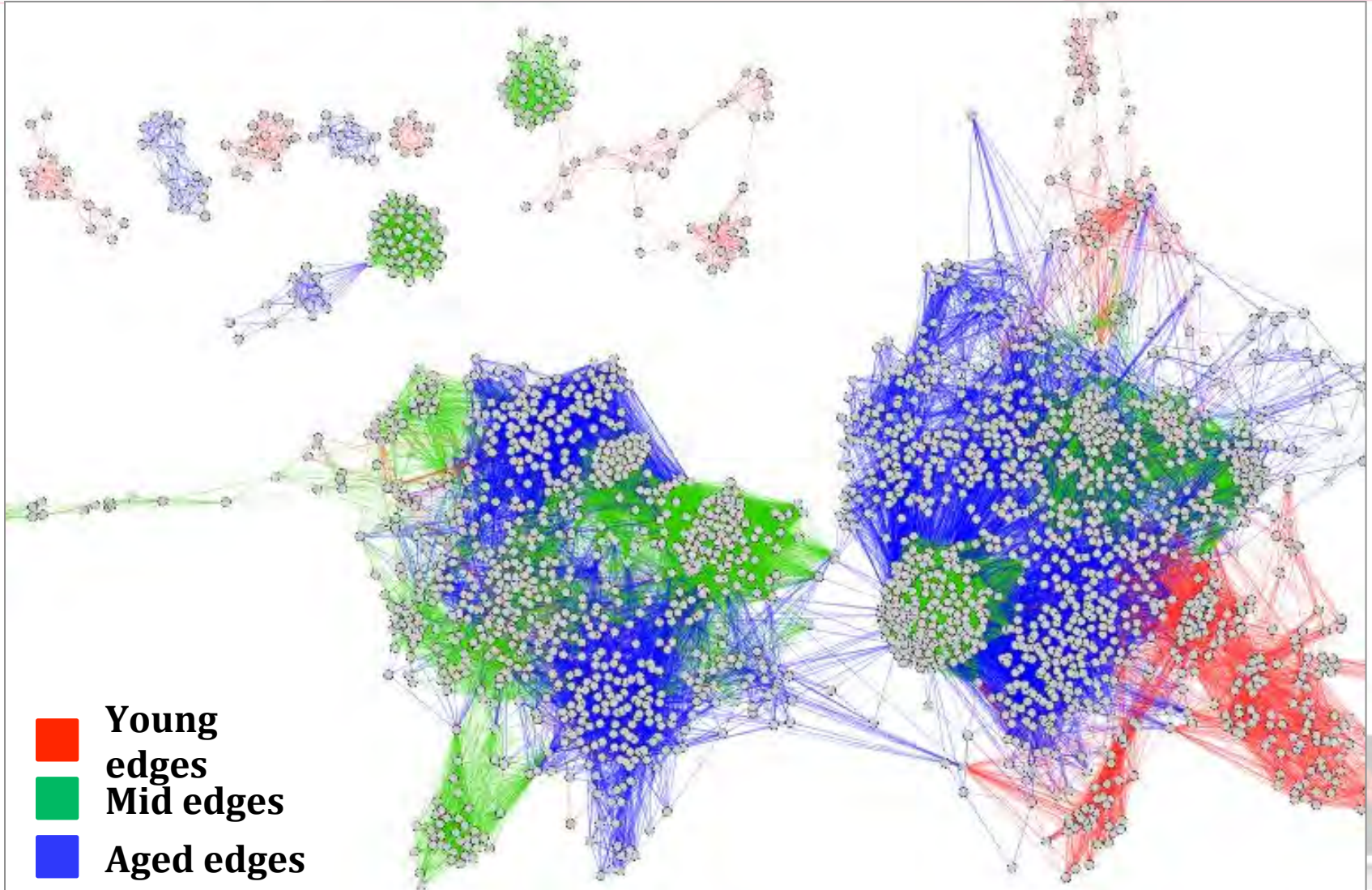





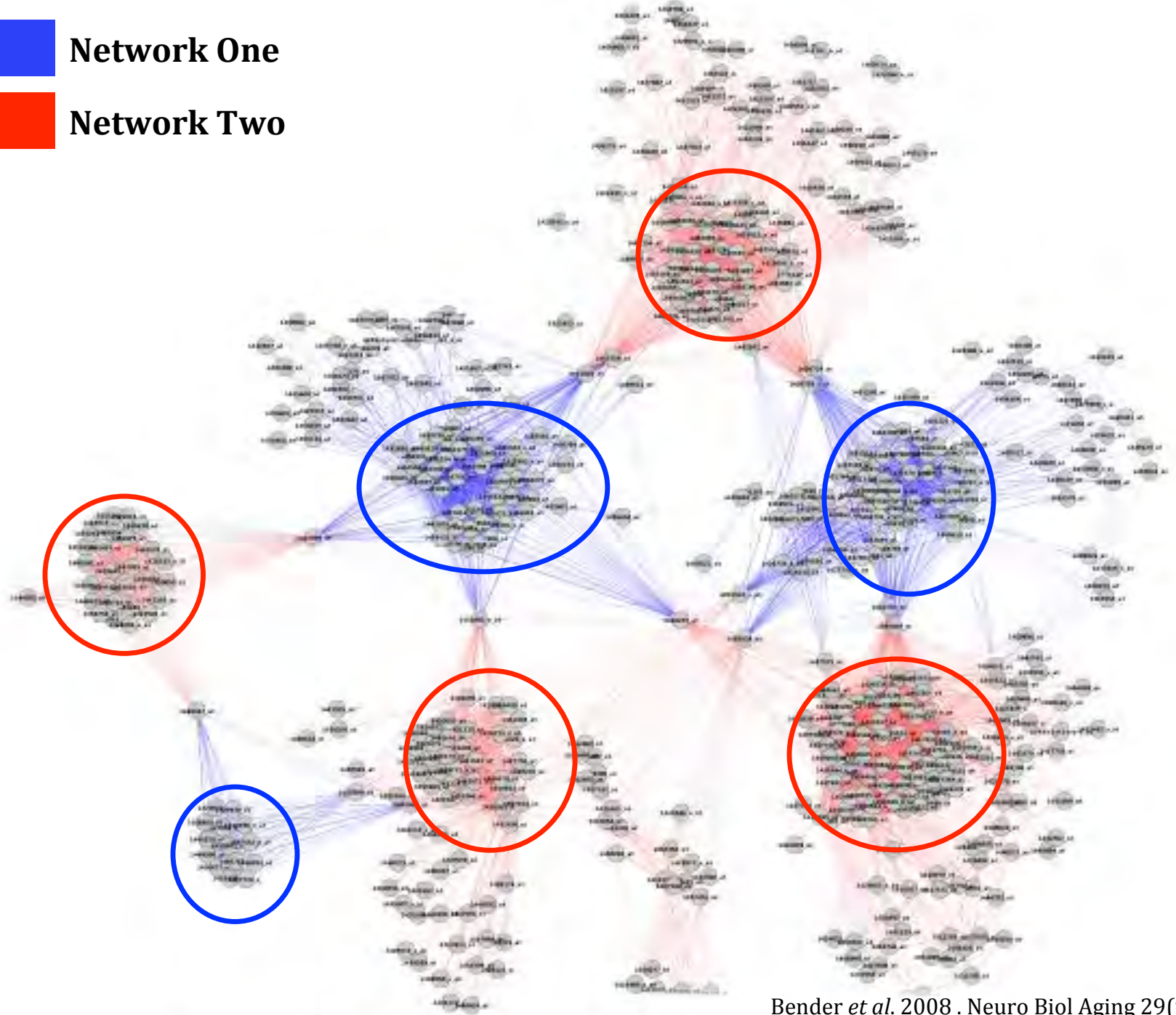
# The Need for HPC



# Aging-Related Networks



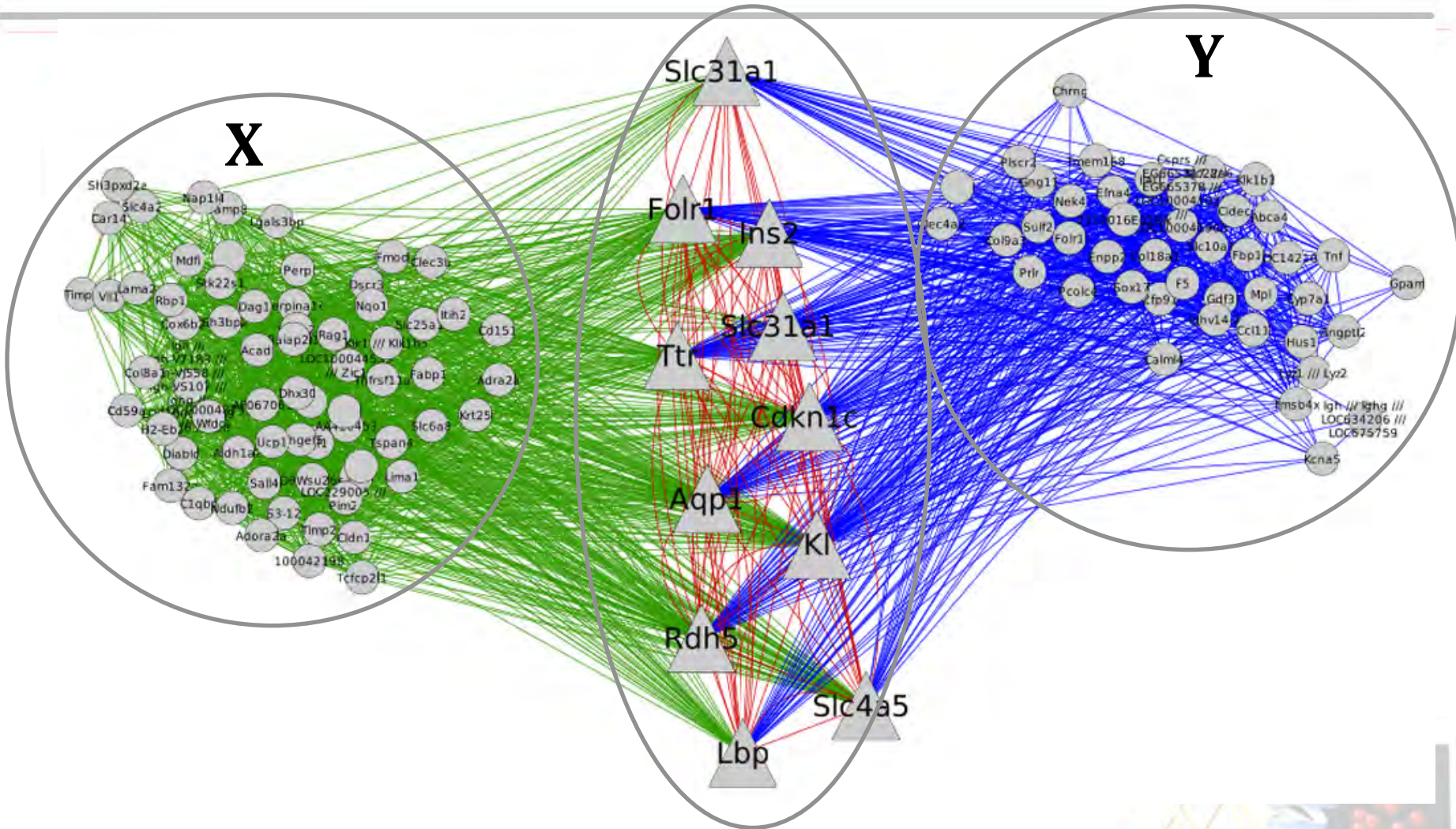
 **Network One**  
 **Network Two**



# S

# Y

# X



# Node Gatewayness

---

- Let undirected graphs  $G1 = (V, E1)$  and  $G2 = (V, E2)$  such that graphs  $G1$  and  $G2$  share same node set  $V$  with different edge sets  $E1$  and  $E2$ .
- For each graph we identify clusters (dense subgraphs ) such that:
  - Cluster  $X$  represents some dense subgraph in  $G1$
  - Cluster  $Y$  represents some dense sub-graph in  $G2$
- Compute  $G'$  such that  $G' = (V, (E1 \cup E2))$



# Node Gatewayness

- Define subset of nodes  $S = V(X) \cap V(Y)$
- For any node  $s$  in  $S$ ,  $E_{(s)}$  is the set of edges connecting  $s$  to any node in the set  $X$  from graph  $G1$  and the set of edges connecting  $s$  to any node in the set  $Y$  from graph  $G2$ .
- Using these definitions we define gatewayness as the following:

$$\text{gatewayness}_s = \frac{E_{(s)}}{(E1(X) + (E2(Y)))}$$

- $E(s)$  = Total edges connecting  $s$  to  $X$  and  $Y$
- $E1(X)|E2(y)$  = Total edges connecting  $S$  to  $X$  and  $Y$



# Incorporating Graph Theoretic Concepts

---

## Elements (Nodes):

Betweenness

Closeness

Degree

BC: Highest betweenness + closeness

CD: Highest degree + closeness

BD: Highest betweenness + degree

BCD: Betweenness + closeness + degree

## Subsystems (Relationships, groups) :

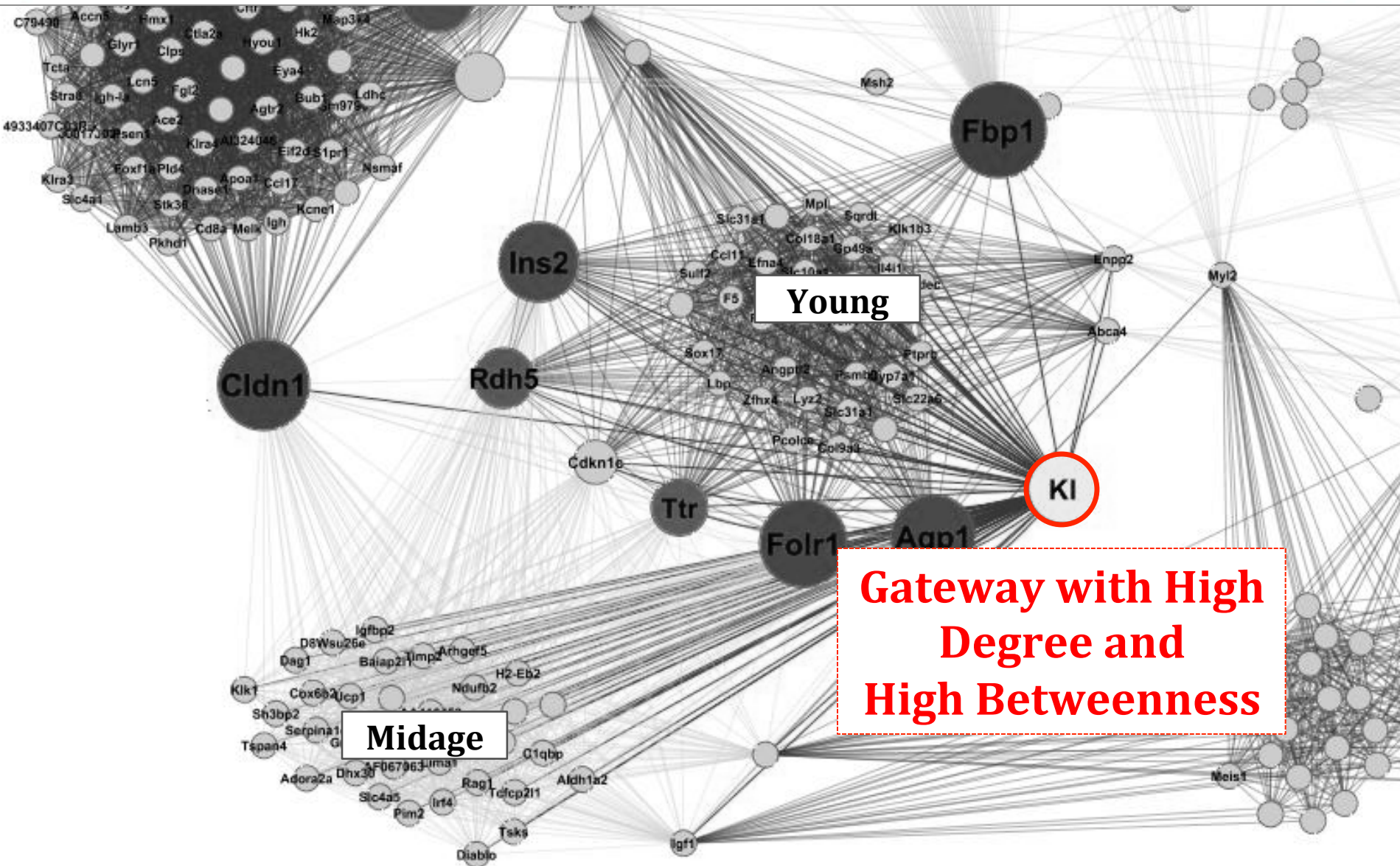
Clusters, cliques

Pathways

Loops/cycles

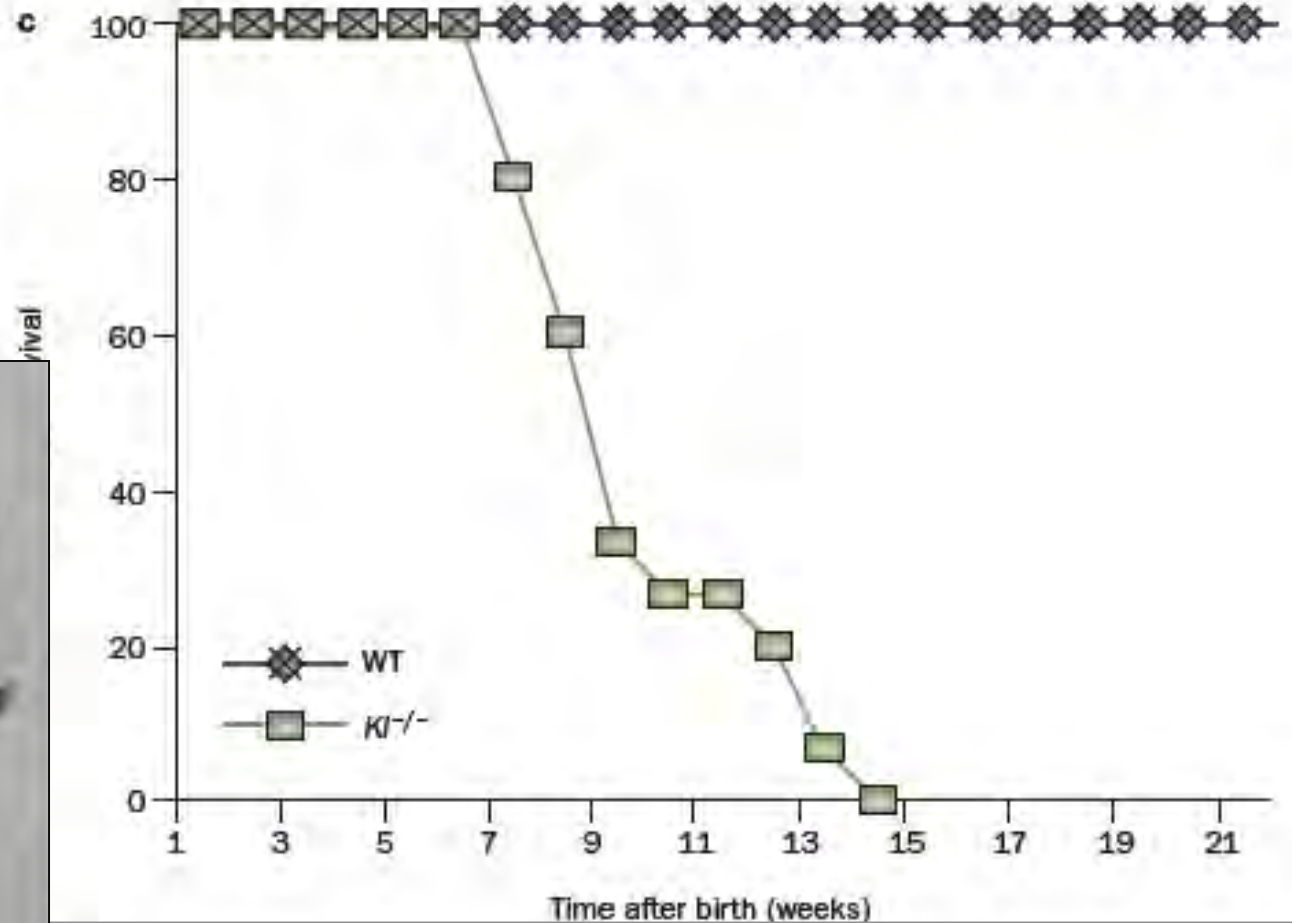


# High BD Node: Klotho

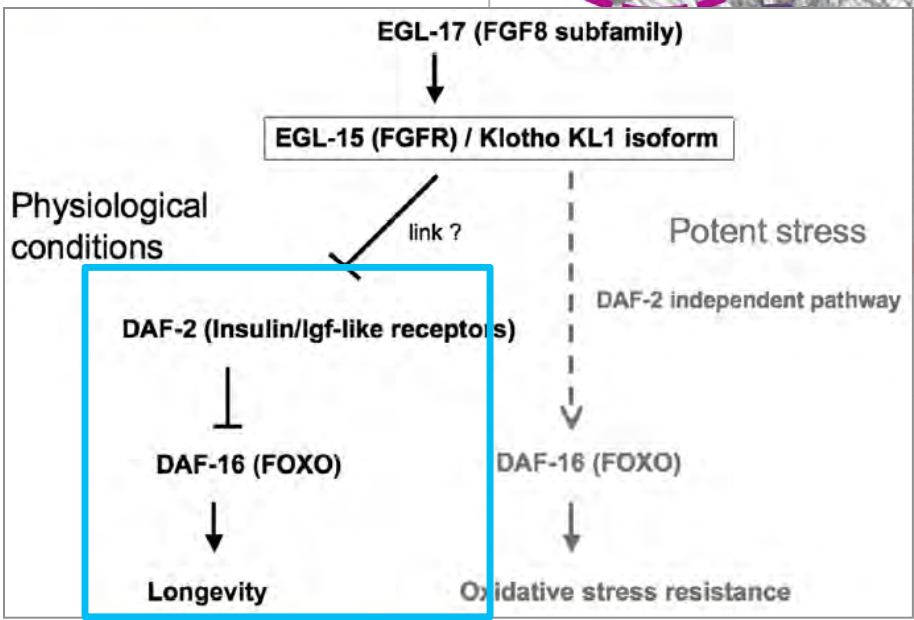
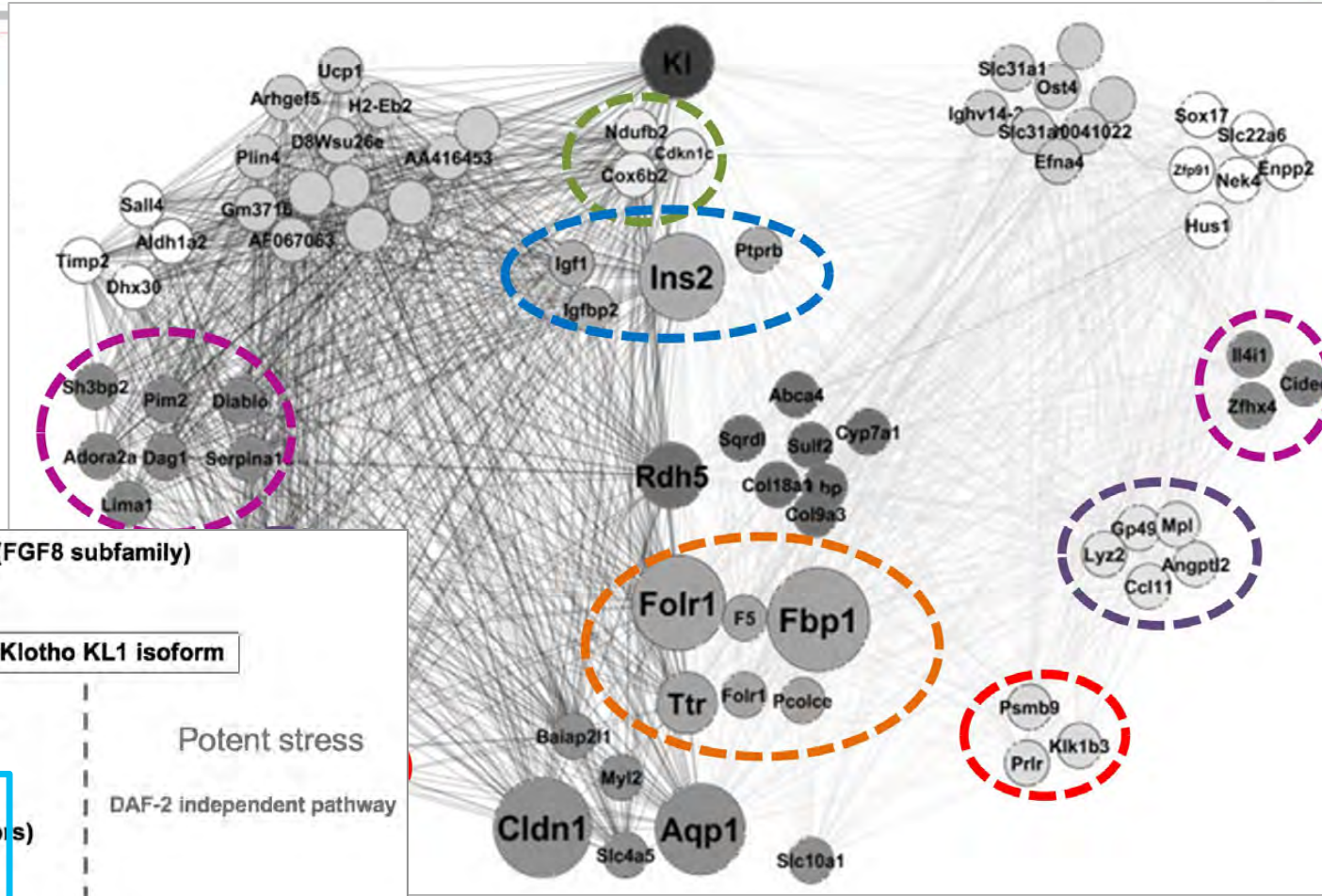




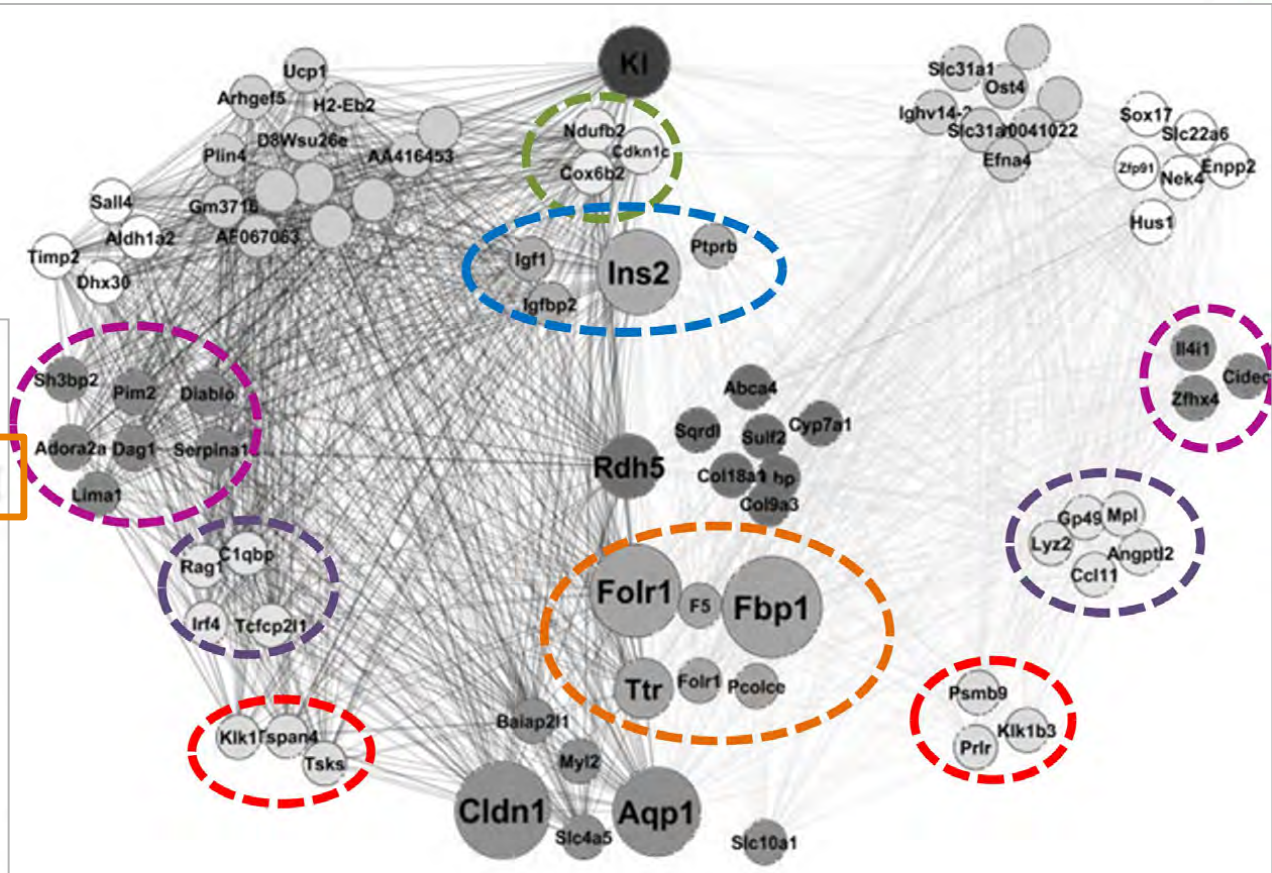
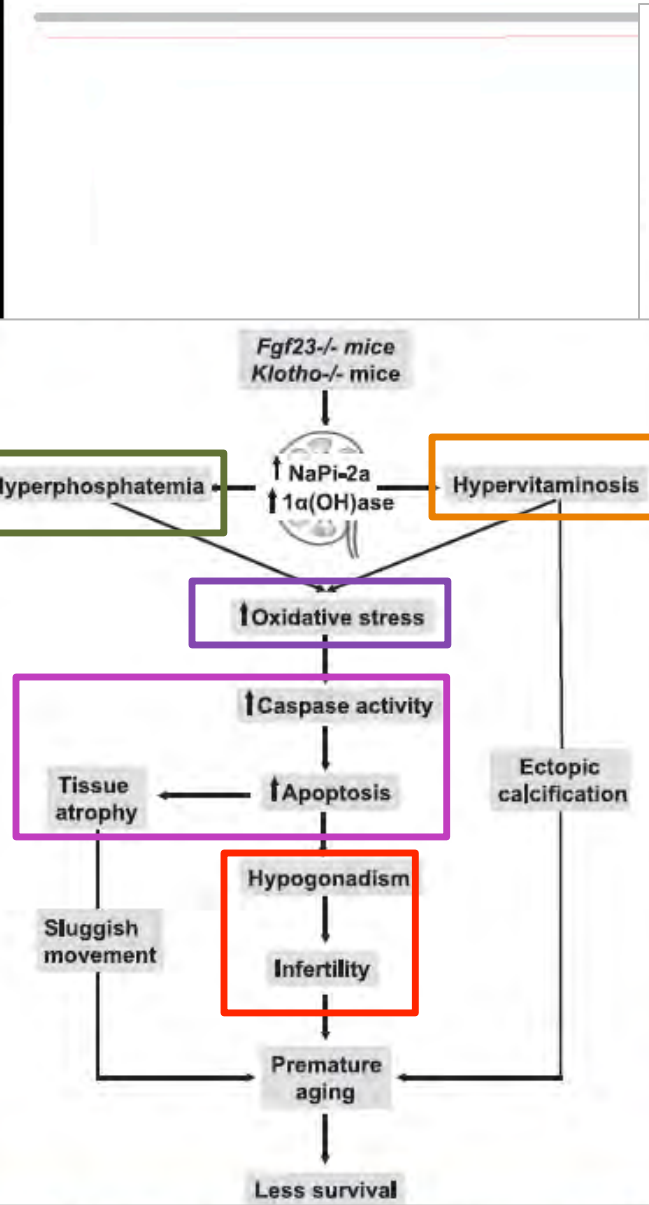
# High BD Node: Validation



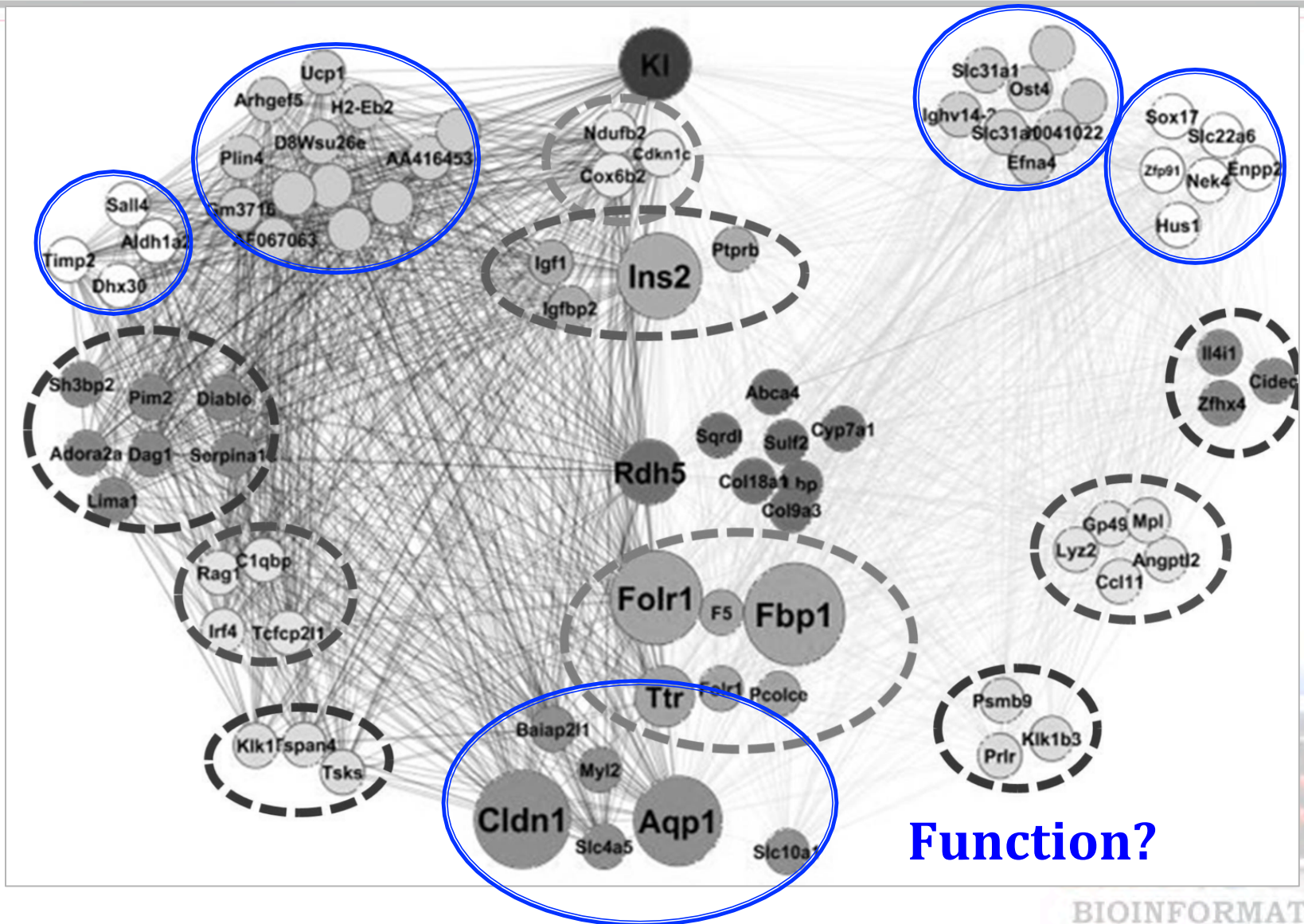
# Validation



# Subsystems Validation



# Discovery



# Case Study: HIV and Drug Addiction

---

Infected	Not Infected
Infected + Combinatorial Drugs	Not Infected + Combinatorial Drugs
Infected + Meth	Not Infected + Meth
Infected + Meth + Combinatorial Drugs	Not Infected + Meth + Combinatorial Drugs



# Role of Methamphetamine

---

- Methamphetamine is a major drug of abuse with reported high use by HIV-infected groups
- Methamphetamine users have higher risk of getting HIV infection
- Impact on nervous system is higher when Methamphetamine is used by HIV infected individual (neuronal injury)

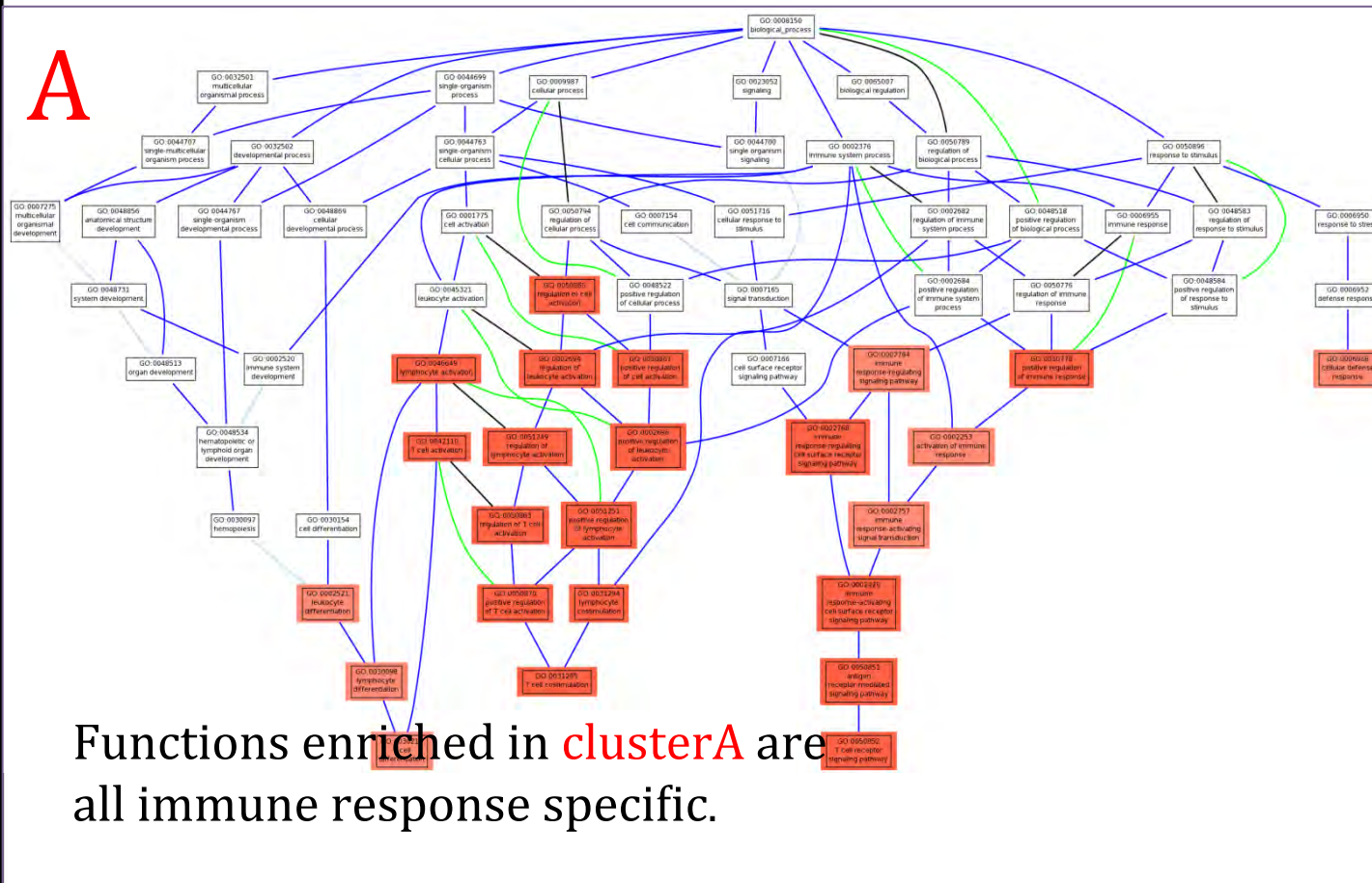


# Clusters are enriched in specific functions

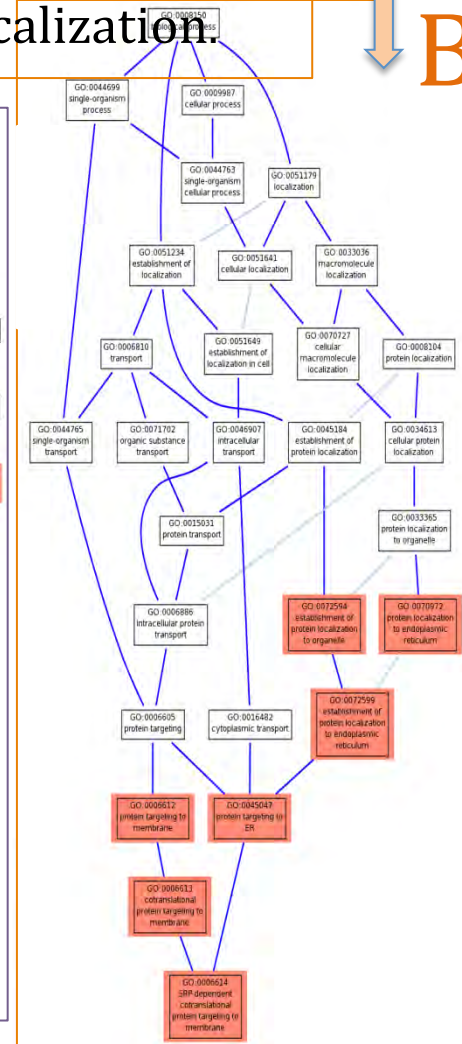
Functions enriched in **clusterB** are protein targeting and localization.



A

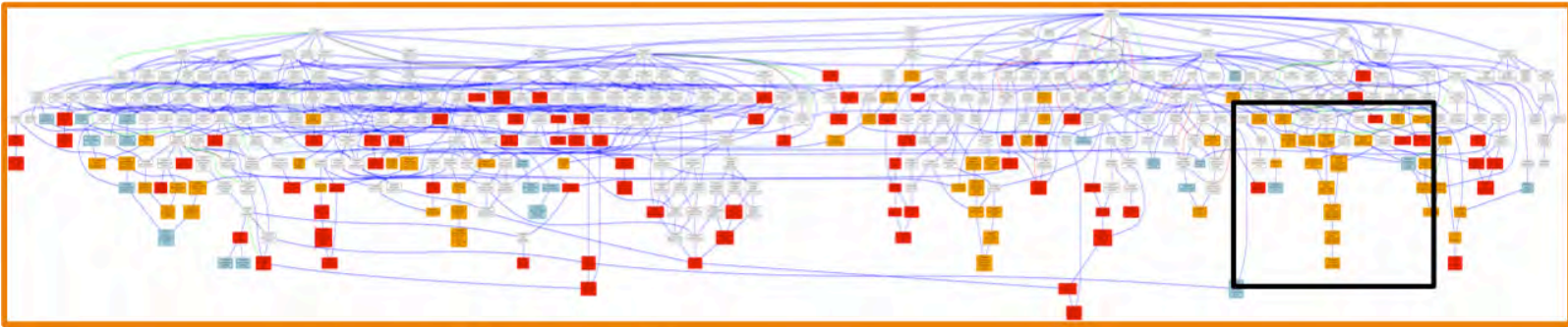


Functions enriched in **clusterA** are all immune response specific.

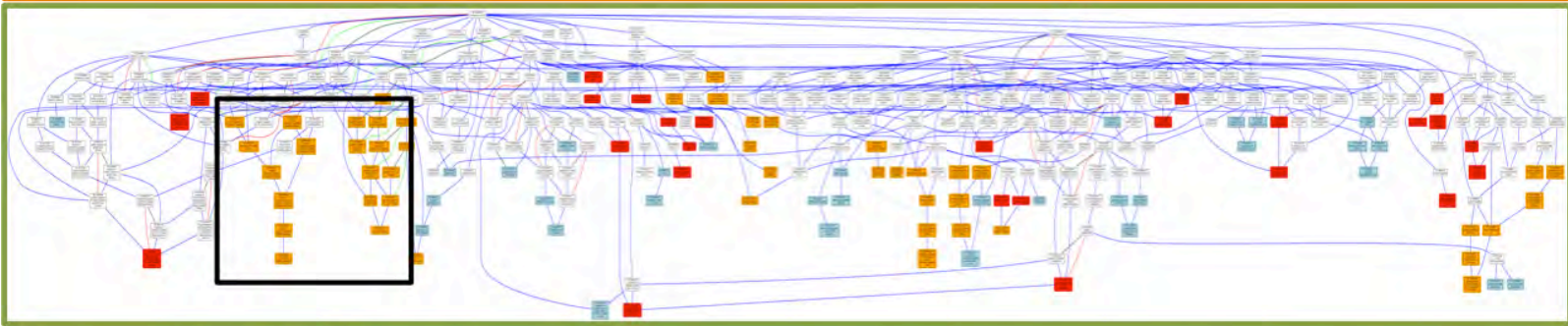


GO:0006124  
SRP-dependent cotranslational protein targeting to membranes

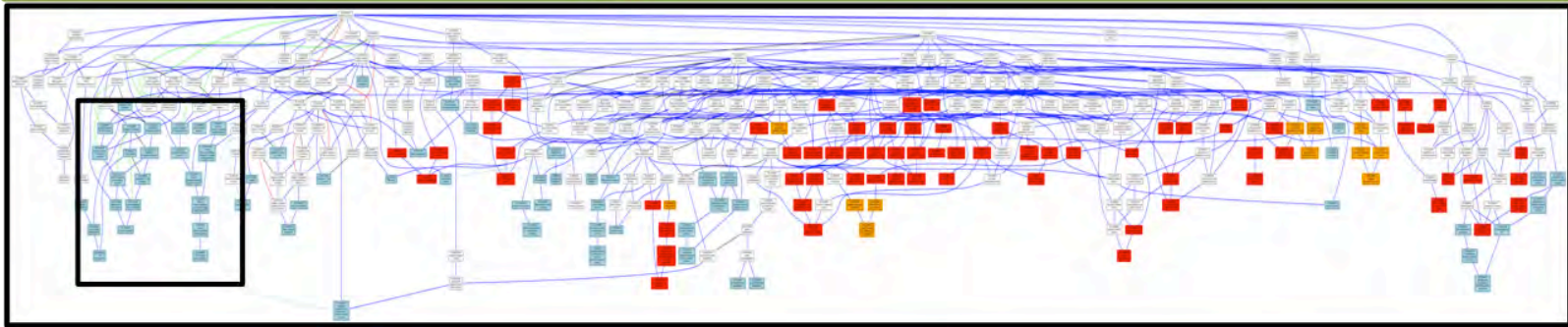
Orange nodes = enriched in both sets; Blue nodes = enriched only in Uninfected



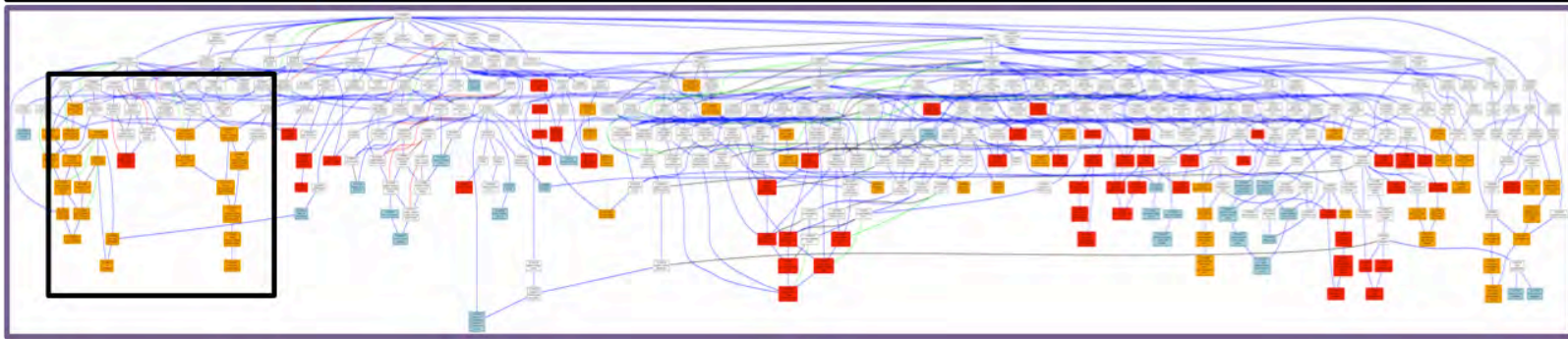
Infected vs.  
Uninfected



HIV  
treatment  
vs.  
Uninfected



Infected +  
Meth vs.  
Uninfected



Infected +  
Meth +  
Treatment  
vs.  
Uninfected



# Results

---

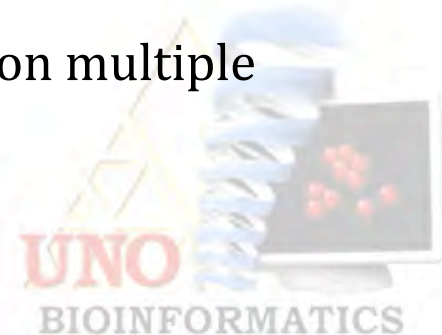
- Large number of nodes are enriched in only one network in Infected + Meth network.
  - Many functions enriched in other conditions have been dropped out in Infected + Meth network.
- Most of the lost functions reappear in Infected + Treated
- Some of these lost functions reappear in Infected + Meth + Treatment

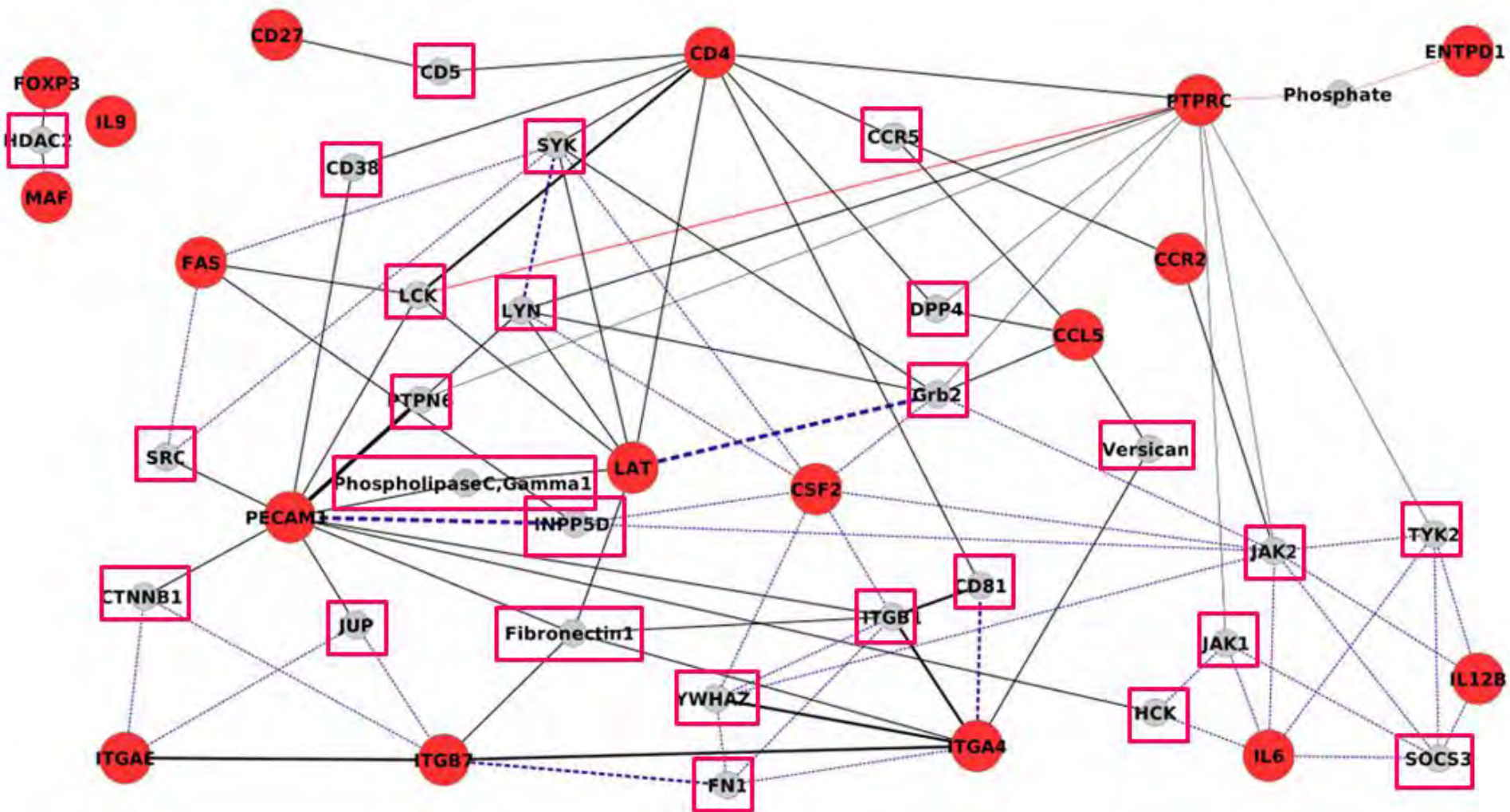


# Case Study: Parkinson's Disease

---

- Data: Flow cytometry markers
  - Parkinsons Disease patients
  - Caretakers (non-Parkinsons)
- Method:
  - Create immediate neighbor (1-hop) interactome
  - Identify targets/interactors
  - Identify “key players” based on iterative marker identification
- Outcome:
  - Identification of new marker targets
  - Notable: Identification of 3 major targets based on multiple evidences (from network integration)





Red nodes:	Original markers
Pink boxes:	Marker targets based on connectivity



<b>1-Hop PPI Targets</b>	<b>1-Hop PPI Connected Targets</b>	<b>Pathway Targets</b>	<b>Reverse 1-Hop PPI Targets</b>	<b>Reverse 1-Hop PPI Targets -</b>	<b>Additional Marker Targets</b>
ITGB1	ITGB1	ITGB1	ITGB1	ITGB1	ITGB1
INPP5D	INPP5D		INPP5D	INPP5D	INPP5D
LCK	LCK		LCK	LCK	LCK
PIK3R1	PIK3R1	PIK3R1	PIK3R1	PIK3R1	
SYK	SYK		SYK	SYK	SYK
CD53	CD53		CD53	CD53	
EED	EED		EED	EED	
FYN	FYN		FYN	FYN	
HCK			HCK	HCK	HCK
JUP			JUP	JUP	JUP

Column 1:	Initial IM network targets
Column 2:	Post-processing IM network targets
Column 3:	Initial Pathway network targets
Column 4:	Reverse - Iterative IM targets - Run 1
Column 5:	Reverse - Iterative IM targets - Run 2
Column 6:	Targets from extraneous data



# How to implement this stuff?

## Computer Science Issues

---

- High Performance Computing
  - Beyond surface-level adaptation of previous algorithms
- Security and Privacy
  - Cloud Security
- Wireless Networks
- Graph Algorithms



# Conclusions

---

- Many Scientific disciplines are now at crossroads
- The proper penetration of IT represent tremendous challenges and great opportunities
- Availability of public data ensures that discoveries are likely to take place at many places
- Interdisciplinary approach is a key for addressing many critical problems – which may lead to major scientific advancement



# Acknowledgments

- UNO Bioinformatics Research Group
  - Kiran Bastola
  - Sanjukta Bhoomwick
  - Kate Dempsey
  - Jasjit Kaur
  - Ramez Mena
  - Sachin Pawaskar
  - Oliver Bonham-Carter
  - Ishwor Thapa
  - Dhawal Verma
  - Julia Warnke
- Former Members of the Group
  - Alexander Churbanov
  - Xutao Deng
  - Huiming Geng
  - Xiaolu Huang
  - Daniel Quest
- Biomedical Researchers
  - Steve Bonasera
  - Richard Hallworth
  - Steve Hinrichs
  - Howard Fox
  - Howard Gendelman
- Funding Sources
  - NIH INBRE
  - NIH NIA
  - NSF EPSCoR
  - NSF STEP
  - Nebraska Research Initiative