University of Nebraska at Omaha

# High Performance Computing in Biomedical Informatics

Hesham H. Ali

UNO Bioinformatics Core Facility

College of Information Science and Technology

# Tutorial Outlines

- Biomedical Informatics - Challenges and Opportunities
- Next Generation Bioinformatics Tools – Focus on Data Analysis and Integration
- Systems Biology and Network Analysis
- Biomedical Informatics and the Cloud: Models and Security
- Case Studies of Discoveries using Biological Networks
- HPC in Network Analysis of High Throughput Biological Data
- Integration of different aspects of Biomedical Informatics
- Next Steps – where to go from here?

# Tutorial Outlines

- *Biomedical Informatics - Challenges and Opportunities*
- Next Generation Bioinformatics Tools – Focus on Data Analysis and Integration
- Systems Biology and Network Analysis
- Biomedical Informatics and the Cloud: Models and Security
- Case Studies of Discoveries using Biological Networks
- HPC in Network Analysis of High Throughput Biological Data
- Integration of different aspects of Biomedical Informatics
- Next Steps – where to go from here?

# Biosciences will never be the same

- IT changed the world forever
- So much biological data is currently available
- The availability of data shifted many branches in Biosciences from pure experimental disciplines to knowledge based disciplines
- Integrating Computational Sciences and Biosciences is not easy
- The answer is Bioinformatics

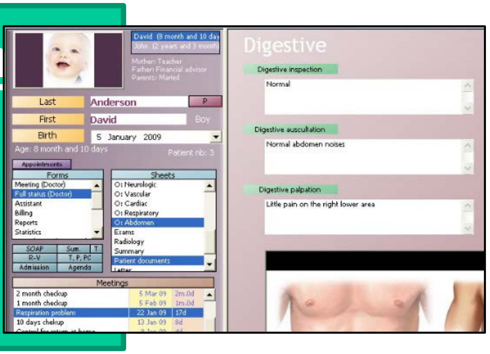
# It is all about the data

- How it all began:
  - Advances in medical instruments and computational technologies led to
  - Massive accumulation of Biomedical data led to
  - The availability of enormous various types of public/private Biomedical data
  - How to take advantage of the available data
- Bioinformatics vs. Health Informatics vs. Biomedical Imaging vs. Public Health Informatics

# Bioinformatics & the Future of Personalized Medicine

## Health Data

- Patient history
- Allergies
- Prescription history
- Family history
- Surgeries



David (8 month and 10 day
John (2 years and 3 month)
Mother: Teacher
Father: Financial advisor
Parents: Maried

| | | |
|---|---|---|
| Last | Anderson | P |
| First | David | Boy |
| Birth | 5 January 2009 | |

Age: 8 month and 10 days
Patient nb: 3

Meetings

| | | |
|---|---|---|
| 2 month checkup | 5 Mar 09 | 2m.0d |
| 1 month checkup | 5 Feb 09 | 1m.0d |
| Respiration problem | 22 Jan 09 | 17d |
| 10 days chekup | 13 Jan 09 | 8d |

Digestive

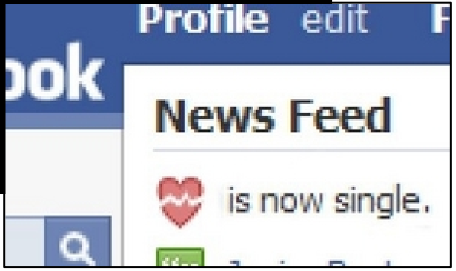
Digestive inspection: Normal

Digestive auscultation: Normal abdomen noises

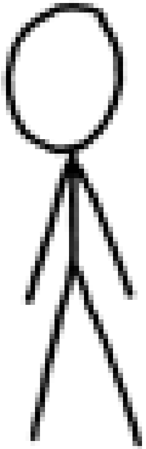
Digestive palpation: Little pain on the right lower area

## Social Data?

- Relationship/friendship data
- Location
- Smog
- Light Exposure



is now single.

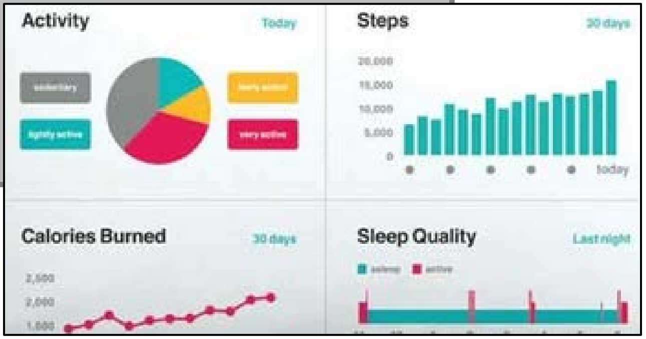


## Genetic Data

- Personalized genome
- Genome *over time?*
- Susceptibilities
- Preventative therapeutics

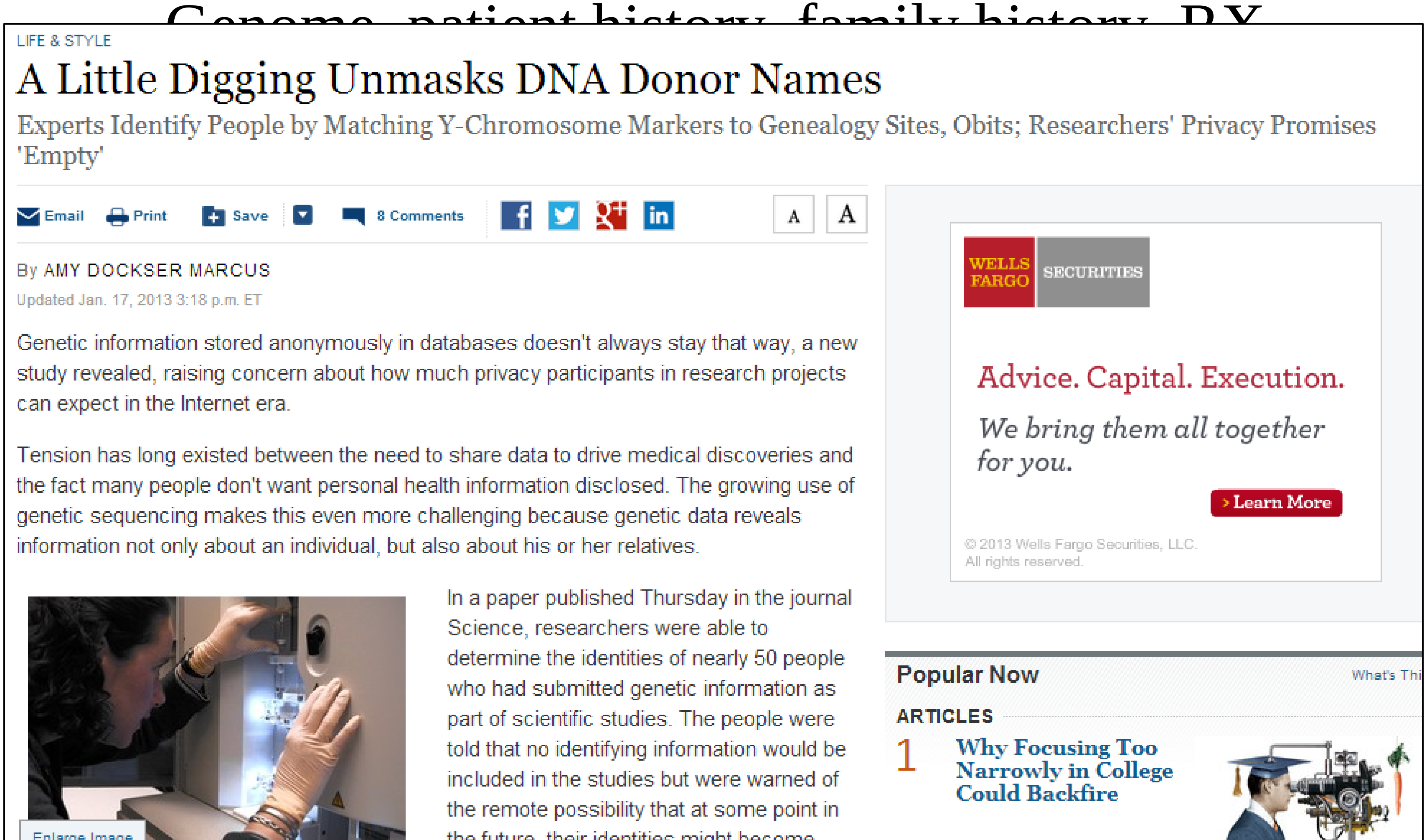
## Wellness Data

- Sleep habits
- Eating habits
- Daily activity
- Stress levels



# Personalized Medicine

- Not so easy to obtain: Health data
  - Genome, patient history, family history, RX



LIFE & STYLE

## A Little Digging Unmasks DNA Donor Names

Experts Identify People by Matching Y-Chromosome Markers to Genealogy Sites, Obits; Researchers' Privacy Promises 'Empty'

By AMY DOCKSER MARCUS

Updated Jan. 17, 2013 3:18 p.m. ET

Genetic information stored anonymously in databases doesn't always stay that way, a new study revealed, raising concern about how much privacy participants in research projects can expect in the Internet era.

Tension has long existed between the need to share data to drive medical discoveries and the fact many people don't want personal health information disclosed. The growing use of genetic sequencing makes this even more challenging because genetic data reveals information not only about an individual, but also about his or her relatives.

In a paper published Thursday in the journal Science, researchers were able to determine the identities of nearly 50 people who had submitted genetic information as part of scientific studies. The people were told that no identifying information would be included in the studies but were warned of the remote possibility that at some point in the future, their identities might become

Popular Now

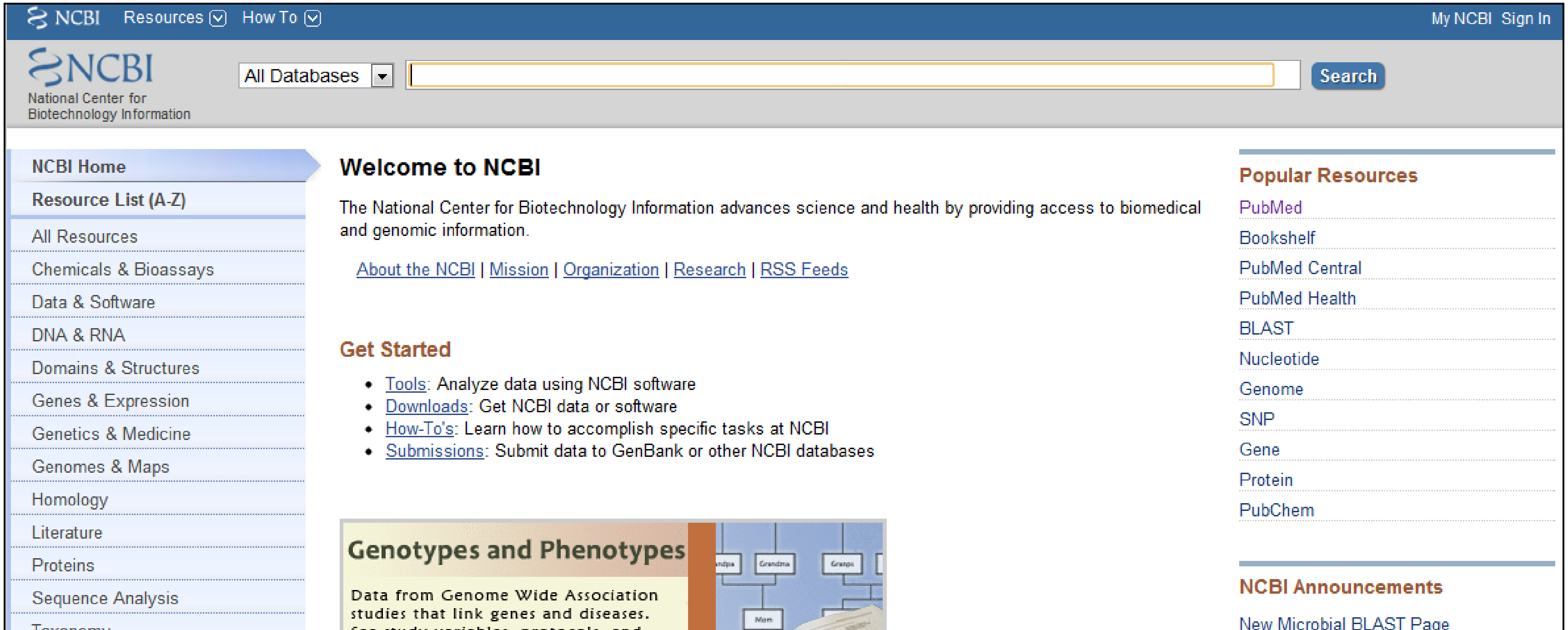
What's Thi

ARTICLES

1 Why Focusing Too Narrowly in College Could Backfire

# What *is* Bioinformatics?

NCBI: “Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.”



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

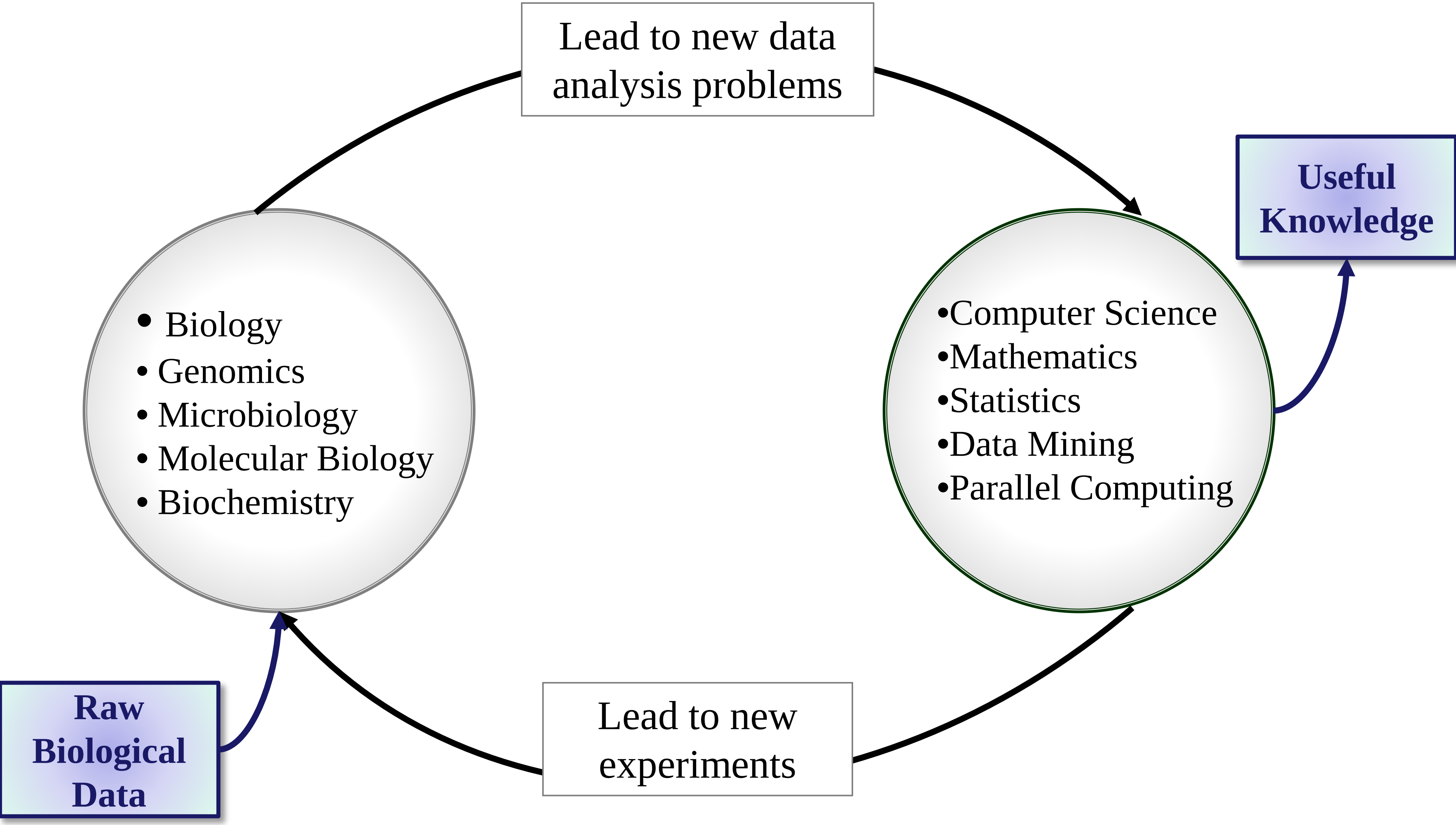
Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Data from Genome Wide Association studies that link genes and diseases. for study variables, protocols, and

# Interactive Aspect of Bioinformatics

UNO
BIOINFORMATICS



Lead to new data analysis problems

- Biology
- Genomics
- Microbiology
- Molecular Biology
- Biochemistry

- Computer Science
- Mathematics
- Statistics
- Data Mining
- Parallel Computing

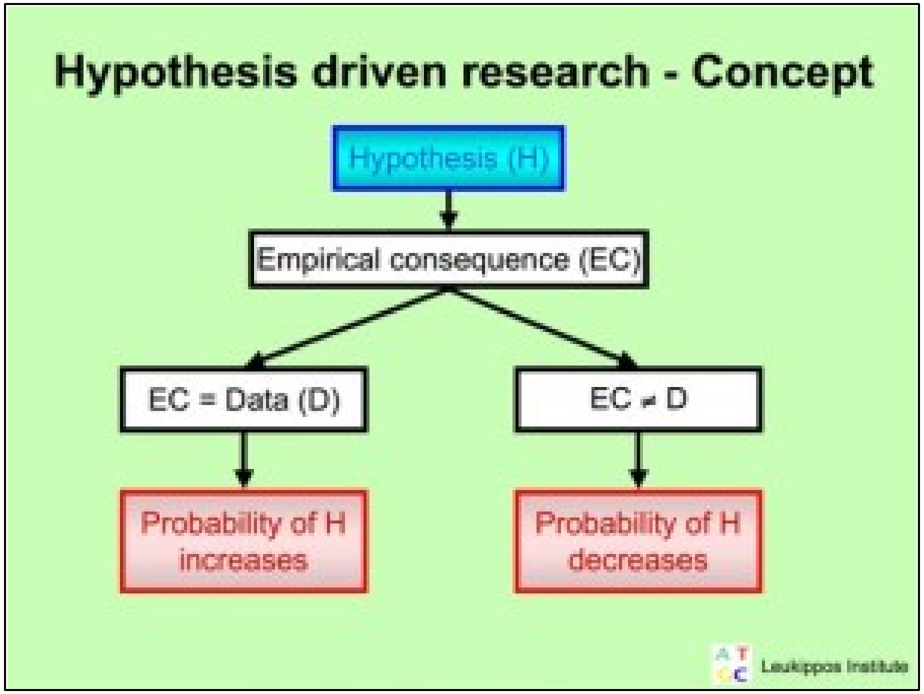
**Useful Knowledge**

**Raw Biological Data**

Lead to new experiments

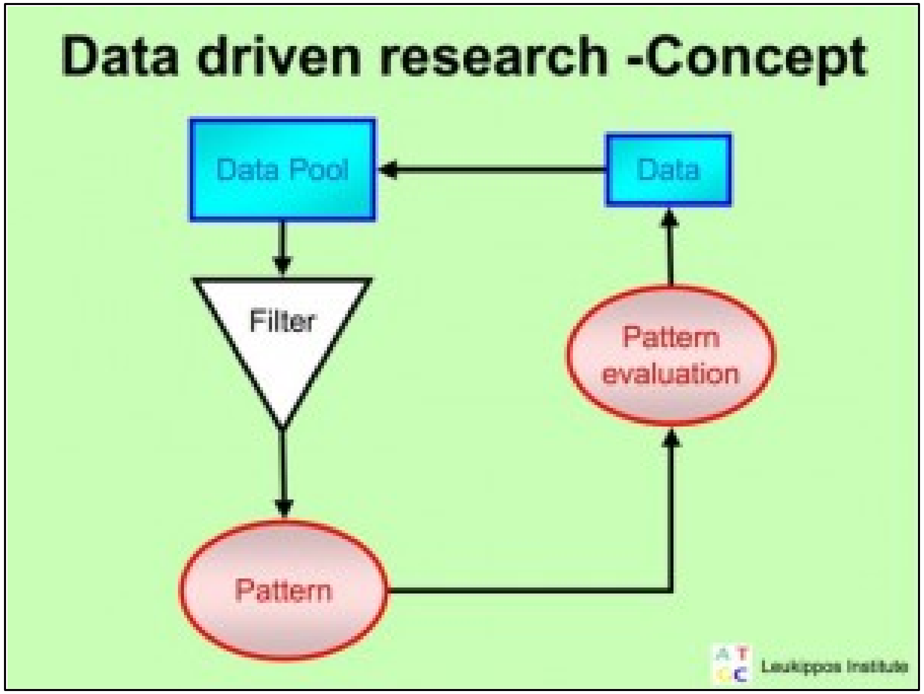
# A Potential Major Change

Data driven research vs. Hypothesis driven research



Hypothesis driven research - Concept

Hypothesis (H)
Empirical consequence (EC)
EC = Data (D)
EC ≠ D
Probability of H increases
Probability of H decreases
Leukippos Institute



Data driven research -Concept

http://www.genome-engineering.com/from-hypothesis-to-data-driven-research.html

# Bioinformatics technology: Where are we?

- High throughput data
- Microarrays
- Next generation sequencing
- Differentially expressed genes
- Biomarkers
- Single position polymorphism and copy number variants
- Genome-wide association studies
- …

# 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
- Computing power
- Lack of integration of data
- …..

# Impact on Industry

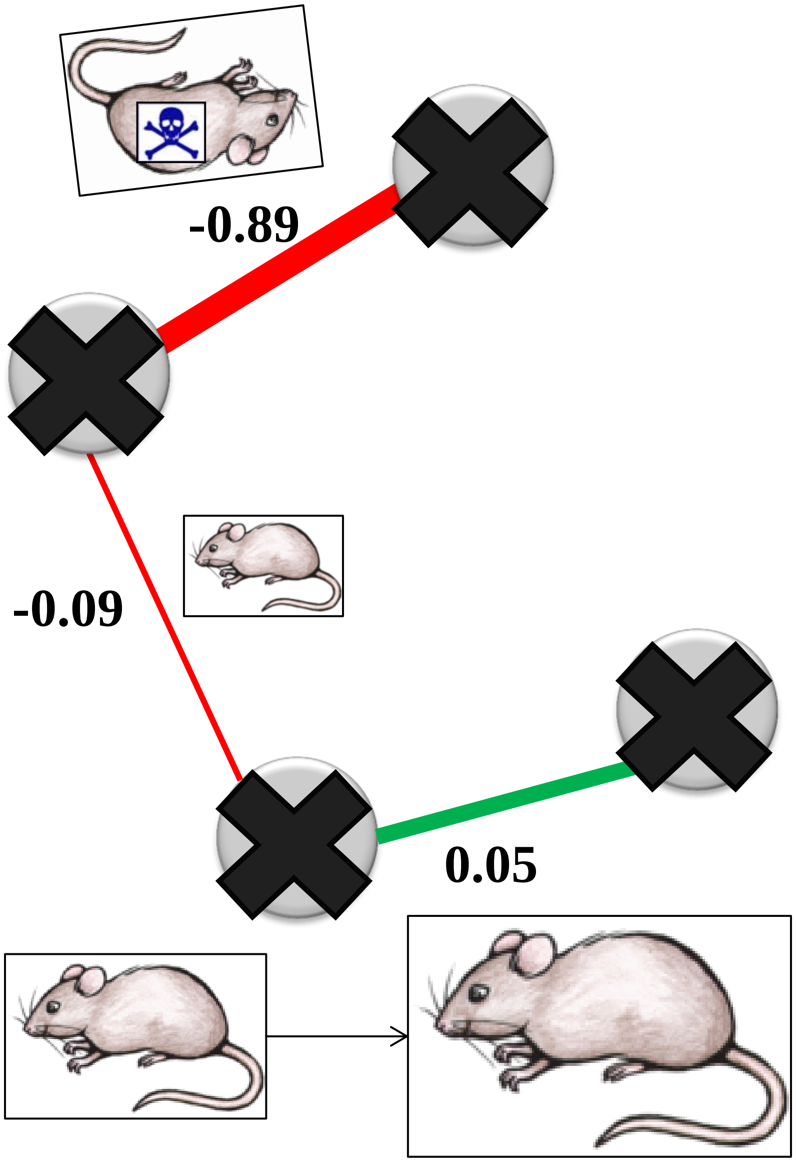
- Increasing number of Biotech companies
- Increasing sales of Biotech drugs
- The emergence of genetic tests
- Emergence of a new paradigm for drugs: right dose of the right drug for the right patient (Pharmacogenomics)

# A Spreading Field

- Although a new field, bioinformatics has spread quickly.  No longer just a theory or an interesting concept, it has moved into the media, research institutes, and threatens to forever alter our lives.
- There is something big related to Bioinformatics almost every day in the digital media:
  - Prince William has a recent Indian ancestry – mitochondrial genome
  - Donation of a mitochondrial genome and legal ramifications
  - Birds with higher number of variants had better survival rate in the Chernobyl crisis
  - There is a group of HIV infected people who don’t develop full fledges AIDS – they are likely to be descendants of those who survived the big Plaque

# Example: Bioinformatics in Health
# The Genetic Interaction Network

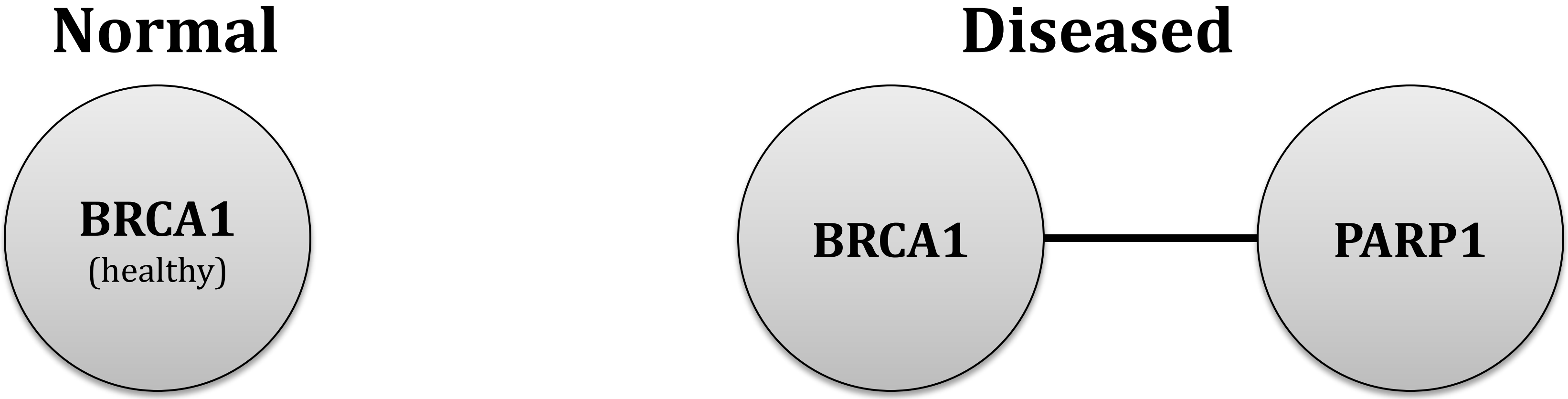
- Protein-protein interaction network
- Metabolome
- Correlation/co-expression network
- Synthetic lethality
- Signal transduction



# Genetic Interaction Networks: Applications

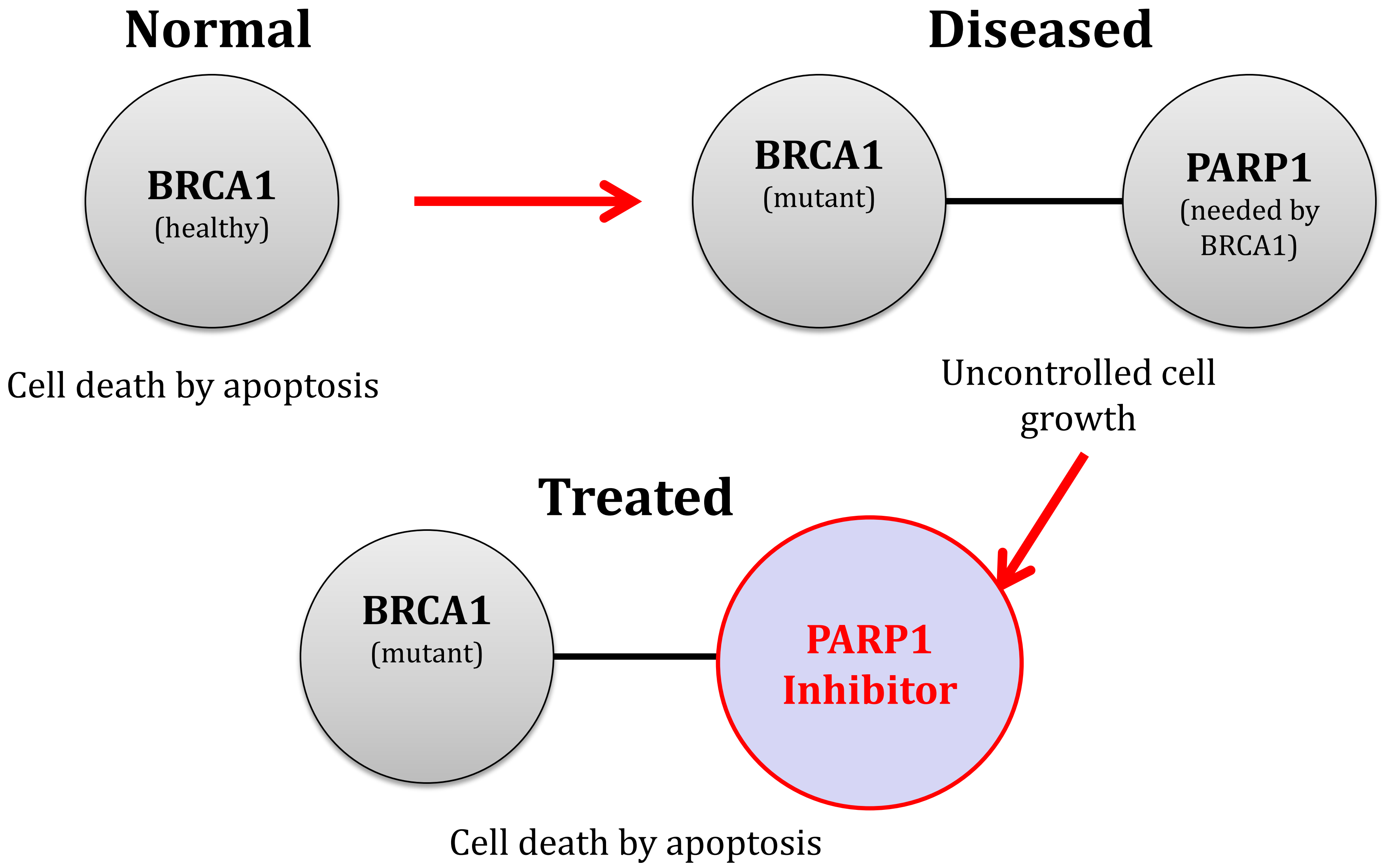
BRCA1 + PARP1 → synthetic lethal interactors

***Tumor cells only***



PARP1 only expressed in tumor cells

# Genetic Interaction Networks: Applications



**Normal**

**BRCA1**
(healthy)

Cell death by apoptosis

**Diseased**

**BRCA1**
(mutant)

**PARP1**
(needed by BRCA1)

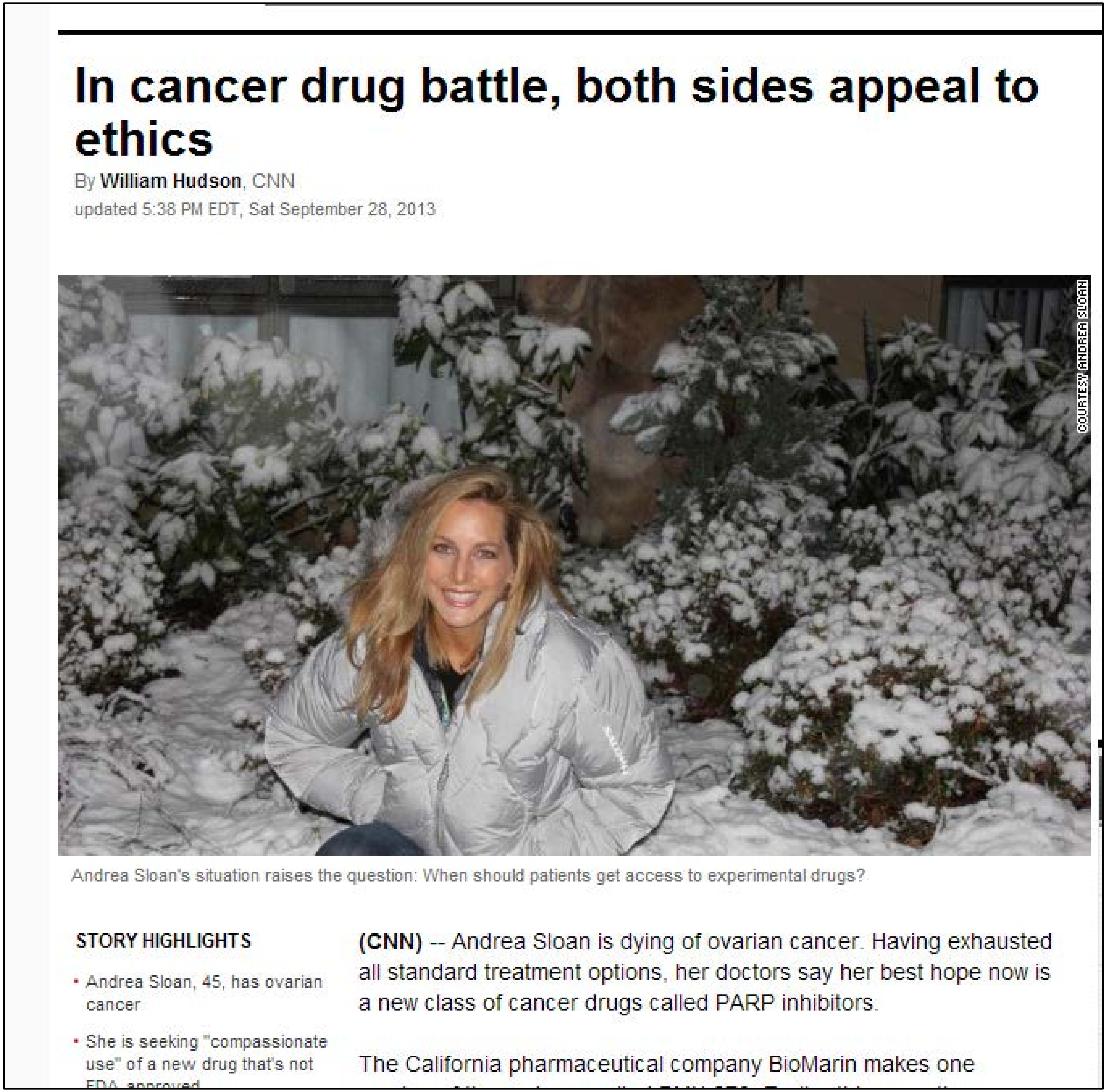
Uncontrolled cell growth

**Treated**

**BRCA1**
(mutant)

**PARP1 Inhibitor**

Cell death by apoptosis



# In cancer drug battle, both sides appeal to ethics

By **William Hudson**, CNN

updated 5:38 PM EDT, Sat September 28, 2013

COURTESY ANDREA SLOAN

Andrea Sloan's situation raises the question: When should patients get access to experimental drugs?

**STORY HIGHLIGHTS**

- Andrea Sloan, 45, has ovarian cancer
- She is seeking "compassionate use" of a new drug that's not FDA approved

**(CNN)** -- Andrea Sloan is dying of ovarian cancer. Having exhausted all standard treatment options, her doctors say her best hope now is a new class of cancer drugs called PARP inhibitors.

The California pharmaceutical company BioMarin makes one

# US Supreme Court says human DNA cannot be patented

Human genes may not be patented, but artificially copied DNA can be claimed as intellectual property, the US Supreme Court has ruled unanimously

AP

Currently, around 40% of the human genome is patented

# Patenting DNA?

Companies attempt to patent DNA they isolate to study disease and cancer, genomes of cancer patients are sequenced to better understand the illness, asthma treatments are catered to different genetic groups, and people marvel at the brilliance of the versatile DNA molecule that has made this all possible.

# More News: Asthma and Genes

- Professors at the Brighton and Sussex Medical School as well as the University of Dundee, did research on severely asthmatic children with a different gene for whom the conventional drug doesn’t work.
- Salmeterol, the drug found in inhalers, acts on beta-2 receptors in the lining of the airways, but as many as 1-in-7 kids have a gene variant: arginine-16 that makes salmeterol ineffective.
- However, for these kids, an anti-inflammatory substitute called montelukast was able to do the job: quality of life increased, attendance at school went up, and visits to out-of-hours GPs’ surgeries went down.

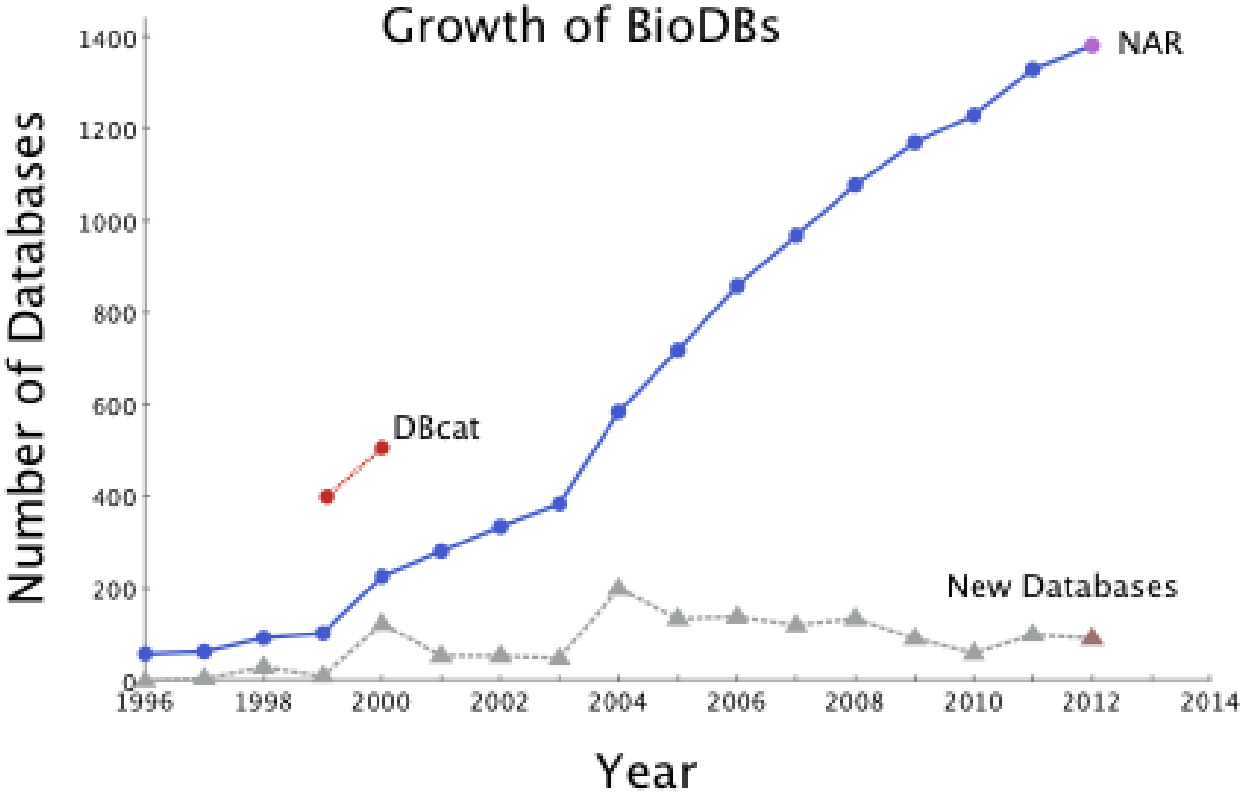
# Potential Implications

- For a small price, the gene of the asthmatic kid can be determined and the proper treatment administered.
- Different treatments for a condition can be tailored to different genetic sub-groups of a population.
- This paves the way for future medicine to be personalized, based on genetic make-up.

# Back to “Biological” Database

- Bioinformatics is a DATA-DRIVEN scientific discipline
- Mainly large set of catalogues sequences
- No extra capabilities of fast access, data sharing or other features found in standard database management systems
- Collection of sequences complemented with additional information such as origin of the data, bibliographic references, sequences function (if known) and others
- Results of many experiments like Microarray Data

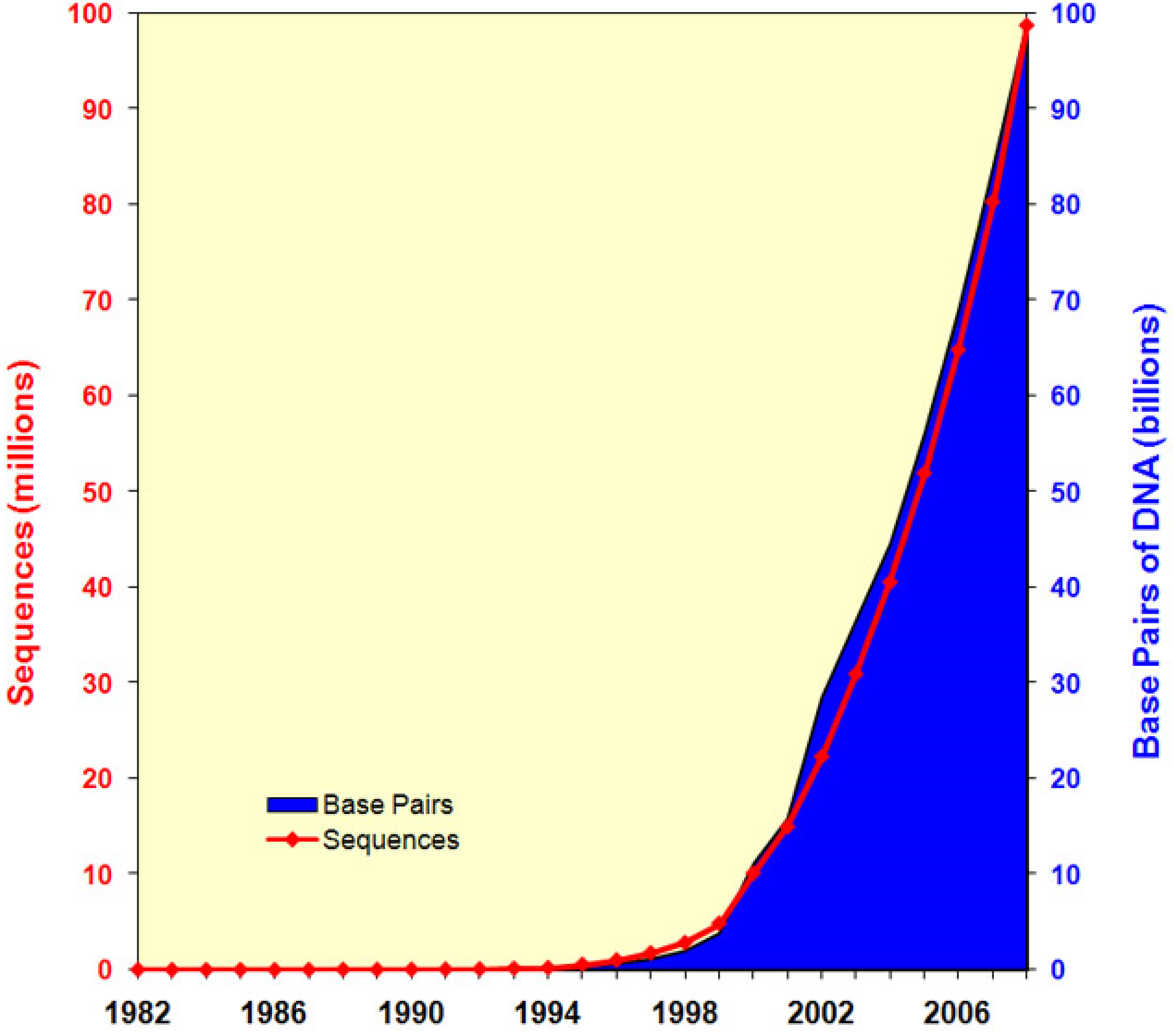
# Growth of Biological Databases



Number of existing (circles) and new databases (triangles) are plotted from 1996 to 2011. New databases are difference between the number of existing databases for each year. DBcat (red) is shown with NAR (blue) counts.

Image Source: http://2.bp.blogspot.com/-o2Tzq-J54ks/TwW006hGADI/AAAAAAAAAYE/caYrHYXfzMg/s1600/db-growth2012.png

# Growth of GenBank
(1982 - 2008)



Source: ncbi.nih.gov

# 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

# Bioinformatics Solutions

- Develop new inventive database models
  - Custom database for specific domains
  - Centralized Structured integrated data
- Develop innovative Bioinformatics tools
  - Clustering/classification algorithms
  - Advanced motif finding approaches
- Systems Biology Approach

# “Big Data”

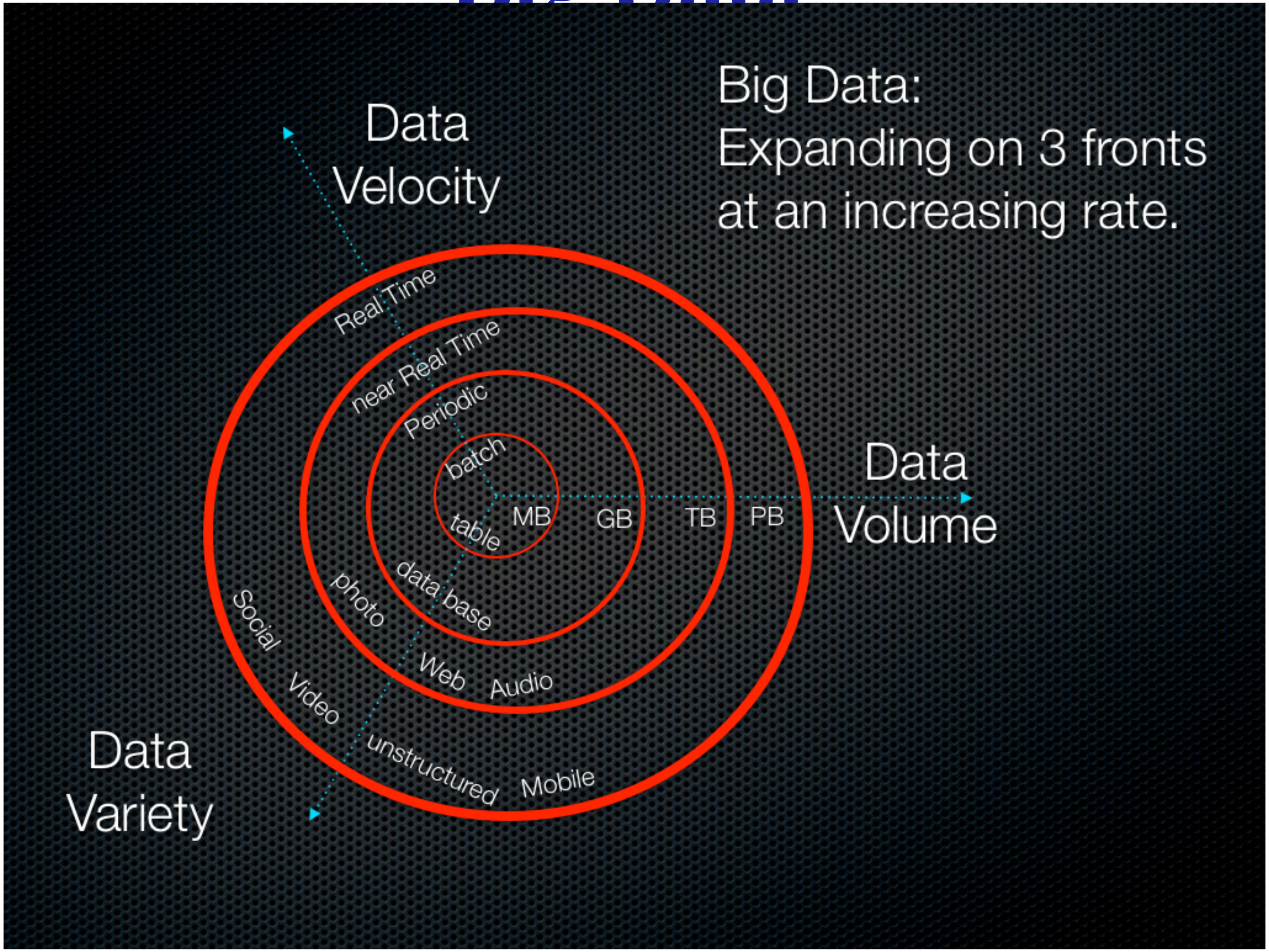


Image Source: http://api.ning.com/files/O6-JQcfS6sxRuZ8I2i5nJVa59xL-krT-a6UqeoLNaHwL2w-JSR-Cy56PmikOywRQgy2gDfYxLAb0Hs*VFr8IePv5QFBJdhDH/BigData.001.jpg

# "Big Data"

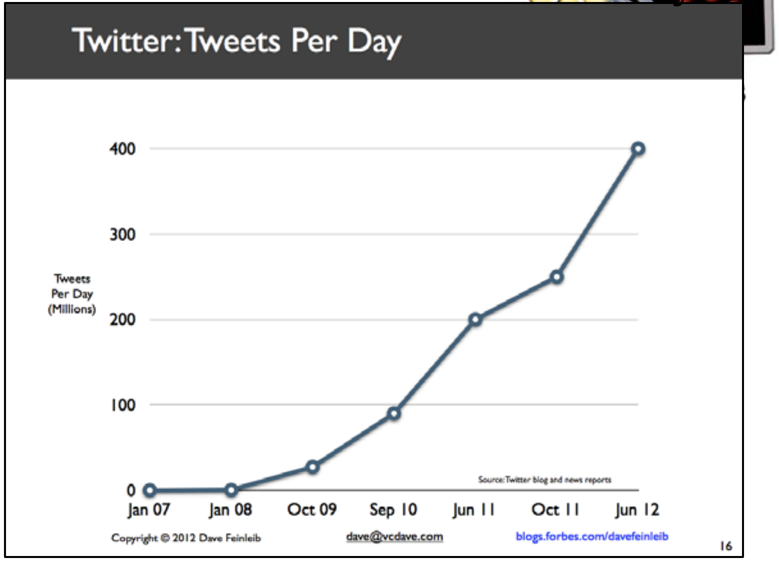
- "Any data too big to be handled by one computer" – Scientist John Rauser[1]

- 90% of worlds data created in last 2 years[2]

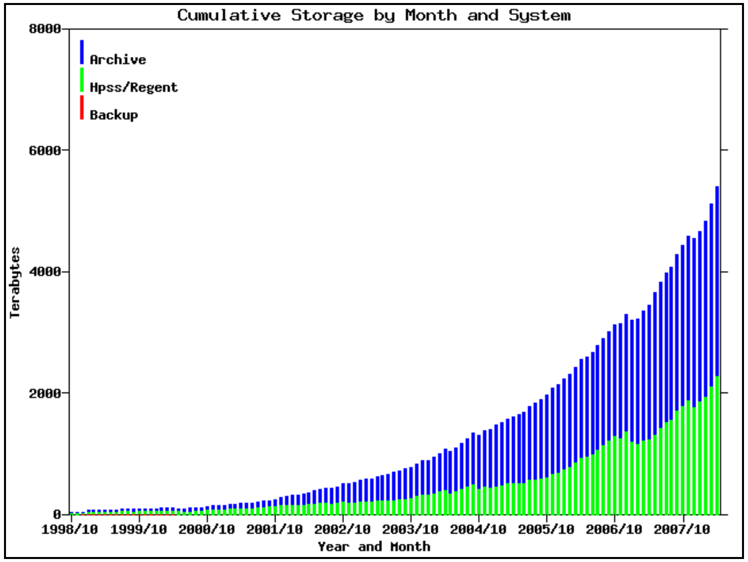
- The four V's of Big Data:
  1. **Volume** – tera to petabytes of info
  2. **Velocity** – Time-sensitive processes
  3. **Variety** – Images, text, database records, ontologies, networks
  4. **Veracity** – Noise vs. signal

- Requires a new set of tools and methods to *store, search, analyze, share, and visualize.*[3]

**Forbes: 400 million Tweets/Day**



NERSC: 6PB of data since 1998



[1] http://www.networkworld.com/news/2012/051012-big-data-259147.html

[2] http://www-01.ibm.com/software/data/bigdata/

[3] http://www.economist.com/node/15557443?story_id=15557443

# What *isn't* Big Data Research?

- Just having more data (inevitable and boring!)
- A problem that can be solved by just more storage space
- Just producing or using large amounts of data
- Traditional schema-aware data and analysis
- Traditional 4-tier Architectures

# What drives Big Data?

- Volume

| | | | |
|---|---|---|---|
| MB | GB | TB | PB |

- Velocity

| | | | |
|---|---|---|---|
| BATCH | PERIODI<br>C | NEAR-REAL<br>TIME | REAL TIME |

- Variety

| | | | |
|---|---|---|---|
| FILES | DB | PHOTO WEB AUDIO | SOCIAL VIDEO MOBILE<br>SENSORS |

- Veracity

| | | | |
|---|---|---|---|
| RAW DATA | RICH<br>ANALYTICS | VALUABLE<br>INTELLIGENCE | ASSURED INFORMATION |

# Big Data Information Systems

- Interactive and Creative
  - Structure and relationships among data are not figured out upfront
  - Data is and its relationships are all unstructured and being produced constantly at a massive rate
  - Model needs to *adapt* which is why networks work so well
- Iterative and value driven
  - What is the business initiative?
  - What are you trying to find out? Use case?

# Tutorial Outlines

- Biomedical Informatics - Challenges and Opportunities
- *Next Generation Bioinformatics Tools – Focus on Data Analysis and Integration*
- Systems Biology and Network Analysis
- Biomedical Informatics and the Cloud: Models and Security
- Case Studies of Discoveries using Biological Networks
- HPC in Network Analysis of High Throughput Biological Data
- Integration of different aspects of Biomedical Informatics
- Next Steps – where to go from here?

# 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

# 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

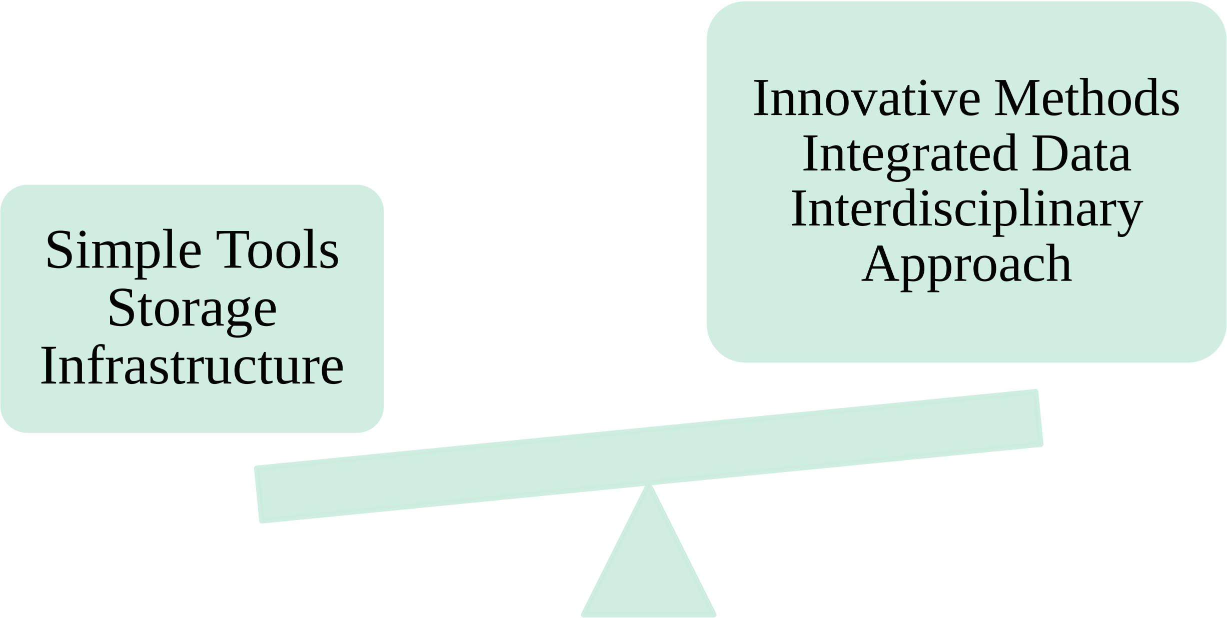
# Bioinformatics Data Cycle

- Data Generation and Collection
- Data Access, Storage and Retrieval
- Data Integration
- Data Visualization
- Analysis and Data Mining
- Decision Support
- Validation and Discovery

# Smart Data Data-Driven Decisions

- 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

# Tipping the Balance



Simple Tools
Storage
Infrastructure

Innovative Methods
Integrated Data
Interdisciplinary
Approach

# 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

# The New Advantage

- The new research paradigm gives a new edge to small to midsize research groups
- Ties between established research groups and medical industries present a two-edge sword
- The focus on new instruments and the contact need to generate new medical data
- Who is better positioned to focus on interdisciplinary scholarly activities
- Emerging Opportunities for everyone!!!

# Focus on Algorithms

- It is all about solving problems – lead to new knowledge
- Integration problem solving techniques with hypothesis based research
- Bioinformatics Algorithms, along with innovative data models, are the central component of Bioinformatics discoveries
- The HPC factor

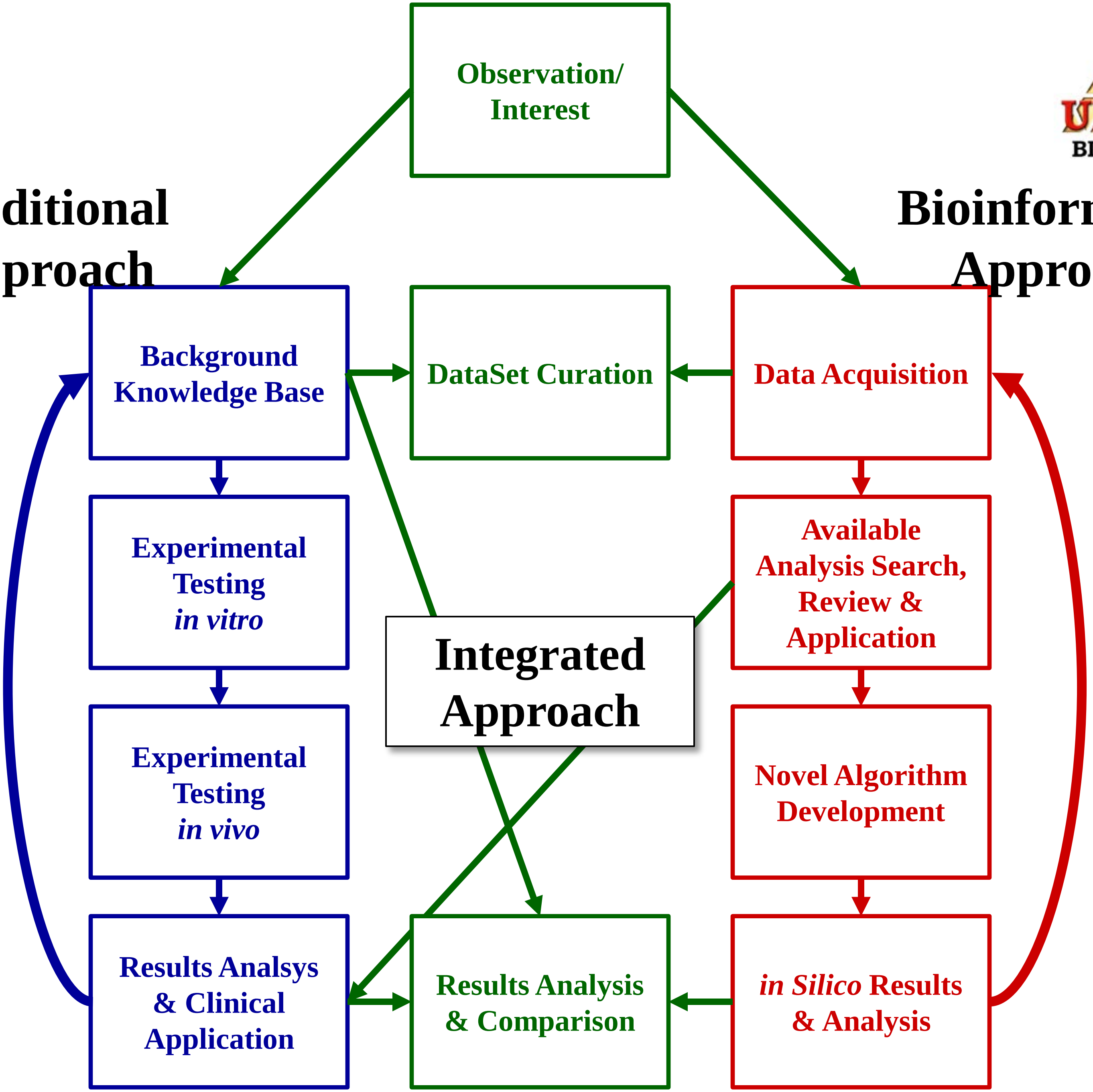
# First Generation Bioinformatics Tools

- 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

# Examples of First Generation Bioinformatics Tools

- Sequence comparison (alignment) tools
- Phylogenetic trees generation tools
- Microarray data statistical tools
- Clustering tools
- Hidden Markov Model (HMM) Based Tools

UNO BIOINFORMATICS



Observation/ Interest

Traditional Approach

Bioinformatics Approach

Background Knowledge Base

DataSet Curation

Data Acquisition

Experimental Testing *in vitro*

Available Analysis Search, Review & Application

Integrated Approach

Experimental Testing *in vivo*

Novel Algorithm Development

Results Analsys & Clinical Application

Results Analysis & Comparison

*in Silico* Results & Analysis

# 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

# A Sample of ICD Tools

- Grammar Based Identification and Classification Tool
- Using Data Compression to Compare Sequences
- Using Cut Orders in the Recognition and Classification of Biological Sequences
- Next Generation Sequencing: A Graph-Theoretic Assembly Tool of Short Reads
- ICD Tools for the Identification of Similarities and Differences in Correlation Networks
- ICD Bioinformatics Tool for Finding Structural Motifs in Proteins

# 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

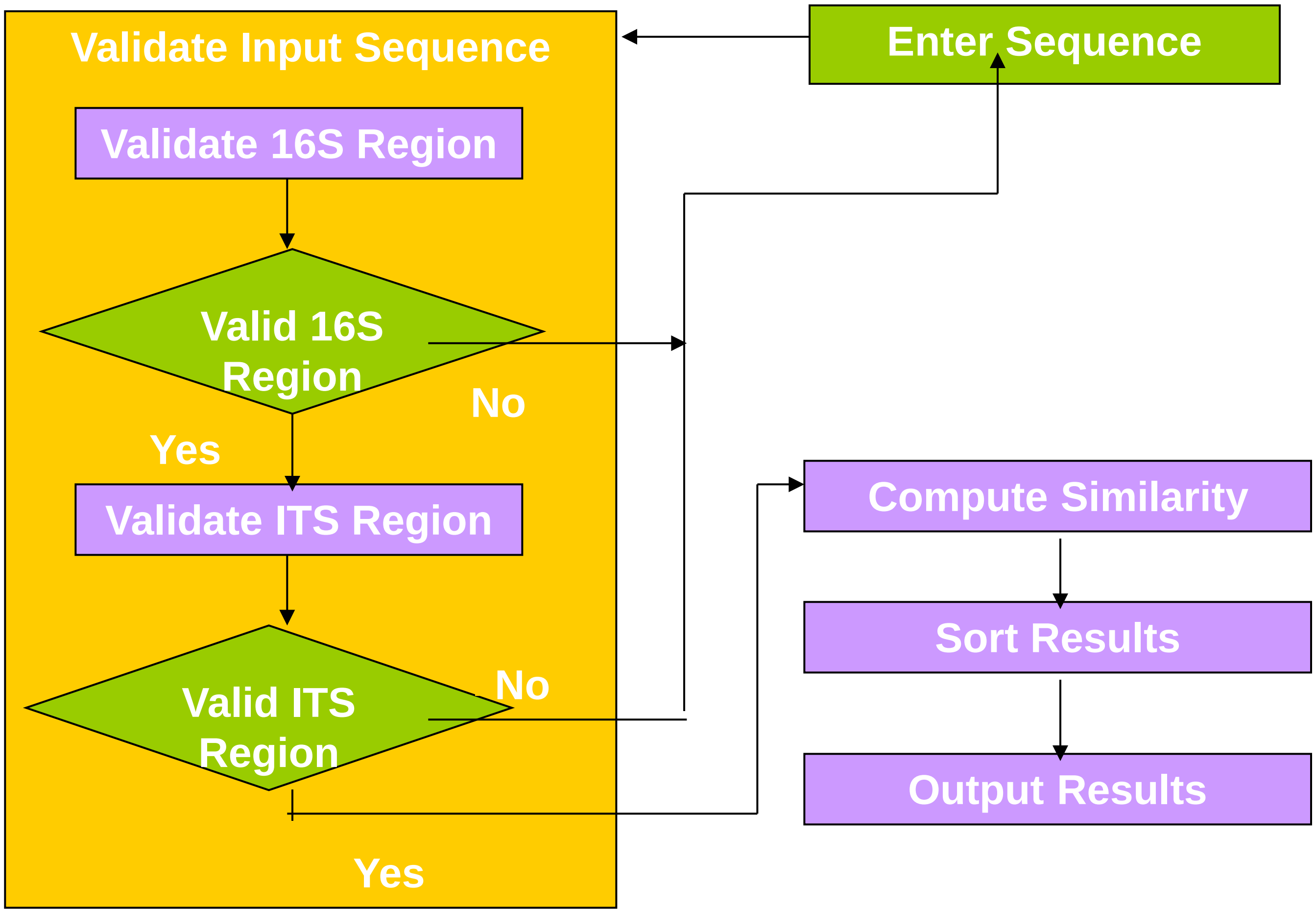
# The Mycobacterium Case Study

- 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

# Case Study: *Mycobacterium*

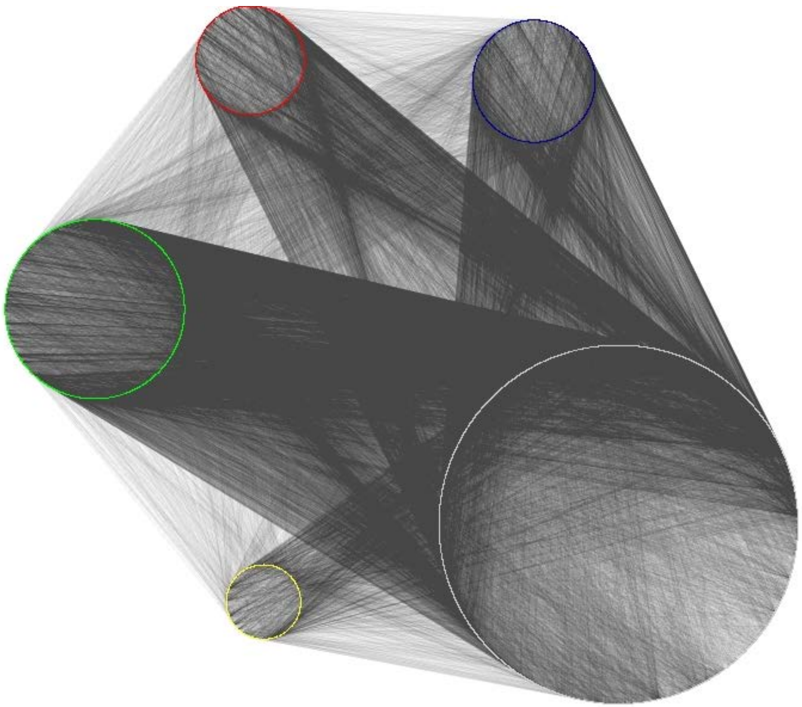


# Nebraska gets its very own organism

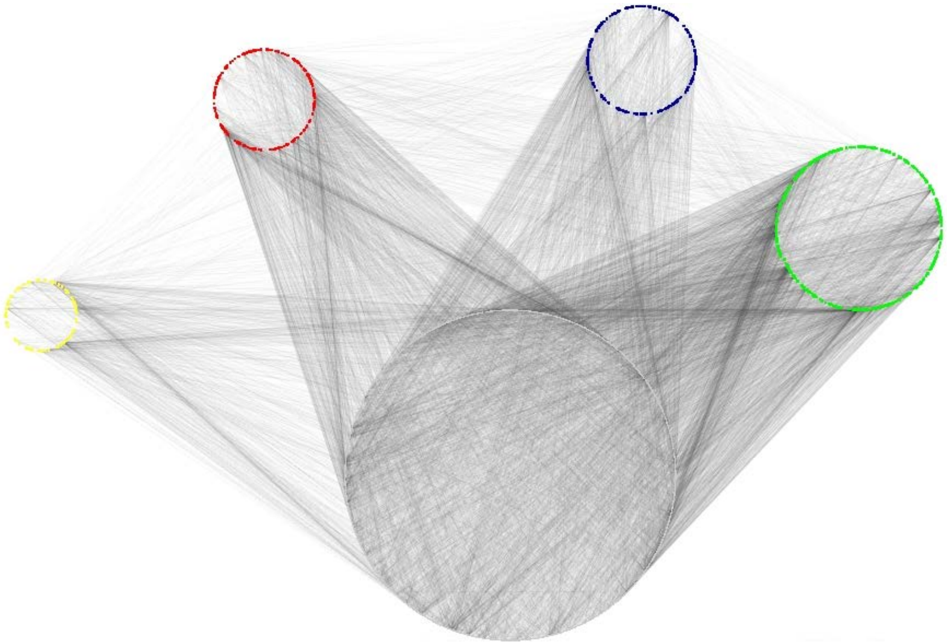
- While trying to pinpoint the cause of a lung infection in local cancer patients, they discovered a previously unknown micro-organism. And they've named it "mycobacterium nebraskense," after the Cornhusker state.
- It was discovered few weeks ago using Mycoalign: A Bioinformatics program developed at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

Source: Omaha World Herald,

# Aging and Biological Networks



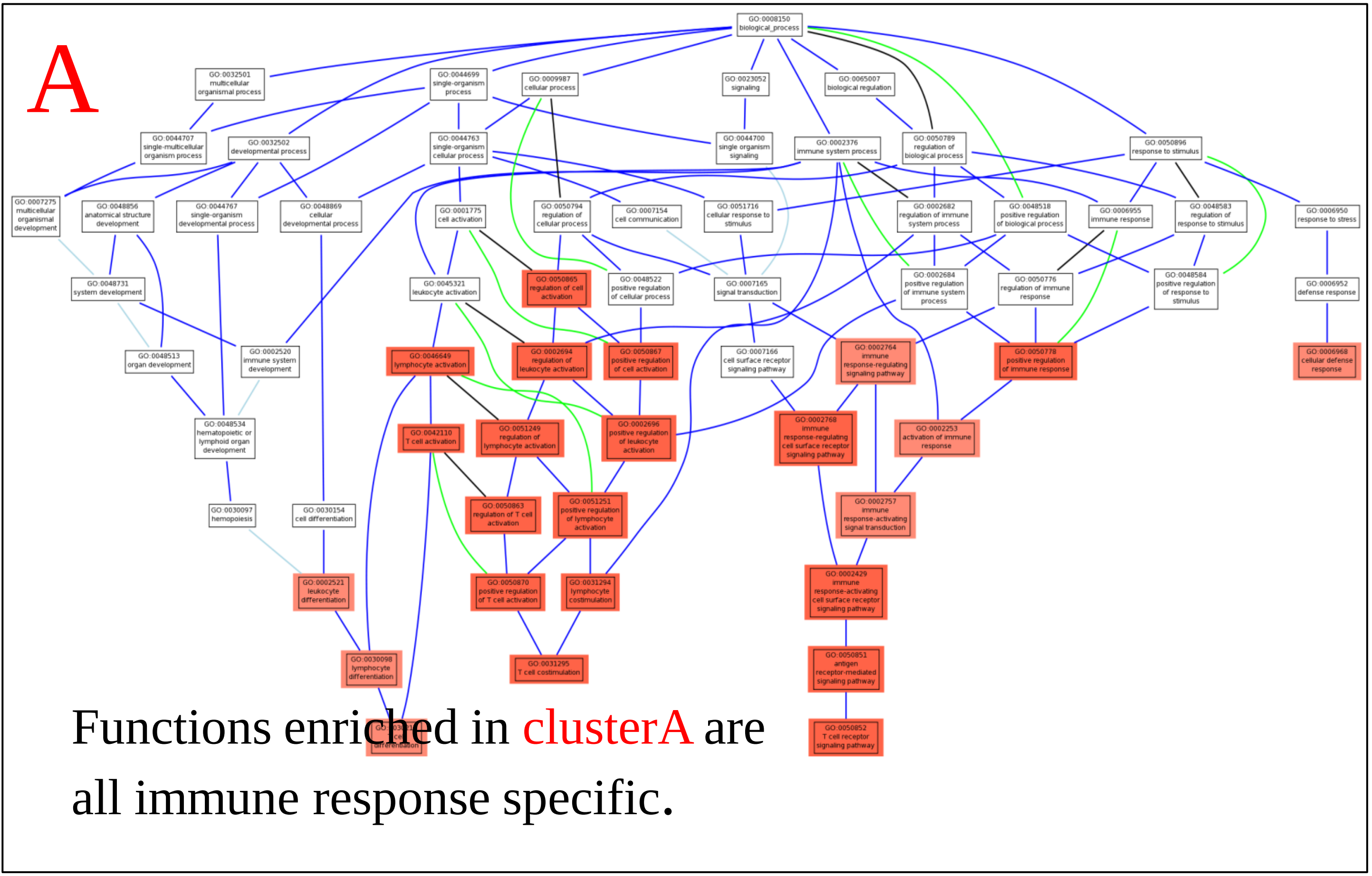
[young]



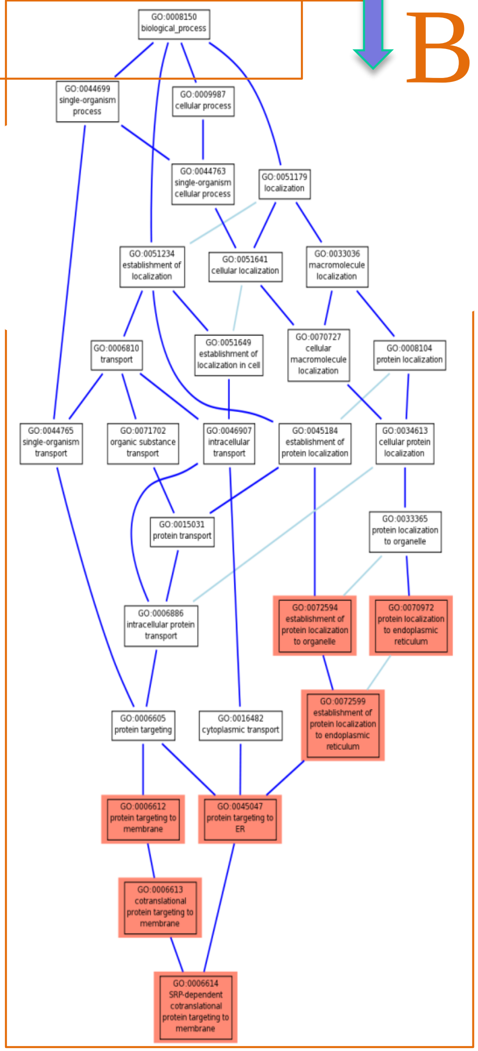
[aged]

# Clusters Enriched in Specific Functions

Functions enriched in clusterB are protein targeting and localization.


A
GO:0008150 biological_process
GO:0032501 multicellular organismal process
GO:0044699 single-organism process
GO:0009987 cellular process
GO:0023052 signaling
GO:0065007 biological regulation
GO:0044707 single-multicellular organism process
GO:0032502 developmental process
GO:0044763 single-organism cellular process
GO:0044700 single organism signaling
GO:0002376 immune system process
GO:0050789 regulation of biological process
GO:0050896 response to stimulus
GO:0007275 multicellular organismal development
GO:0048856 anatomical structure development
GO:0044767 single-organism developmental process
GO:0048869 cellular developmental process
GO:0001775 cell activation
GO:0050794 regulation of cellular process
GO:0007154 cell communication
GO:0051716 cellular response to stimulus
GO:0002682 regulation of immune system process
GO:0048518 positive regulation of biological process
GO:0006955 immune response
GO:0048583 regulation of response to stimulus
GO:0006950 response to stress
GO:0048731 system development
GO:0045321 leukocyte activation
GO:0050865 regulation of cell activation
GO:0048522 positive regulation of cellular process
GO:0007165 signal transduction
GO:0002684 positive regulation of immune system process
GO:0050776 regulation of immune response
GO:0048584 positive regulation of response to stimulus
GO:0006952 defense response
GO:0048513 organ development
GO:0002520 immune system development
GO:0046649 lymphocyte activation
GO:0002694 regulation of leukocyte activation
GO:0050867 positive regulation of cell activation
GO:0007166 cell surface receptor signaling pathway
GO:0002764 immune response-regulating signaling pathway
GO:0050778 positive regulation of immune response
GO:0006968 cellular defense response
GO:0048534 hematopoietic or lymphoid organ development
GO:0042110 T cell activation
GO:0051249 regulation of lymphocyte activation
GO:0002696 positive regulation of leukocyte activation
GO:0002768 immune response-regulating cell surface receptor signaling pathway
GO:0002253 activation of immune response
GO:0030097 hemopoiesis
GO:0030154 cell differentiation
GO:0050863 regulation of T cell activation
GO:0051251 positive regulation of lymphocyte activation
GO:0002757 immune response-activating signal transduction
GO:0002521 leukocyte differentiation
GO:0050870 positive regulation of T cell activation
GO:0031294 lymphocyte costimulation
GO:0002429 immune response-activating cell surface receptor signaling pathway
GO:0030098 lymphocyte differentiation
GO:0031295 T cell costimulation
GO:0050851 antigen receptor-mediated signaling pathway
differentiation
GO:0050852 T cell receptor signaling pathway

Functions enriched in clusterA are all immune response specific.


B
GO:0008150 biological_process
GO:0044699 single-organism process
GO:0009987 cellular process
GO:0044763 single-organism cellular process
GO:0051179 localization
GO:0051234 establishment of localization
GO:0051641 cellular localization
GO:0033036 macromolecule localization
GO:0006810 transport
GO:0051649 establishment of localization in cell
GO:0070727 cellular macromolecule localization
GO:0008104 protein localization
GO:0044765 single-organism transport
GO:0071702 organic substance transport
GO:0046907 intracellular transport
GO:0045184 establishment of protein localization
GO:0034613 cellular protein localization
GO:0015031 protein transport
GO:0033365 protein localization to organelle
GO:0006886 intracellular protein transport
GO:0072594 establishment of protein localization to organelle
GO:0070972 protein localization to endoplasmic reticulum
GO:0006605 protein targeting
GO:0016482 cytoplasmic transport
GO:0072599 establishment of protein localization to endoplasmic reticulum
GO:0006612 protein targeting to membrane
GO:0045047 protein targeting to ER
GO:0006613 cotranslational protein targeting to membrane
GO:0006614 SRP-dependent cotranslational protein targeting to membrane

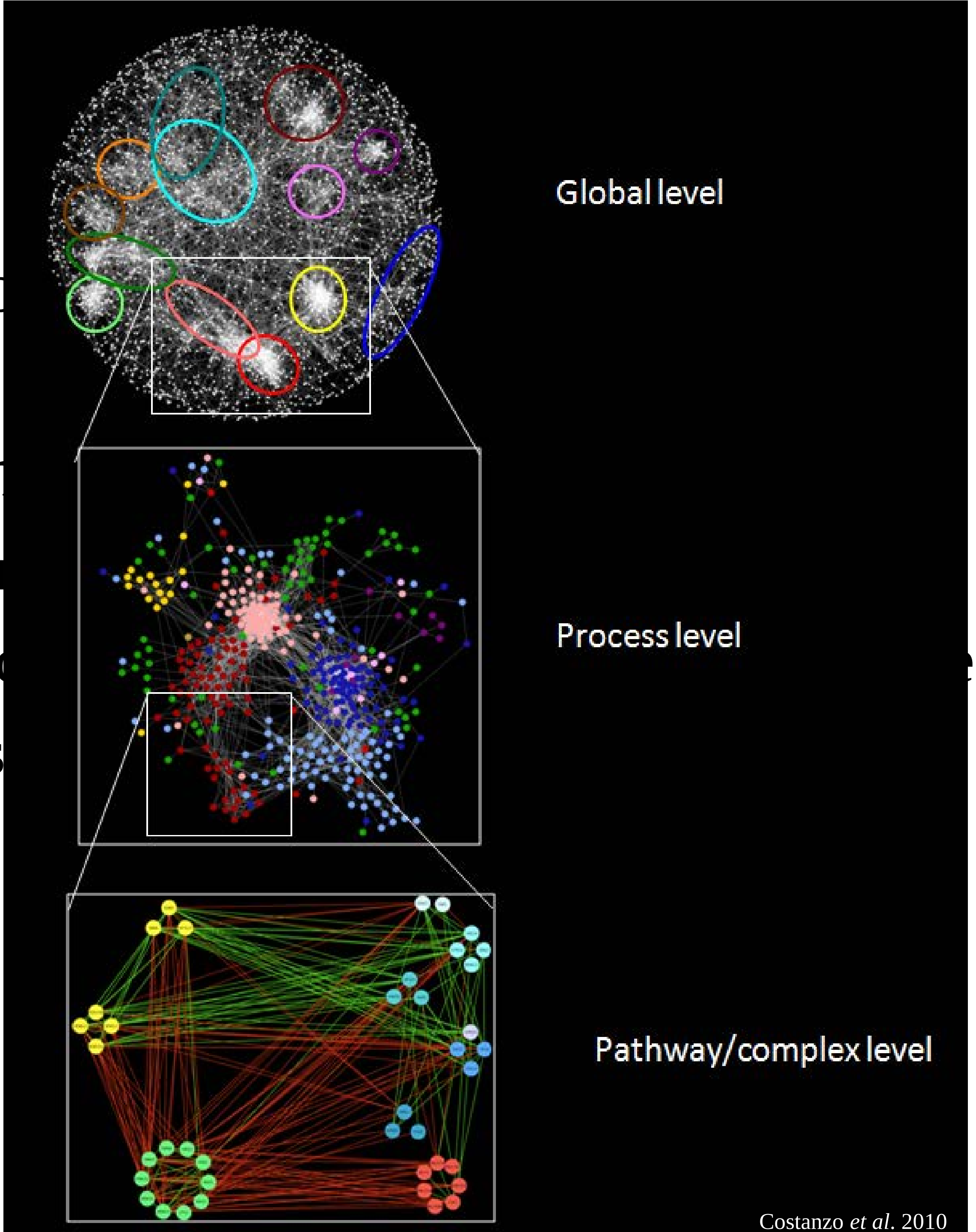
# Tutorial Outlines

- Biomedical Informatics - Challenges and Opportunities
- Next Generation Bioinformatics Tools – Focus on Data Analysis and Integration
- *Systems Biology and Network Analysis*
- Biomedical Informatics and the Cloud: Models and Security
- Case Studies of Discoveries using Biological Networks
- HPC in Network Analysis of High Throughput Biological Data
- Integration of different aspects of Biomedical Informatics
- Next Steps – where to go from here?

# 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

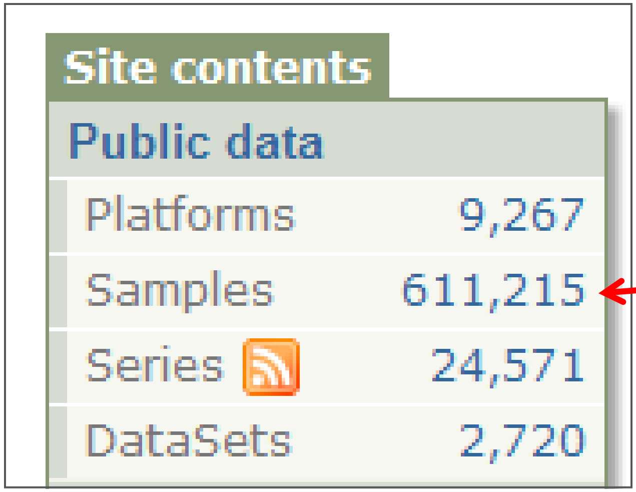
- Netwo
- Systen
  - Inter
  - Usec e
  - Pers



Costanzo *et al*. 2010

# Why Networks?

- Explosion of biological data



Site contents

| Public data | |
|---|---|
| Platforms | 9,267 |
| Samples | 611,215 |
| Series | 24,571 |
| DataSets | 2,720 |

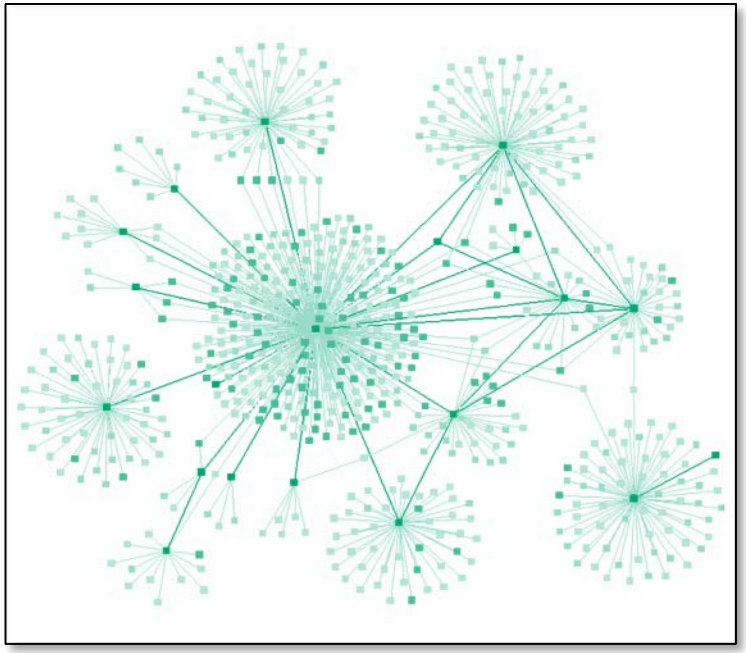
**Each sample can have over 40,000 genes**

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

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

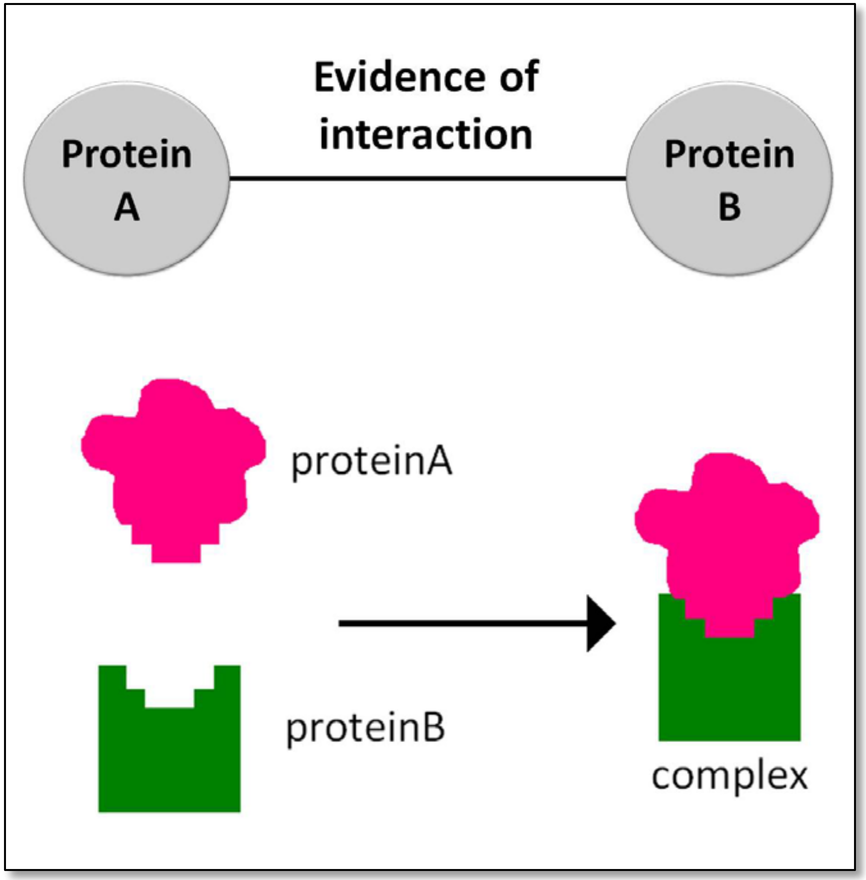
# Biological Networks

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



# Types of Biological Networks

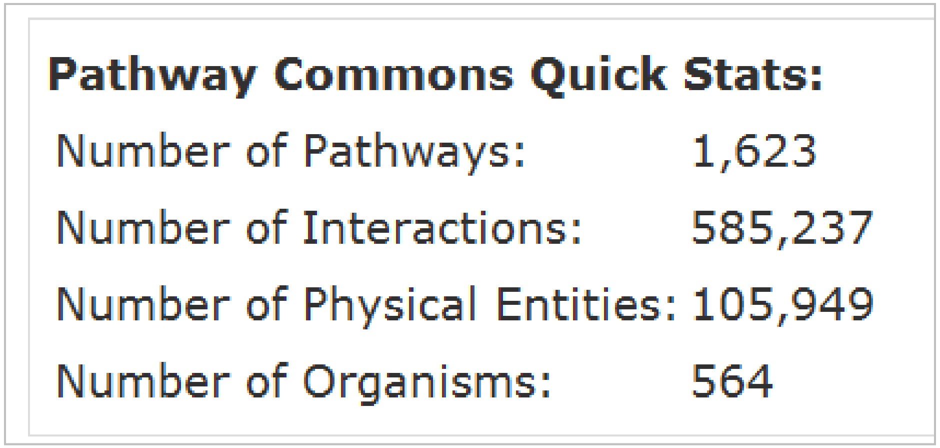
- Protein-protein interaction network
- Metabolome
- Correlation/co-expression network
- Synthetic lethality
- Signal transduction



# PPI Networks

- Built directly from Y2H, Co-IP, TAP
  - Physical detection of interactions

- Databases house PPI data
  - Pathway commons (warehouse)
  - BioGrid
  - HumanCyc
  - ……



**Pathway Commons Quick Stats:**

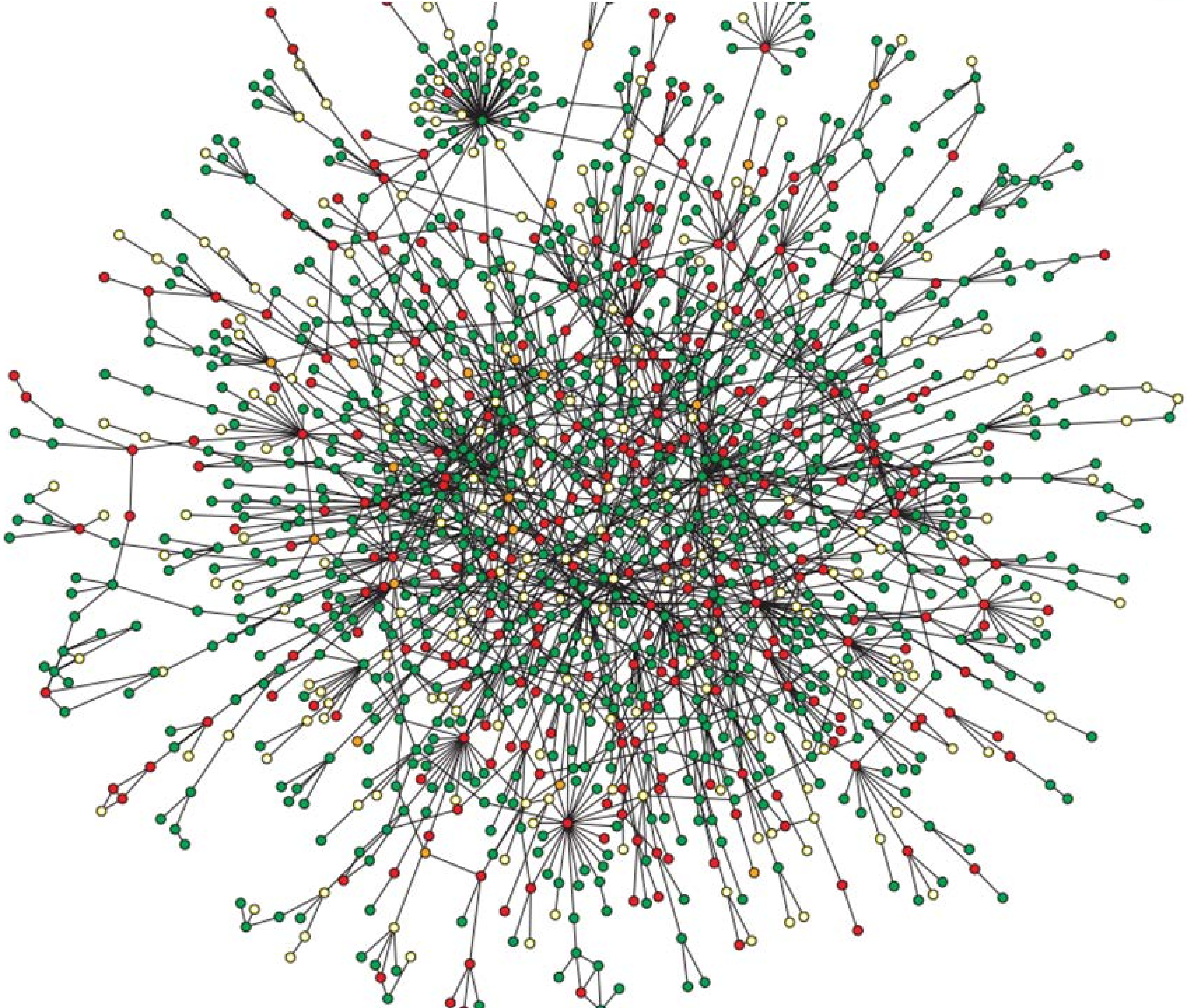
| | |
|---|---|
| Number of Pathways: | 1,623 |
| Number of Interactions: | 585,237 |
| Number of Physical Entities: | 105,949 |
| Number of Organisms: | 564 |

# PPI Networks

- "Hub" proteins in biological networks began from the study of PPI's

- Study done by Jeong (2001)
    - 1870 proteins (nodes)
    - 2240 interactions (edges)

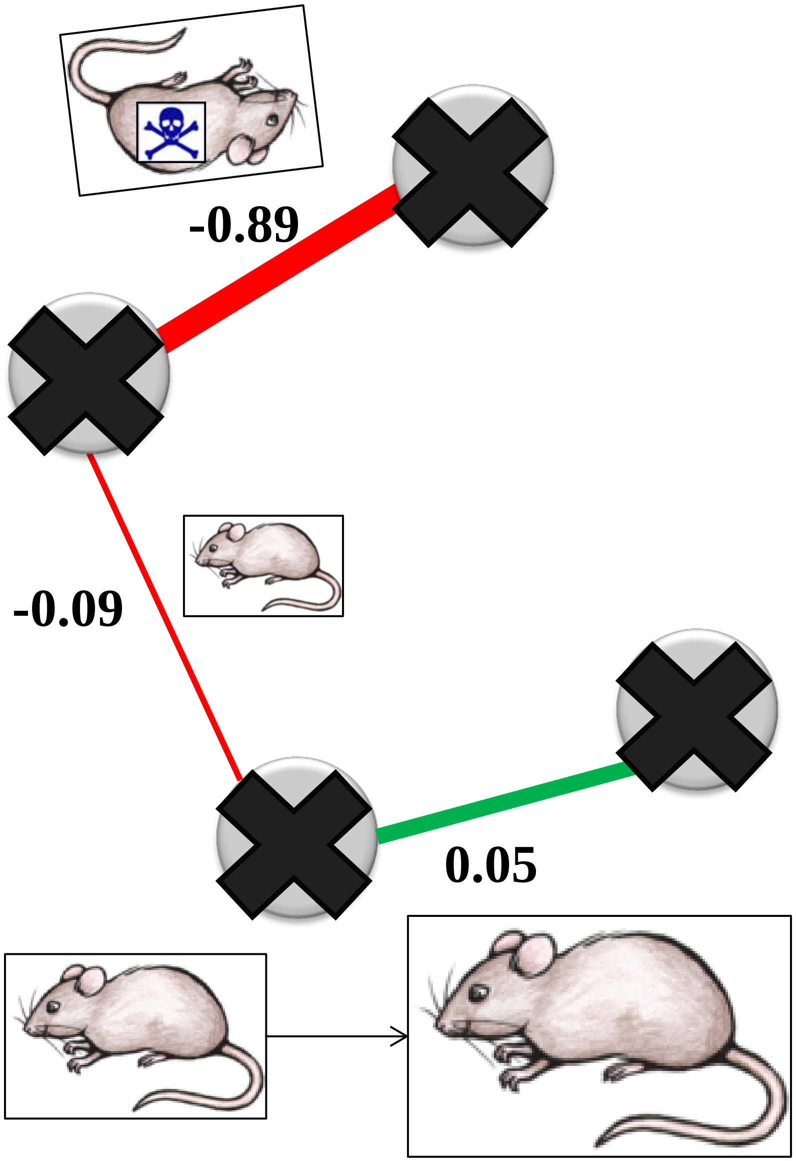
- Forms a scale-free network (*incomplete*)



ICS

# Types of Biological Networks

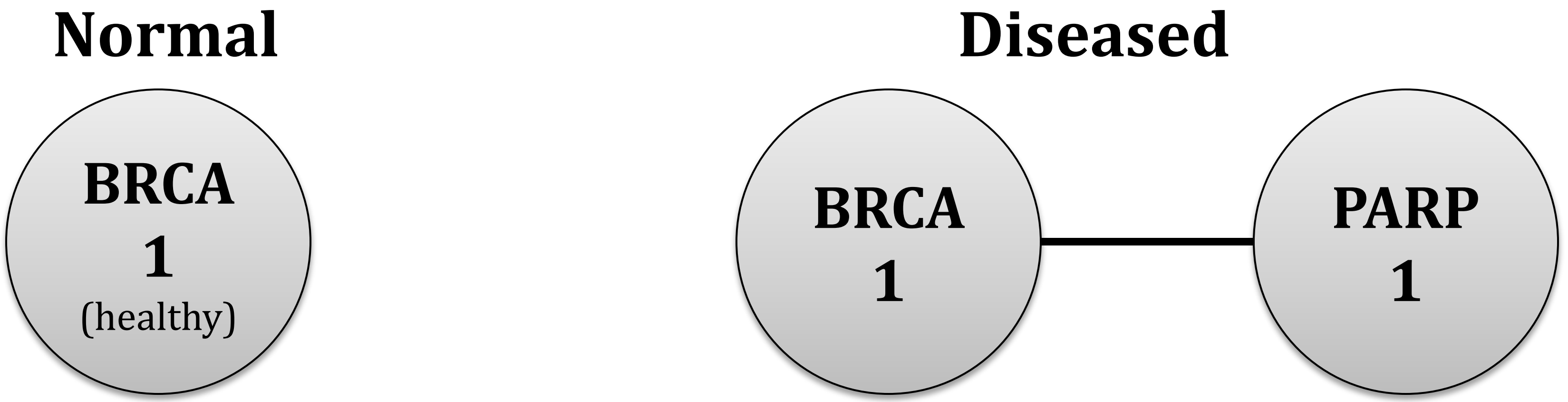
- Protein-protein interaction network
- Metabolome
- Correlation/co-expression network
- Synthetic lethality
- Signal transduction



# Genetic Interaction Networks: Applications

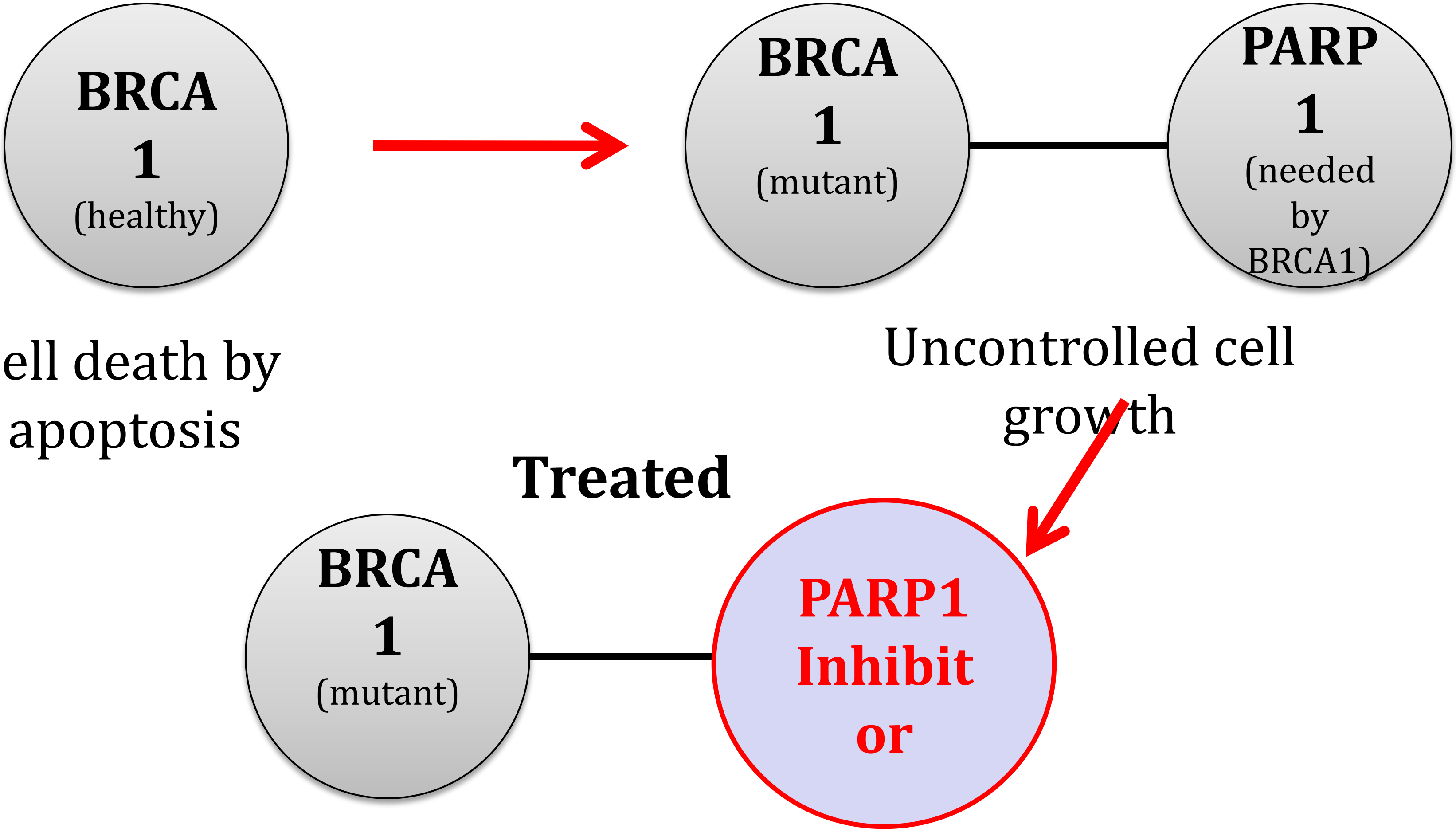
BRCA1 + PARP1 → synthetic lethal interactors

***Tumor cells only***



PARP1 only expressed in tumor cells

# Genetic Interaction Networks: Applications



**Normal**

**BRCA 1**
(healthy)

Cell death by apoptosis

**Diseased**

**BRCA 1**
(mutant)

**PARP 1**
(needed by BRCA1)

Uncontrolled cell growth

**Treated**

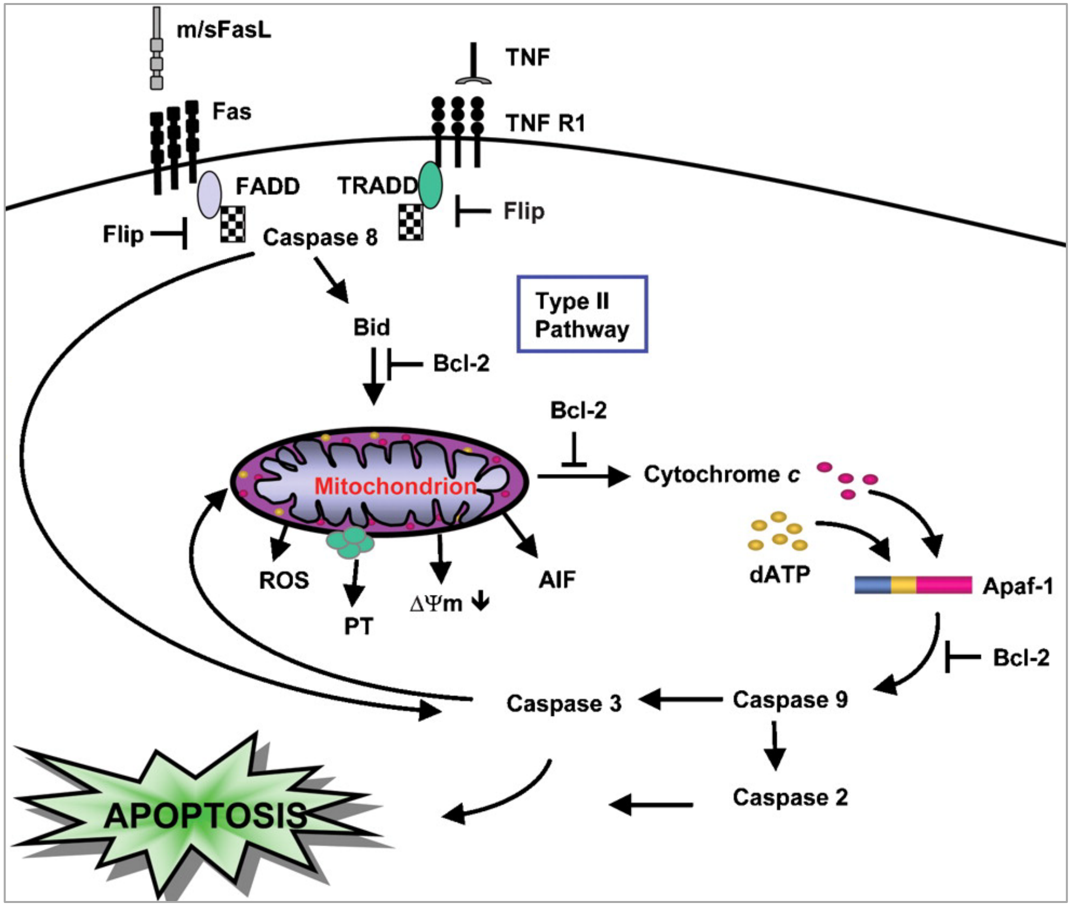
**BRCA 1**
(mutant)

**PARP1 Inhibitor**

Cell death by apoptosis

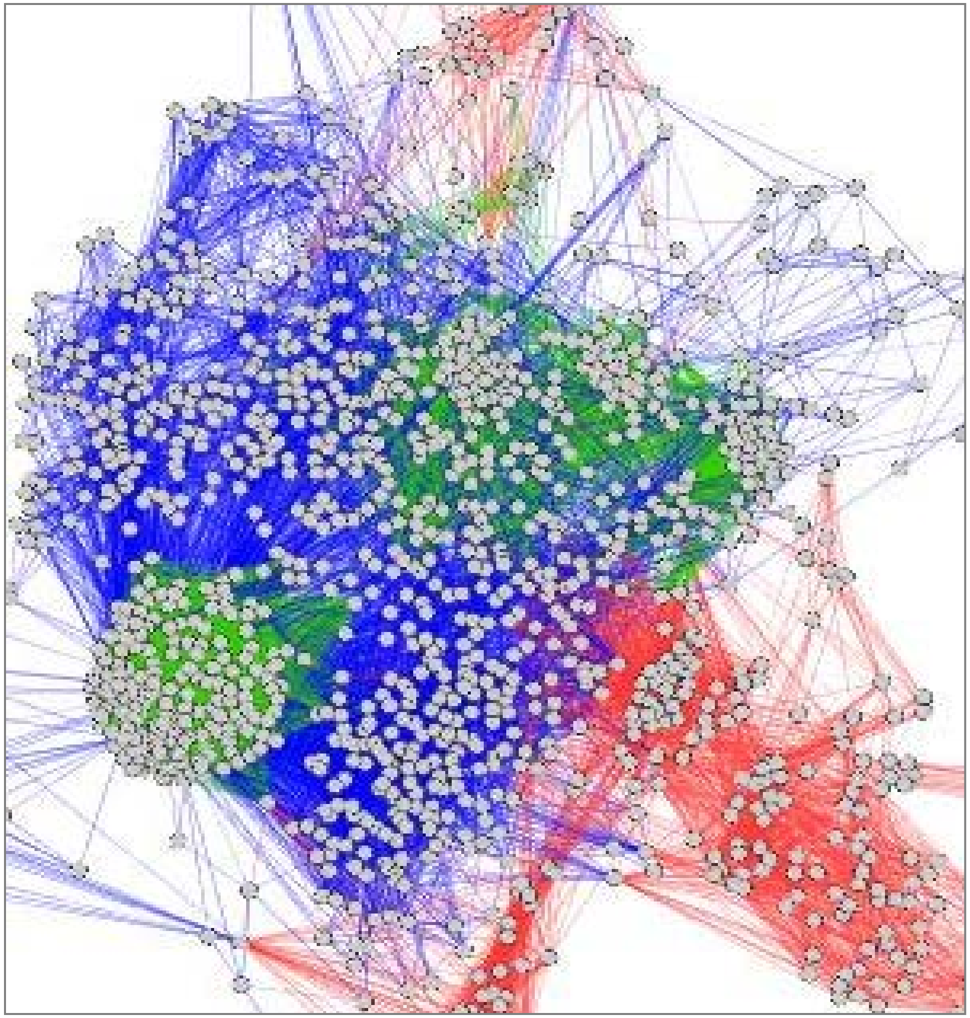
# Types of Biological Networks

- Protein-protein interaction network
- Metabolome
- Correlation/co-expression network
- Synthetic lethality
- Signal transduction



# Types of Biological Networks

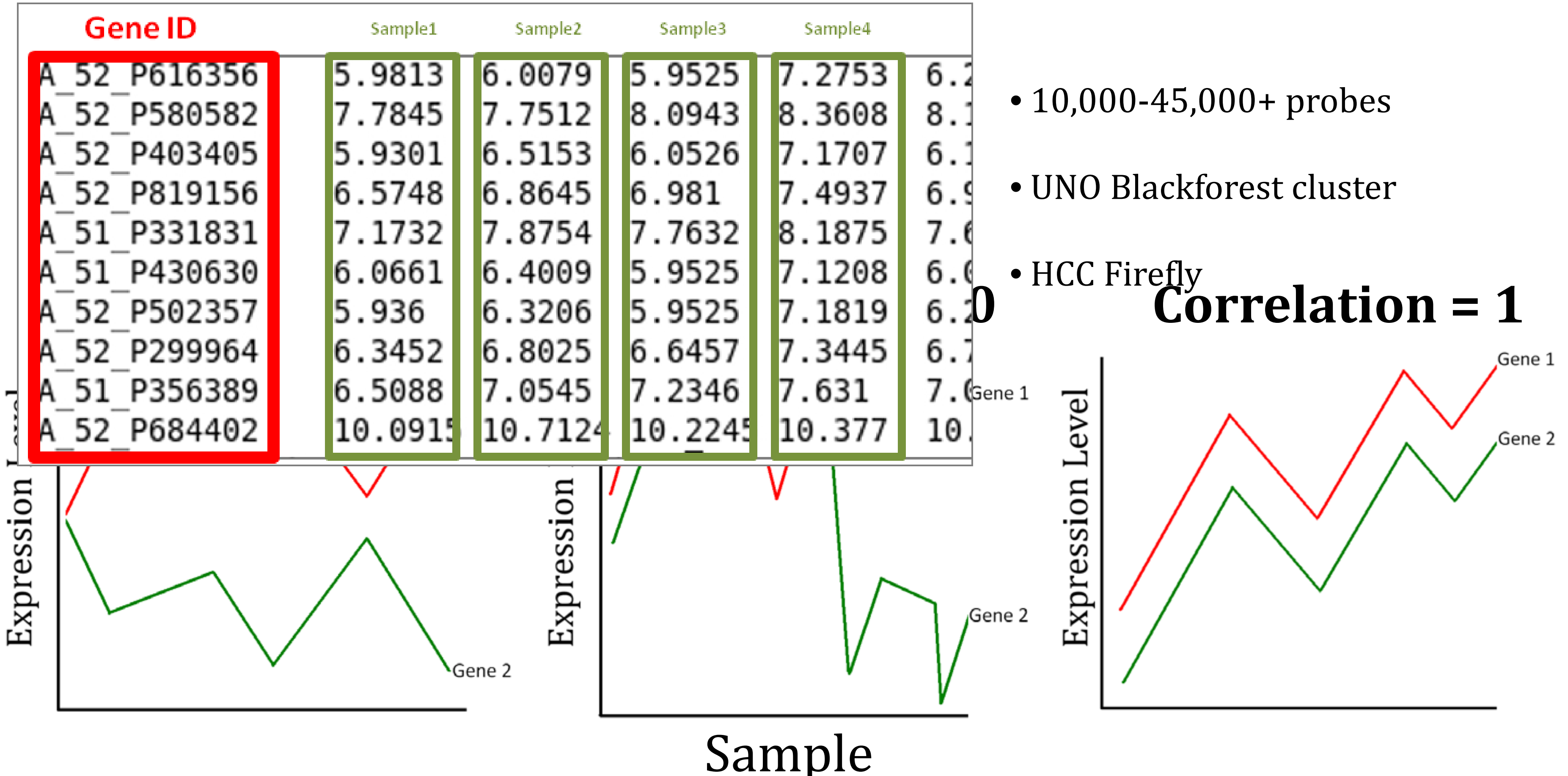
- Protein-protein interaction network
- Metabolome
- Correlation/co-expression network
- Synthetic lethality
- Signal transduction



- Correlation networks
  - What are they? How are they made?
  - How can they be used in biomedical research?
- Network Comparison
  - Identifying common & unique network elements
  - Filtering noise from causative relationships
- Case study
  - Proof of concept using sample expression data
  - What kinds of questions can we ask?

# Correlation Networks

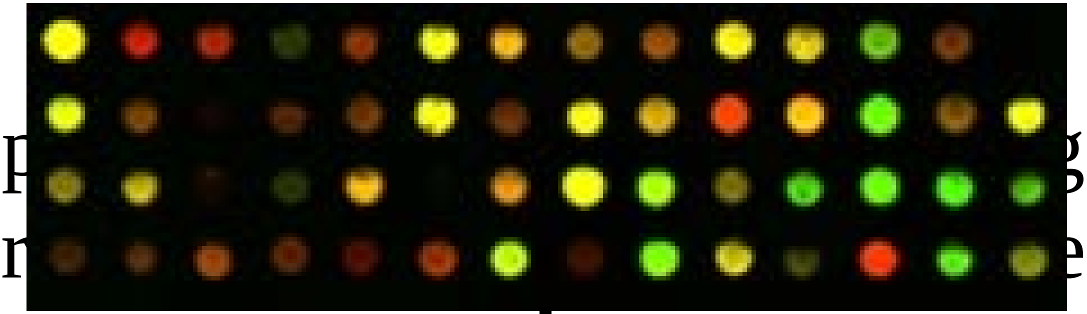
- 10,000-45,000+ probes
- UNO Blackforest cluster
- HCC Firefly



| Gene ID | Sample1 | Sample2 | Sample3 | Sample4 | |
|---|---|---|---|---|---|
| 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 | 6.0 |
| 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. |

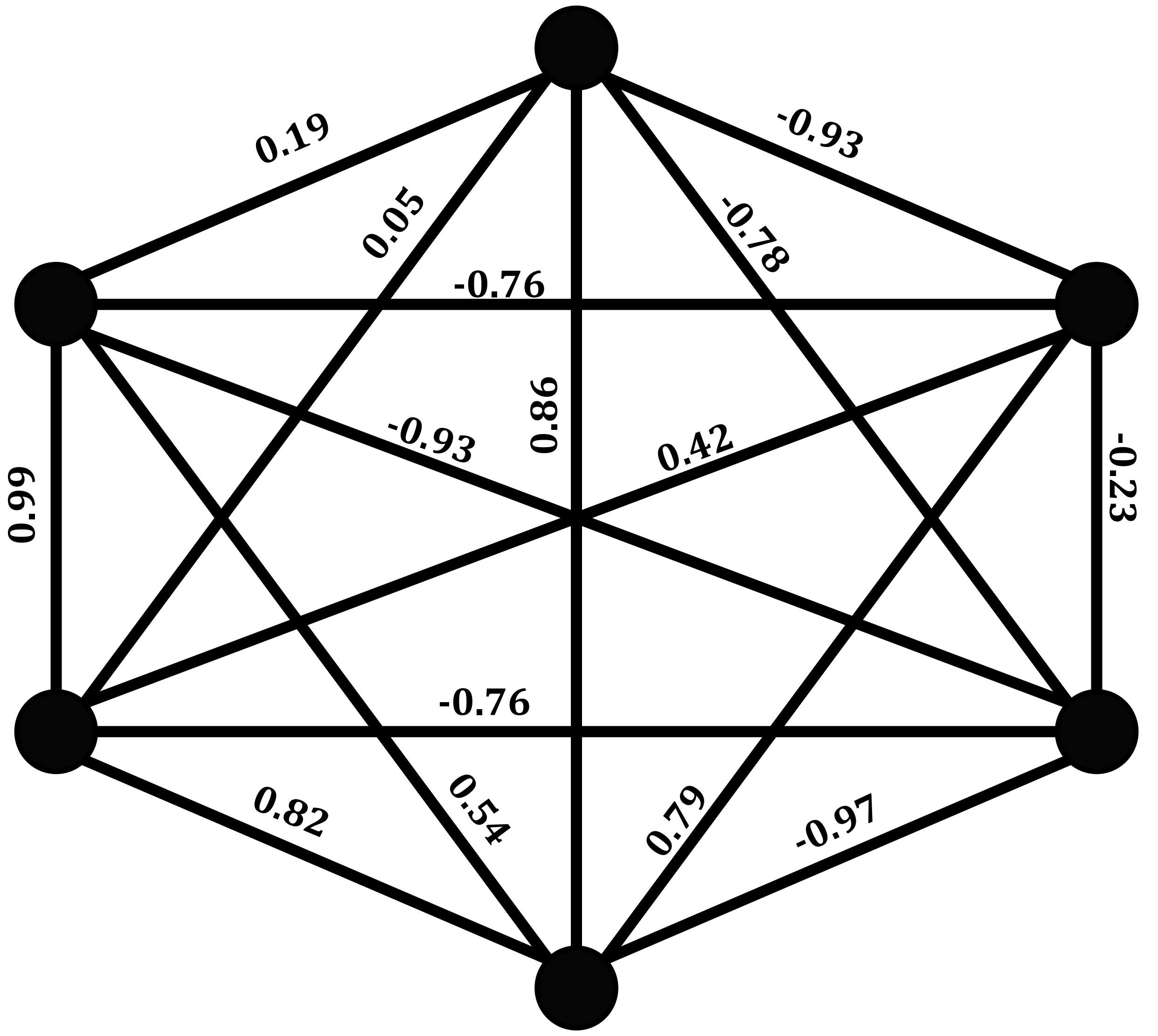
# Correlation Networks

A grap gree of
co entity



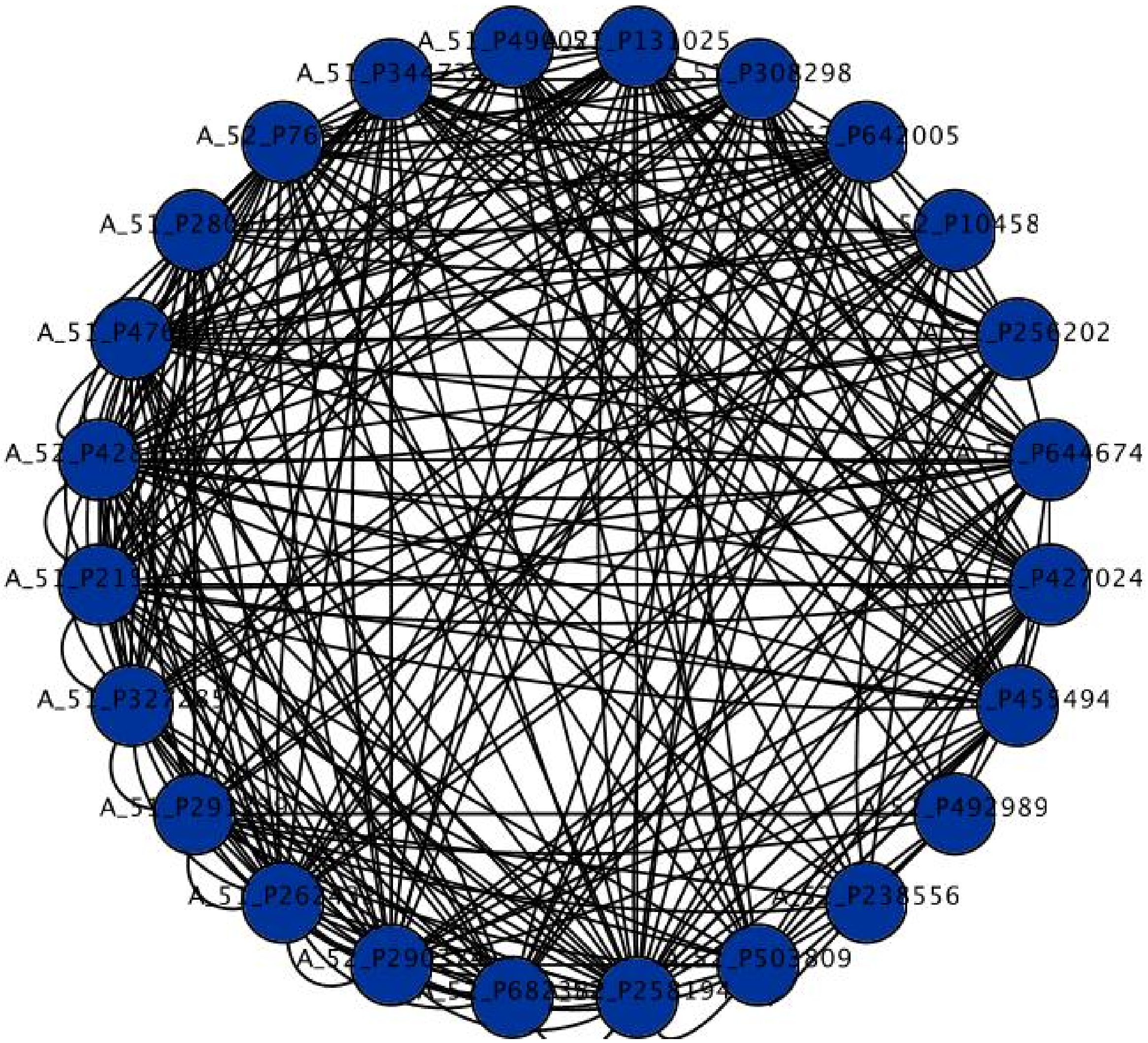
|  | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| **Gene 1** | 10.5 | 11.0 | 12.1 |
| **Gene 2** | 3.2 | 3.3 | 2.9 |
| **Gene 3** | 1.4 | 1.5 | 0.9 |
| **Gene 4** | 7.8 | 7.1 | 8.2 |

# Correlation Networks



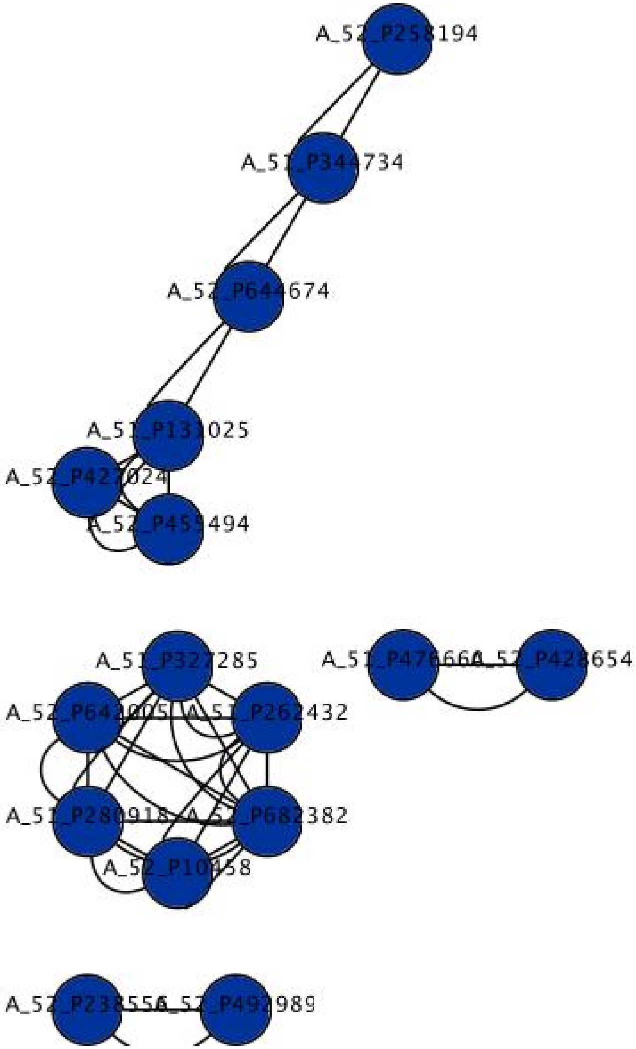
# Correlation Networks

24 node sample
Threshold: 0.00-1.00



# Correlation Networks

24 node sample
Threshold: 0.80-1.00



# Correlation Network Applications

- "Versus" analysis
  - Normal vs. disease
  - Times/environments

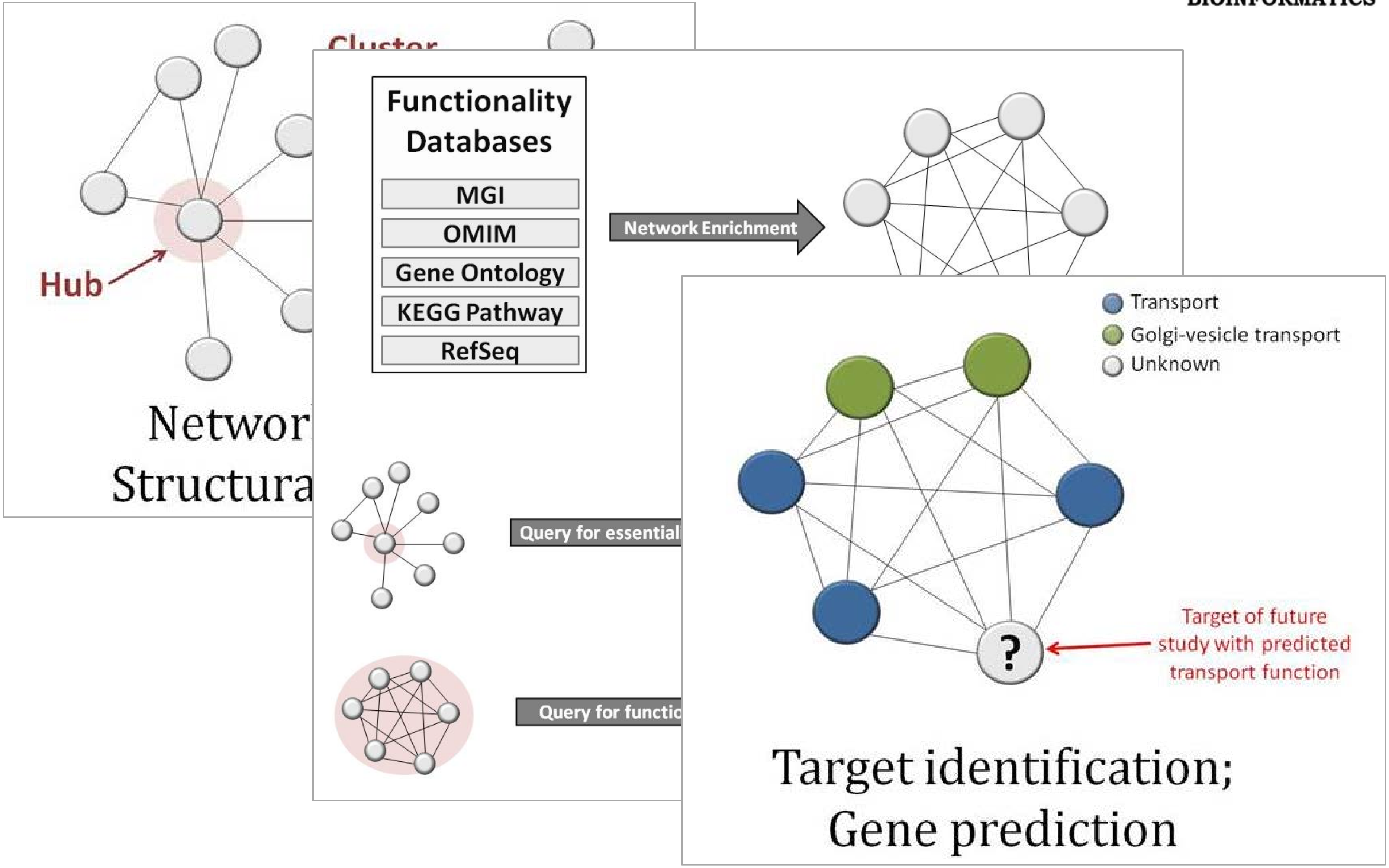
- Model for high-throughput data
  - Especially useful in microarrays

- Identification of groups of causative genes
  - Ability to rank based on graph structure
  - Identify sets of co-regulated, co-expressed genes

# Power of Correlation Analysis

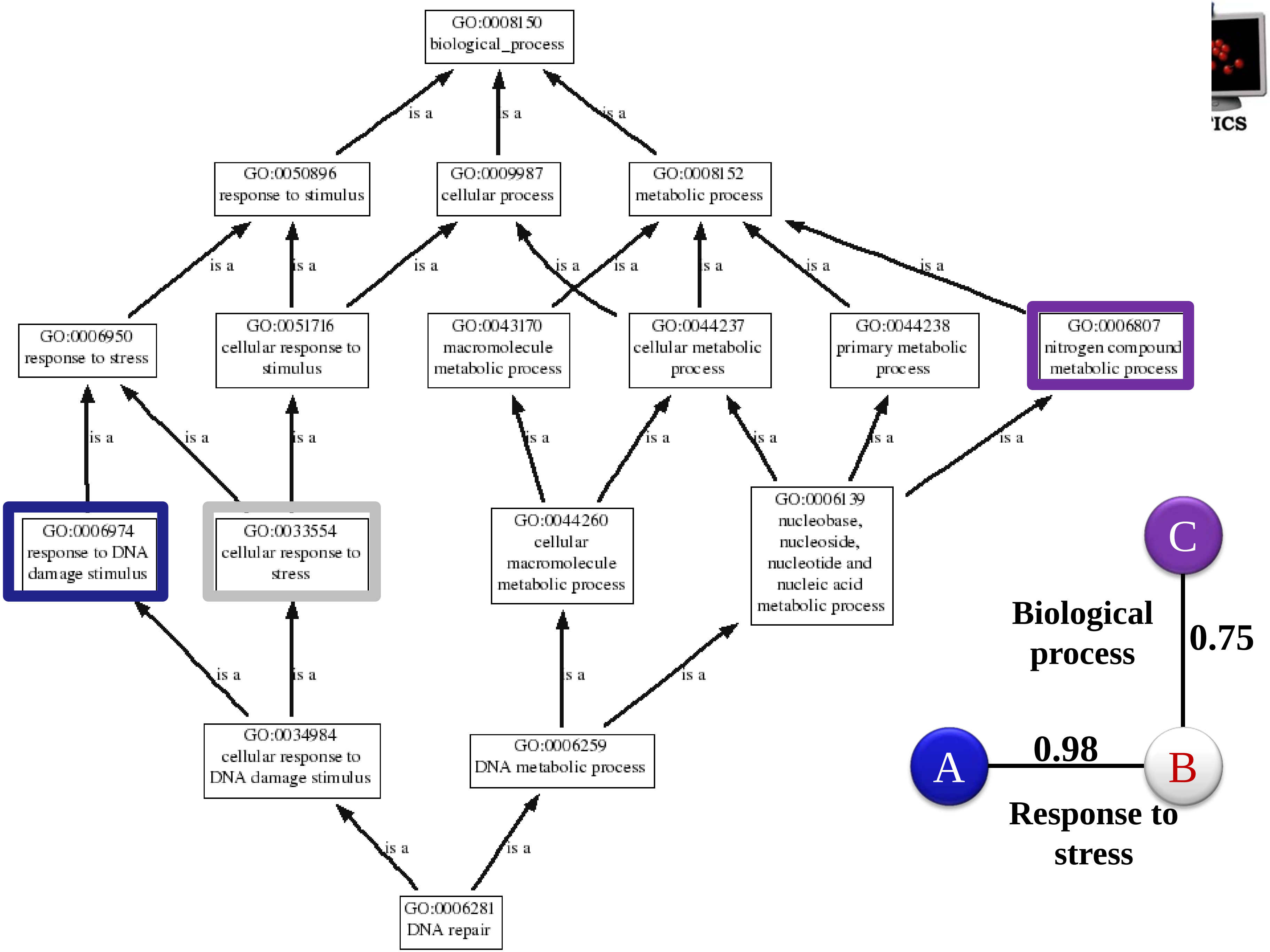
- Correlation versus Causation
- Correlation networks
- Casting the net wide – signal and noise
- The use of enrichment before obtaining information and after for validation

# Integrated Data Model



Target of future study with predicted transport function

Target identification; Gene prediction


GO:0008150
biological_process
is a
is a
is a
GO:0050896
response to stimulus
GO:0009987
cellular process
GO:0008152
metabolic process
is a
is a
is a
is a
is a
is a
is a
is a
GO:0006950
response to stress
GO:0051716
cellular response to stimulus
GO:0043170
macromolecule metabolic process
GO:0044237
cellular metabolic process
GO:0044238
primary metabolic process
GO:0006807
nitrogen compound metabolic process
is a
is a
is a
is a
is a
is a
is a
is a
GO:0006974
response to DNA damage stimulus
GO:0033554
cellular response to stress
GO:0044260
cellular macromolecule metabolic process
GO:0006139
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
is a
is a
is a
is a
GO:0034984
cellular response to DNA damage stimulus
GO:0006259
DNA metabolic process
is a
is a
GO:0006281
DNA repair
C
Biological process
0.75
A
0.98
B
Response to stress

# Tutorial Outlines

- Biomedical Informatics - Challenges and Opportunities
- Next Generation Bioinformatics Tools – Focus on Data Analysis and Integration
- Systems Biology and Network Analysis
- *Biomedical Informatics and the Cloud: Models and Security*
- Case Studies of Discoveries using Biological Networks
- HPC in Network Analysis of High Throughput Biological Data
- Integration of different aspects of Biomedical Informatics
- Next Steps – where to go from here?

# 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

# HPC and the Cloud

- Cloud for storage versus cloud for computation
  - Network analysis is now possible for everyone

- The dynamic nature of the required infrastructure
  - Outsource versus build

- Wise utilization of cloud computing facilities

# HPC and Big Data in Biological Networks

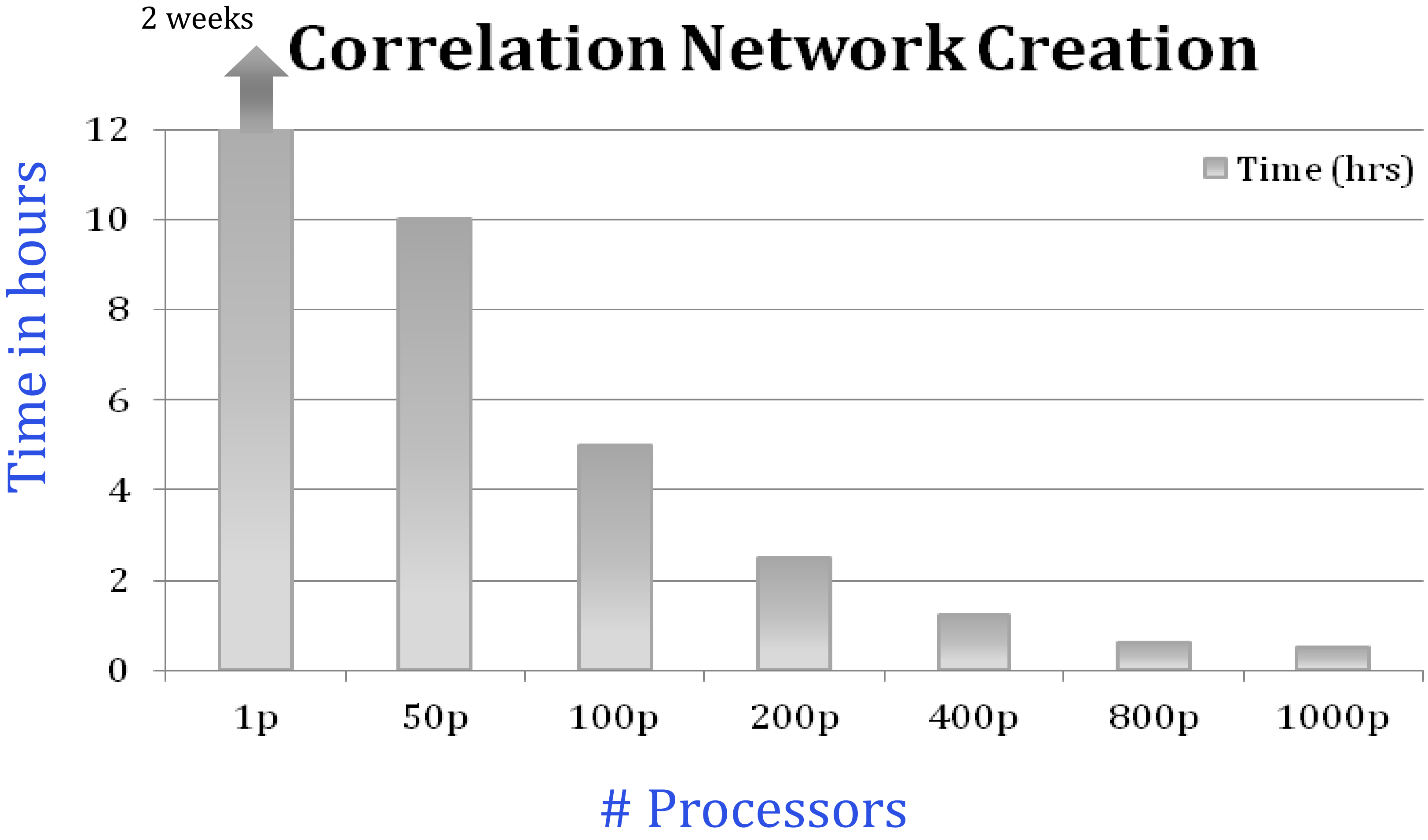
UNO BIOINFORMATICS

- 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

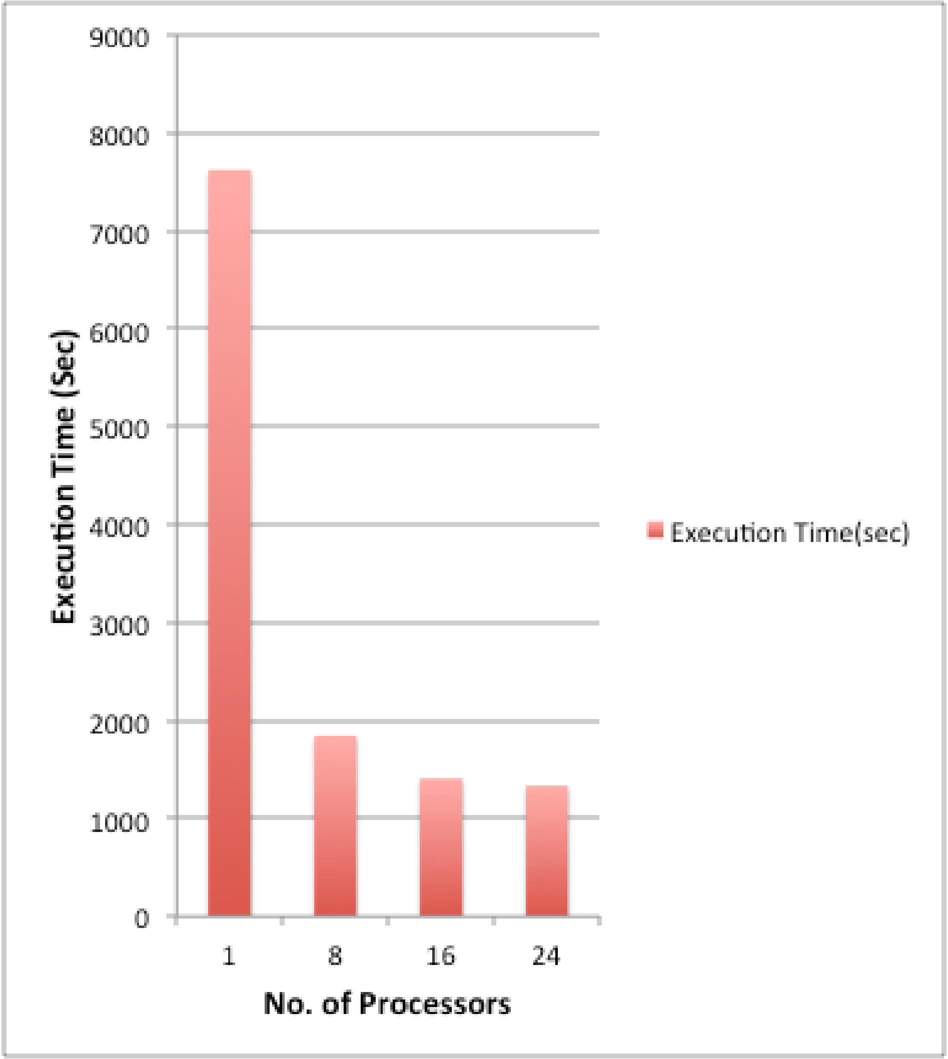
- Holland Computing Center: Firefly 1150 8-core cluster – from weeks to hours/minutes

# The Need for HPC



# Differentially Expressed Genes Analysis

- Implement baySeq method to find differentially expressed genes
- Input: 19K genes, HIV Infected and Uninfected RNA seq counts
- Significantly faster when processors are added. After 16 processors, not much of improvement



# mpiBLAST

- Parallel implementation of NCBI BLAST

- Option1: Query Segmentation
  - One query can have many sequences
  - mpiBLAST divides the input query to send to separate processors
  - Small input means faster result

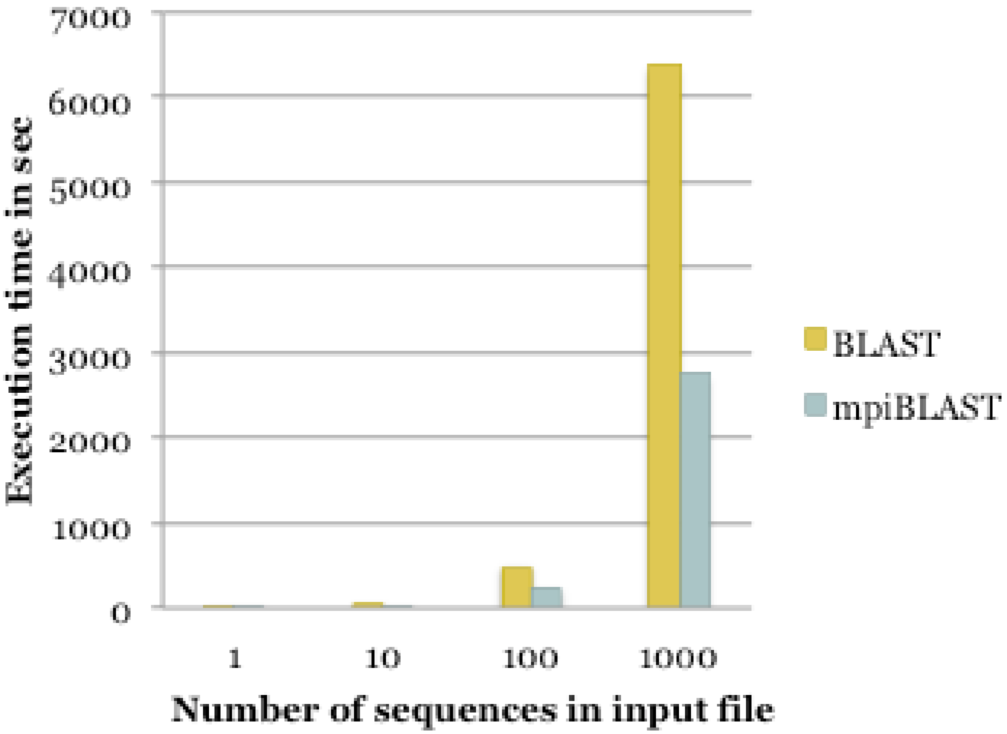
- Option2: Database Fragmentation
  - Database is created using separate method to produce ‘n’ fragments of database
  - Each processor gets small fragment of the database ( reduces page swaps)

# Input Size vs. Execution Time

- Input Database: 1 GB sequence data for all the viruses
- For large input, mpiBLAST is significantly faster than the regular BLAST
- Parameters:

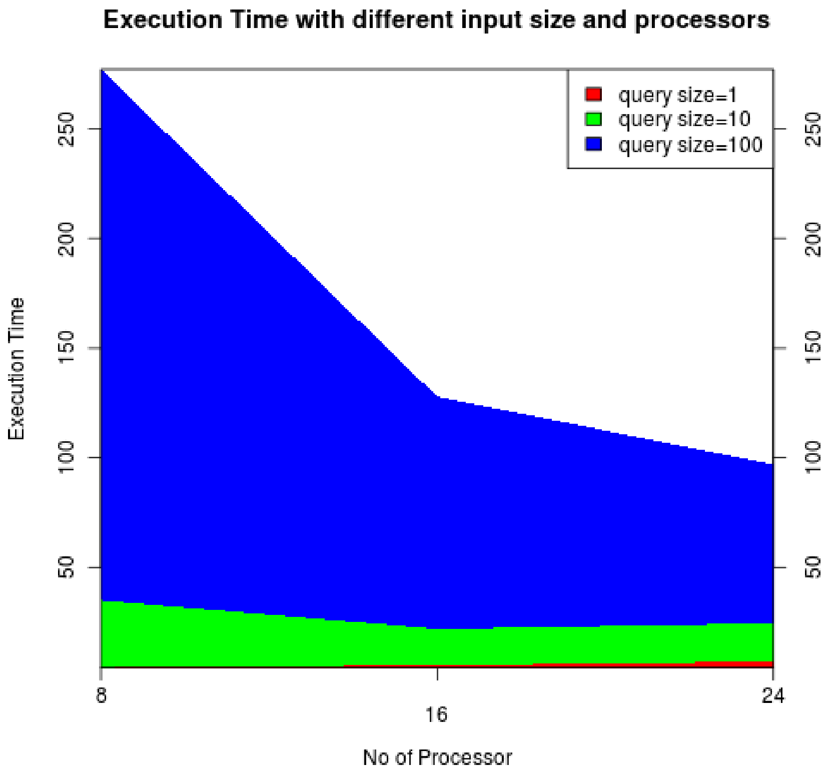
  No. of Processors = 8

  Database fragmentation = 20



# Does it scale up?

- When query size = 1:
  - Increasing Processors -> slower
    - More communication between processes
- When query size = 10:
  - Increasing from 8 to 16 processors reduced total time
  - Increasing from 16 to 24 processors increased 1 sec of execution time
- When query size = 100:
  - Increasing from 8 to 16 reduced execution time
  - Increasing from 16 to 24 only produced slight improvement
- Increasing Processors does NOT always imply faster results

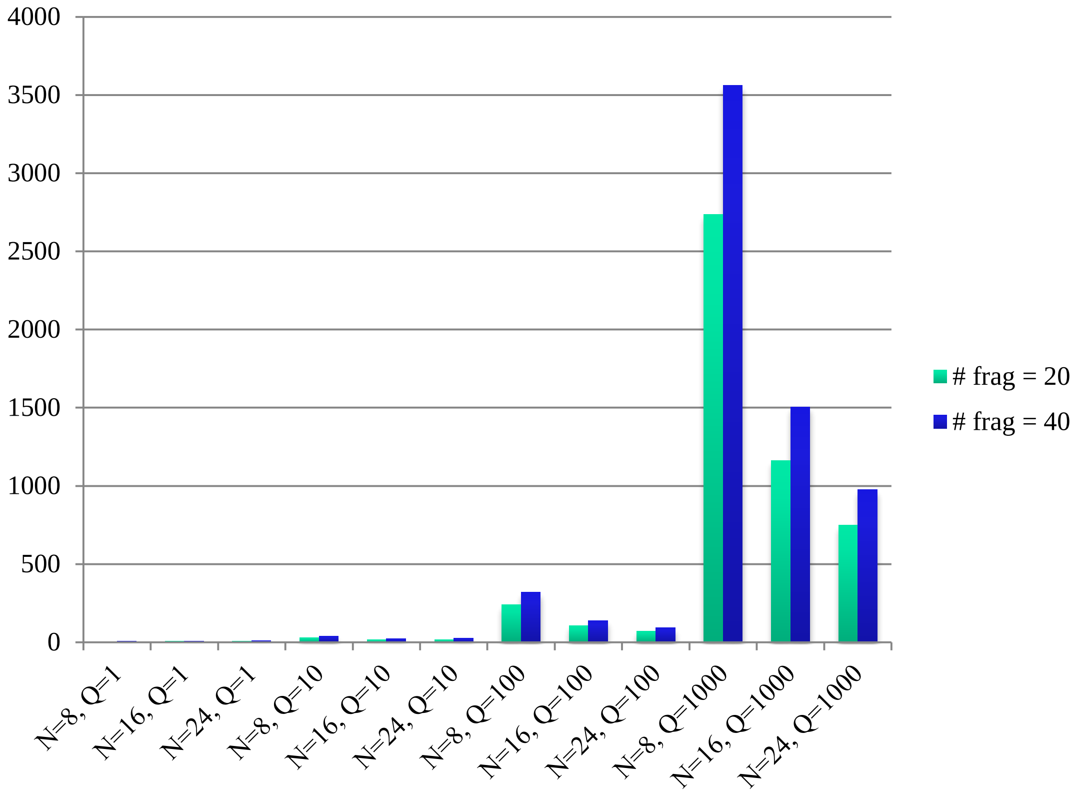


# Database Fragments vs. Execution Time

- The objective is to check if making more fragments of the database gives faster results

- higher fragmentation resulted slower performance

- N= # of processors
- Q= Input # of Sequences

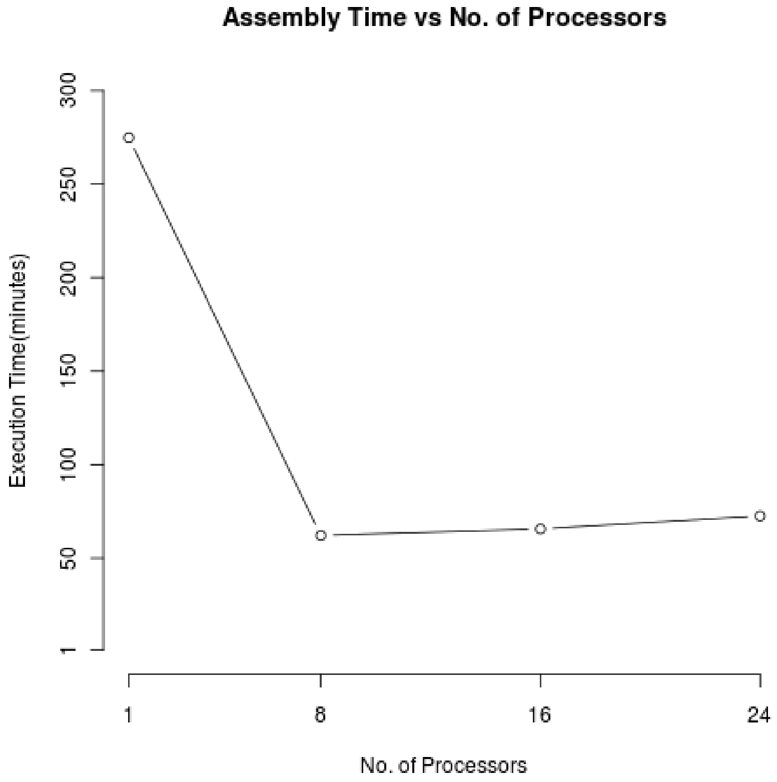


# Assembly of Next Gen Sequence Data

- High throughput sequencing generates millions to billions of reads
  - Each read can be sub-divided into k-mers
- Two main approaches:
  - de Bruijn Graphs
    - Compact representation of reads based on k-mers
    - Parallelize overlap calculations of k-mers
    - Example: Velvet, SOAPdenovo, ABySS
  - Overlap-layout consensus
    - Find overlap, build unitig based on overlap graph and then contigs based on unitig graph
    - Example: celera, edena
- Distributed joining of k-mer to give contigs

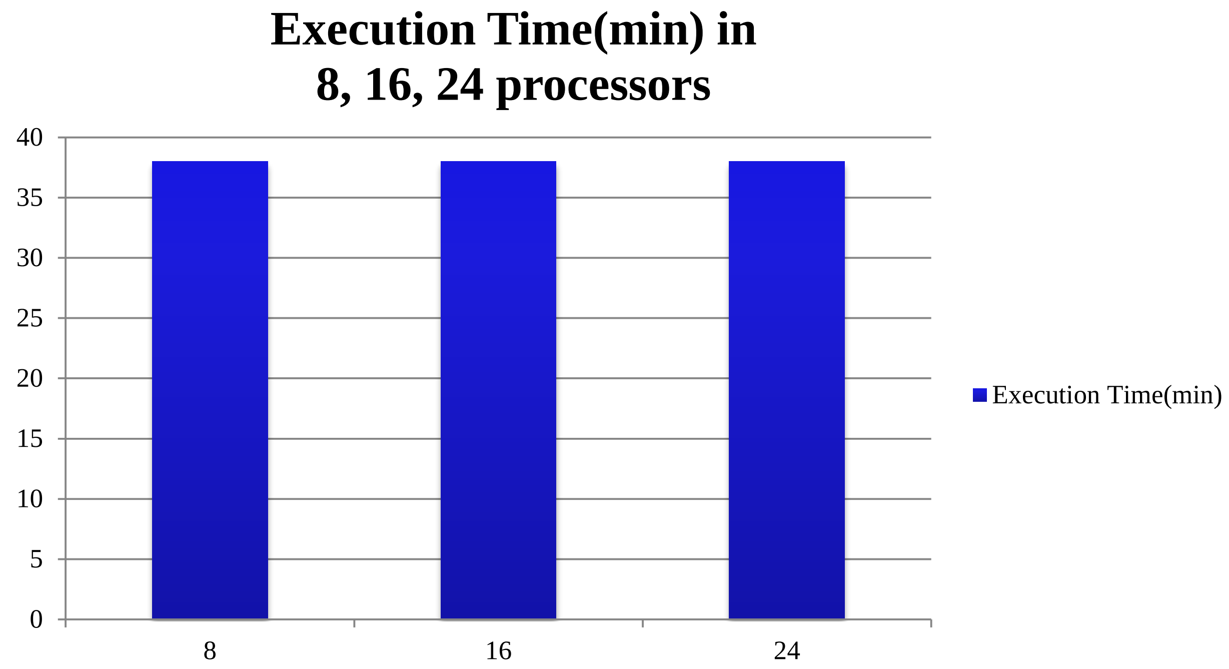
# ABySS: Speed-up Pattern

- Assembled 49 Million reads
- Great improvement from single processor to 8.
- Not much improvement adding processors from 8 to 16 or 16 to 24.



# SOAPdenovo: Speed-up Pattern

- 49 Million reads
- No. of processors 8, 16, 24
- No improvement from adding processors from 8 to 24.



# Celera, Velvet and Edena

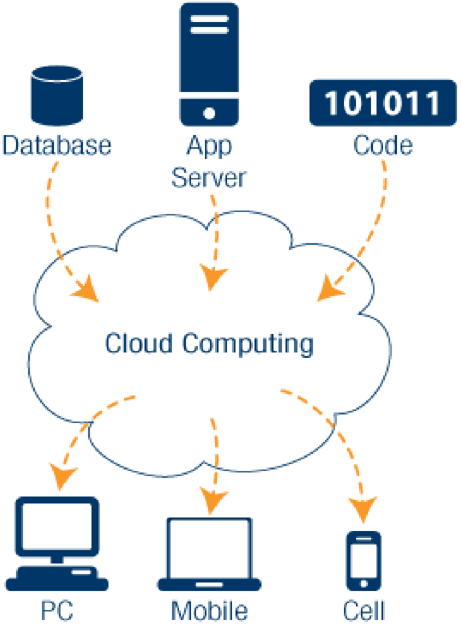
- Celera
  - Execution Time (16 threads): 3days 3 hrs and 10 mins
    - Time and space requirement is very high as compared to others, hence didn’t try with different number of threads
- Velvet and Edena are Serial-Only implementations
- Velvet: Execution Time = 2hr 9min 31sec
- Edena: Execution Time = 8hr 1min 19sec

# To Build or to Out-source?

- Utilize public resources or build a private infrastructure?
- What are the deciding factors?
  - Cost
  - Security/Privacy
  - Development ability
  - Customization
  - Nature of utilization
  - Systems administration
  - …

# Working in the Cloud

- Cloud computing is Web-based processing and storage. Software and equipment are offered as a service over the Web.
  - Data and applications can be accessed from any location
  - Data and applications can easily be shared through a common platform
  - Clouds need not be public; companies can introduce private cloud computing solutions

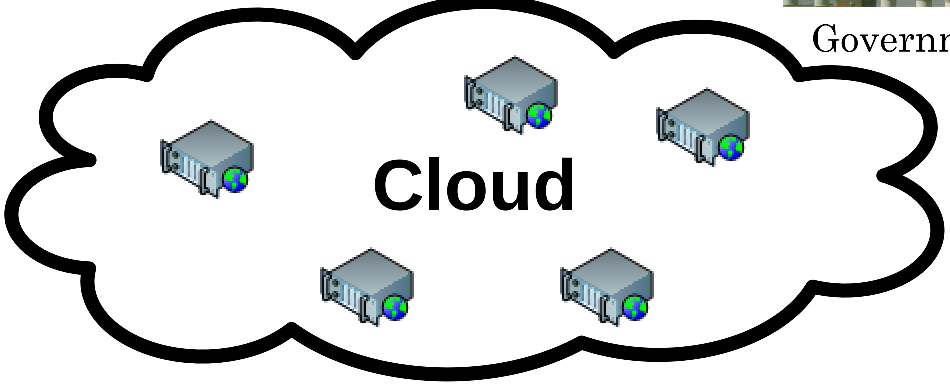


# Cost Reduction & Convenience

Small Business

Government Offices

Multinational Corporations



Homes

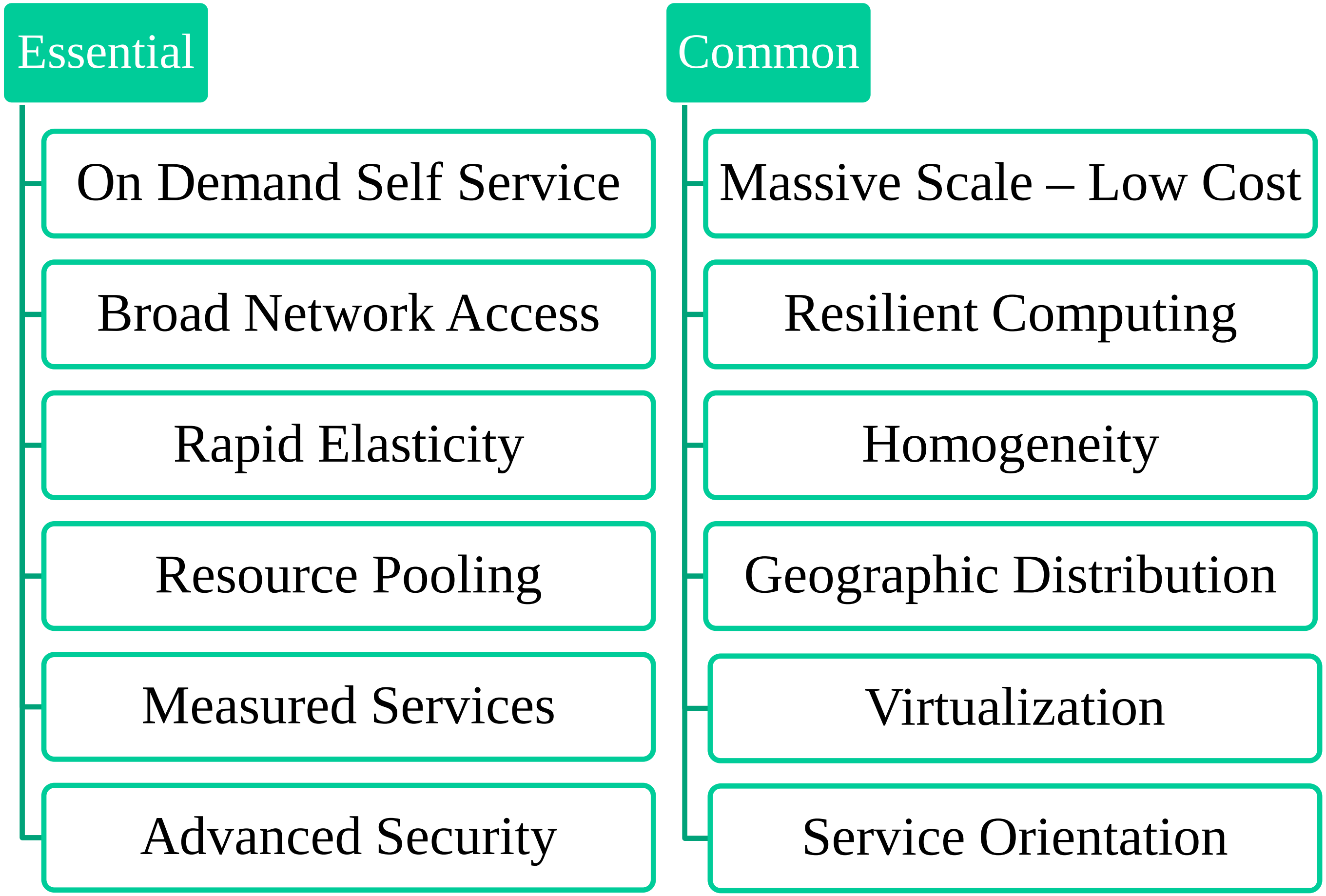
- Flexible availability of resources
- Opportunity for developers to easily push their applications
- Targeted advertising
- Easy Software Upgrades for customers
  - Example: Webmail

Cloud

# What is Cloud Computing?

- **Cloud Computing** is a general term used to describe a new class of network based computing that takes place over the Internet,
  - **Commoditised -** basically a step on from Utility Computing
  - a collection/group of integrated and networked hardware, software and Internet infrastructure (platform).
  - Using the Internet for communication and transport provides hardware, software and networking services to clients
- **Abstraction –** They hide the complexity and details of the underlying infrastructure from users and applications by providing very simple graphical interface or API.
- **Ubiquitous -** on demand services, always on, anywhere, anytime and any place.
- **Elastic** - Pay for use and as needed - scale up and down in capacity and functionalities

# Cloud Computing - Characteristics



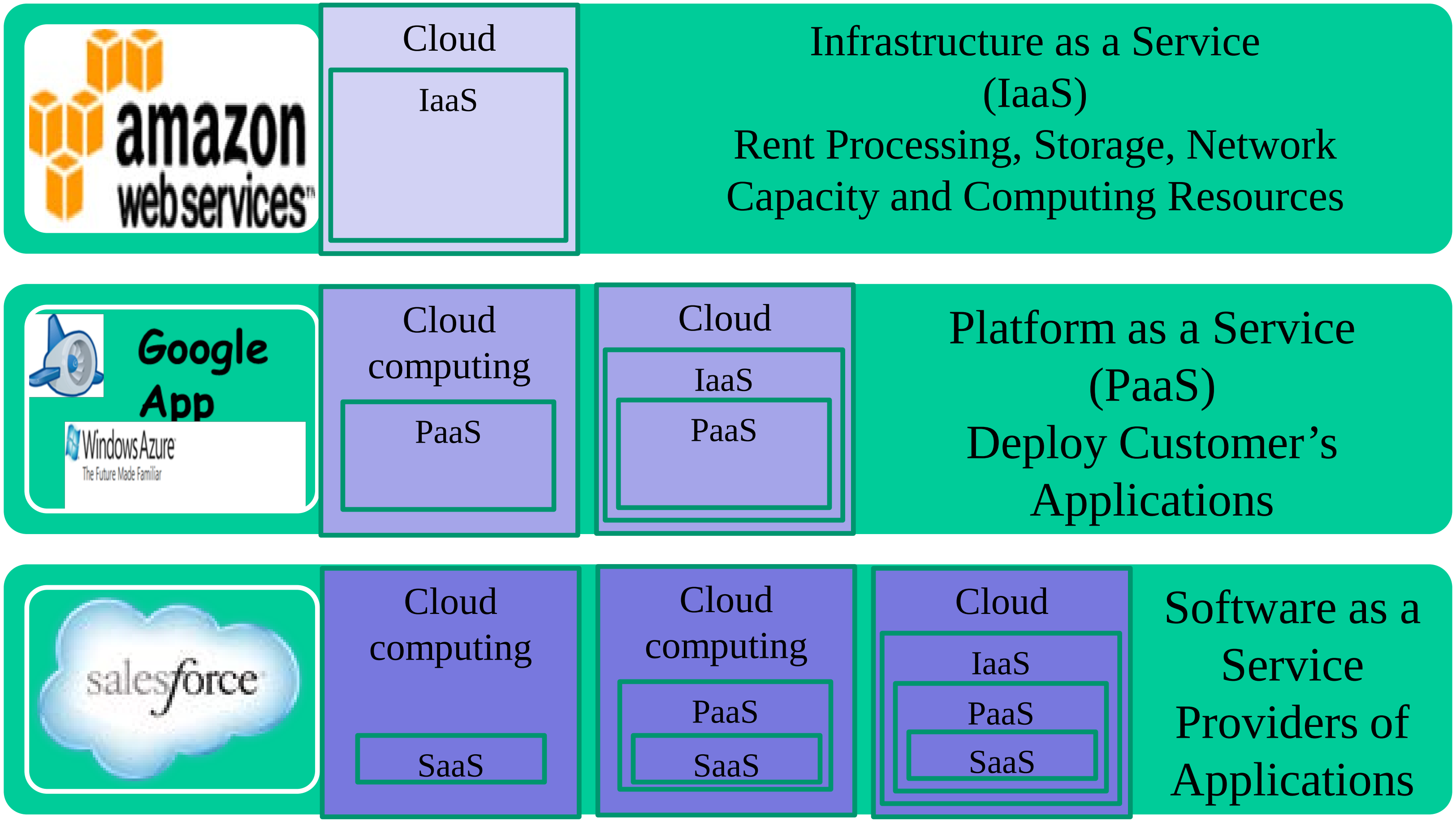
Essential

- On Demand Self Service
- Broad Network Access
- Rapid Elasticity
- Resource Pooling
- Measured Services
- Advanced Security

Common

- Massive Scale – Low Cost
- Resilient Computing
- Homogeneity
- Geographic Distribution
- Virtualization
- Service Orientation

# Cloud Computing - Models



Infrastructure as a Service
(IaaS)
Rent Processing, Storage, Network
Capacity and Computing Resources

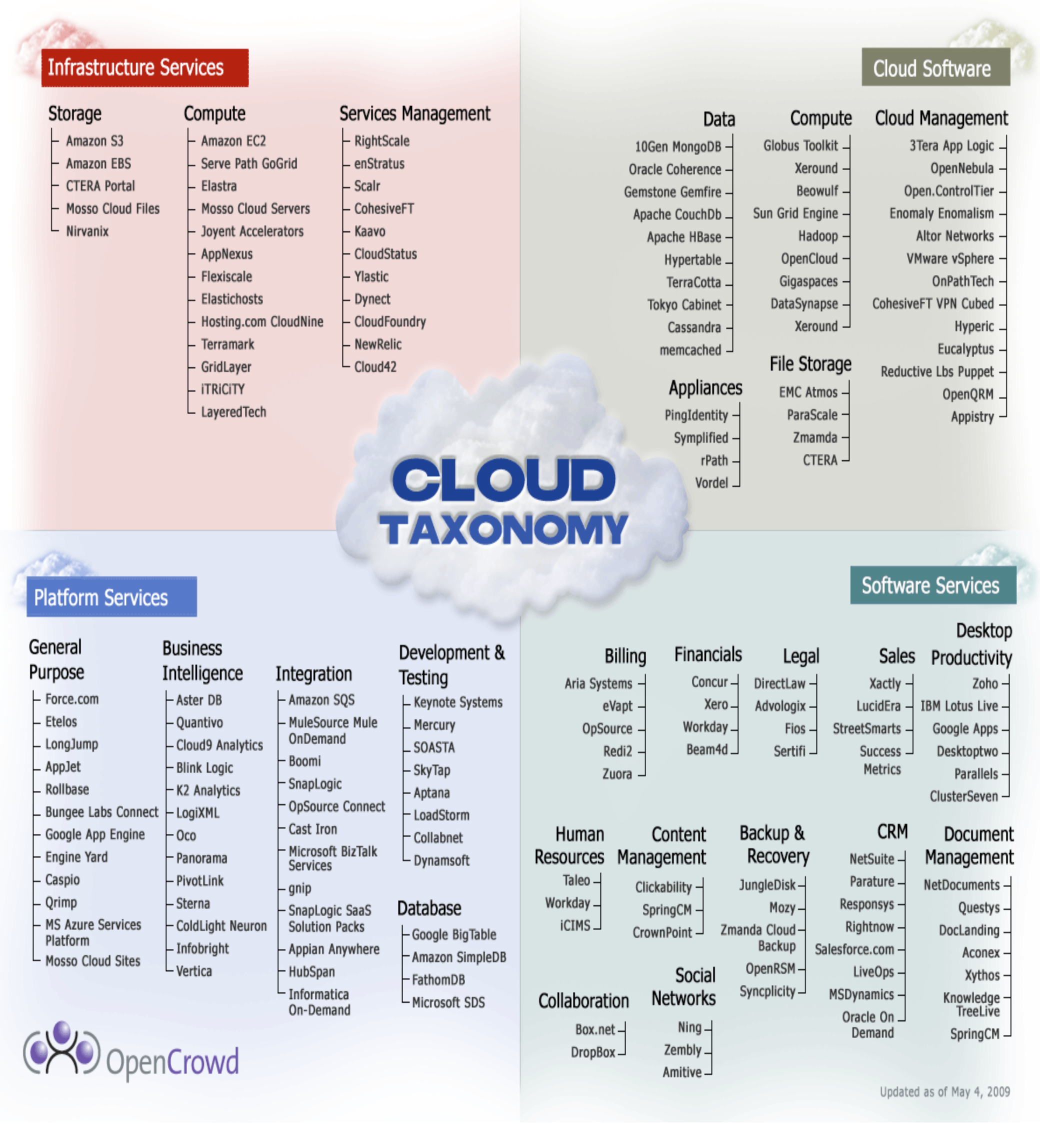
Platform as a Service
(PaaS)
Deploy Customer's
Applications

Software as a
Service
Providers of
Applications

# Cloud Computing – Service Layers

| | Services | Description |
|---|---|---|
| Application Focused | Services | Services – Complete business services such as PayPal, OpenID, OAuth, Google Maps, Alexa |
| | Application | Application – Cloud based software that eliminates the need for local installation such as Google Apps, Microsoft Online |
| | Development | Development – Software development platforms used to build custom cloud based applications (PAAS & SAAS) such as SalesForce |
| Infrastructure Focused | Platform | Platform – Cloud based platforms, typically provided using virtualization, such as Amazon ECC, Sun Grid |
| | Storage | Storage – Data storage or cloud based NAS such as CTERA, iDisk, CloudNAS |
| | Hosting | Hosting – Physical data centers such as those run by IBM, HP, NaviSite, etc. |

# Cloud – Taxonomy



CLOUD TAXONOMY

## Infrastructure Services

### Storage
- Amazon S3
- Amazon EBS
- CTERA Portal
- Mosso Cloud Files
- Nirvanix

### Compute
- Amazon EC2
- Serve Path GoGrid
- Elastra
- Mosso Cloud Servers
- Joyent Accelerators
- AppNexus
- Flexiscale
- Elastichosts
- Hosting.com CloudNine
- Terramark
- GridLayer
- iTRiCiTY
- LayeredTech

### Services Management
- RightScale
- enStratus
- Scalr
- CohesiveFT
- Kaavo
- CloudStatus
- Ylastic
- Dynect
- CloudFoundry
- NewRelic
- Cloud42

## Cloud Software

### Data
- 10Gen MongoDB
- Oracle Coherence
- Gemstone Gemfire
- Apache CouchDb
- Apache HBase
- Hypertable
- TerraCotta
- Tokyo Cabinet
- Cassandra
- memcached

### Appliances
- PingIdentity
- Symplified
- rPath
- Vordel

### Compute
- Globus Toolkit
- Xeround
- Beowulf
- Sun Grid Engine
- Hadoop
- OpenCloud
- Gigaspaces
- DataSynapse
- Xeround

### File Storage
- EMC Atmos
- ParaScale
- Zmamda
- CTERA

### Cloud Management
- 3Tera App Logic
- OpenNebula
- Open.ControlTier
- Enomaly Enomalism
- Altor Networks
- VMware vSphere
- OnPathTech
- CohesiveFT VPN Cubed
- Hyperic
- Eucalyptus
- Reductive Lbs Puppet
- OpenQRM
- Appistry

## Platform Services

### General Purpose
- Force.com
- Etelos
- LongJump
- AppJet
- Rollbase
- Bungee Labs Connect
- Google App Engine
- Engine Yard
- Caspio
- Qrimp
- MS Azure Services Platform
- Mosso Cloud Sites

### Business Intelligence
- Aster DB
- Quantivo
- Cloud9 Analytics
- Blink Logic
- K2 Analytics
- LogiXML
- Oco
- Panorama
- PivotLink
- Sterna
- ColdLight Neuron
- Infobright
- Vertica

### Integration
- Amazon SQS
- MuleSource Mule OnDemand
- Boomi
- SnapLogic
- OpSource Connect
- Cast Iron
- Microsoft BizTalk Services
- gnip
- SnapLogic SaaS Solution Packs
- Appian Anywhere
- HubSpan
- Informatica On-Demand

### Development & Testing
- Keynote Systems
- Mercury
- SOASTA
- SkyTap
- Aptana
- LoadStorm
- Collabnet
- Dynamsoft

### Database
- Google BigTable
- Amazon SimpleDB
- FathomDB
- Microsoft SDS

OpenCrowd

## Software Services

### Billing
- Aria Systems
- eVapt
- OpSource
- Redi2
- Zuora

### Financials
- Concur
- Xero
- Workday
- Beam4d

### Legal
- DirectLaw
- Advologix
- Fios
- Sertifi

### Sales
- Xactly
- LucidEra
- StreetSmarts
- Success Metrics

### Desktop Productivity
- Zoho
- IBM Lotus Live
- Google Apps
- Desktoptwo
- Parallels
- ClusterSeven

### Human Resources
- Taleo
- Workday
- iCIMS

### Content Management
- Clickability
- SpringCM
- CrownPoint

### Backup & Recovery
- JungleDisk
- Mozy
- Zmanda Cloud Backup
- OpenRSM
- Syncplicity

### CRM
- NetSuite
- Parature
- Responsys
- Rightnow
- Salesforce.com
- LiveOps
- MSDynamics
- Oracle On Demand

### Document Management
- NetDocuments
- Questys
- DocLanding
- Aconex
- Xythos
- Knowledge TreeLive
- SpringCM

### Collaboration
- Box.net
- DropBox

### Social Networks
- Ning
- Zembly
- Amitive

Updated as of May 4, 2009

# Cloud - Opportunities and Challenges

- Opportunities:
  - It enables services to be used without any understanding of their infrastructure.
  - Cloud computing works using economies of scale:
  - Cost would be by on-demand pricing.
  - Data and services are stored remotely but accessible from “anywhere”.

- Challenges:
  - Use of cloud computing means dependence on others and that could possibly limit flexibility and innovation:
  - Security could prove to be a big issue:
    - It is still unclear how safe out-sourced data is and when using these services ownership of data is not always clear.
  - There are also issues relating to policy and access:
    - If your data is stored abroad whose policy do you adhere to?
    - What happens if the remote server goes down?
    - There have been cases of users being locked out of accounts and losing access to data.

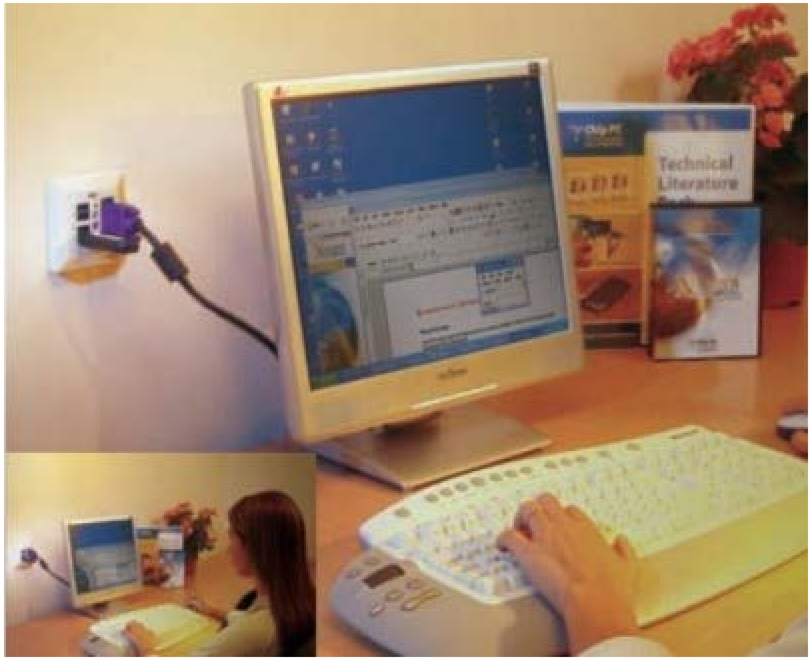
# Cloud - Opportunities and Challenges

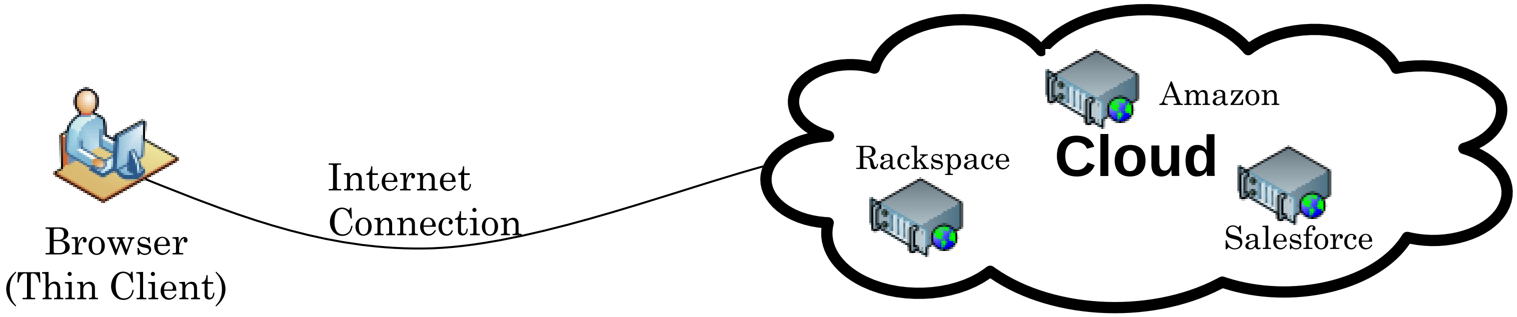
- HPC Systems:
  - Not clear that you can run compute-intensive HPC applications that use MPI/OpenMPI in domain's such as such as Bio-Informatics, Healthcare, etc.
  - Scheduling is important with this type of application
    - As you want all the VM to be co-located to minimize communication latency!
  - How do you minimize energy costs to keep costs down.
- General Concerns:
  - Each cloud systems uses different protocols and different APIs - may not be possible to run applications between cloud based systems.
  - Many new open source systems appearing that you can install and run on your local cluster - should be able to run a variety of applications on these systems

# The Future of Thin Clients

Jack PC with US Frame

Jack PC with EU Frame





Browser (Thin Client)

Internet Connection

Cloud

Amazon

Rackspace

Salesforce

# Cloud Security Challenges

- Who controls the encryption/decryption keys?
  - Customer / cloud vendor
- Storage services provided by one vendor may be incompatible with another vendor’s services should you decide to switch vendors
  - Example: Amazon’s Simple Storage Service (S3) is incompatible with IBM’s Blue Cloud, or Google, or Dell
- Customers want:
  - SSL both ways across the Internet
  - Data encryption when data is at rest
  - Ideally customer must control the encryption/decryption keys

# Cloud Security Challenges

- Data Integrity – assurance that data is identically maintained during any operation (such as transfer, storage, or retrieval)
  - Consistency and correctness

- Data must change only in response to authorized transactions
  - Unfortunately, there are no common standards

# Best Practices

- Separation of special project central storage shares
- Use of front-end web-based servers to enable workflow execution within cluster.  Having these servers hosted from a Virtual Private Cloud is the way to go!
- Identify opportunities for designing embarrassingly parallel applications – simpler to code.
- Combine MPI and OpenMP to make more efficient parallel programs.

# Virtual Cloud Security

- Good old-fashioned basics:
  - Access Control (OS-level/Application-level)
  - Server/Machine firewall (OS-level)
  - Antivirus (OS-level)
  - Strong passwords (OS-level/Application-level)
  - VPN/IDS (Network-level)
- Private on-premise cloud
- Security Hardened virtual OS template
- Hypervisor security (cloud engine security)
- Encryption
  - Data In Transit (SSL/TLS, PKE)
  - Data At Rest (File system Encryption, Data Partitioning)
- Regulations:
  - HIPAA
  - HITECH
  - FIPS 140-2
- Benefits: Scalability, quicker resource allocation, shared usage

# Virtual Cloud Security Basics

- Strong passwords (OS-level/Application-level):

  Best practices in creating, storing and periodically updating passwords.

- Access Control (OS-level/Application-level):

  Control access to system or application using central Directory services like AD.

- Server/Machine firewall (OS-level):

  Only services that need to be visible from outside should be opened up and by default, everything else should be inaccessible.

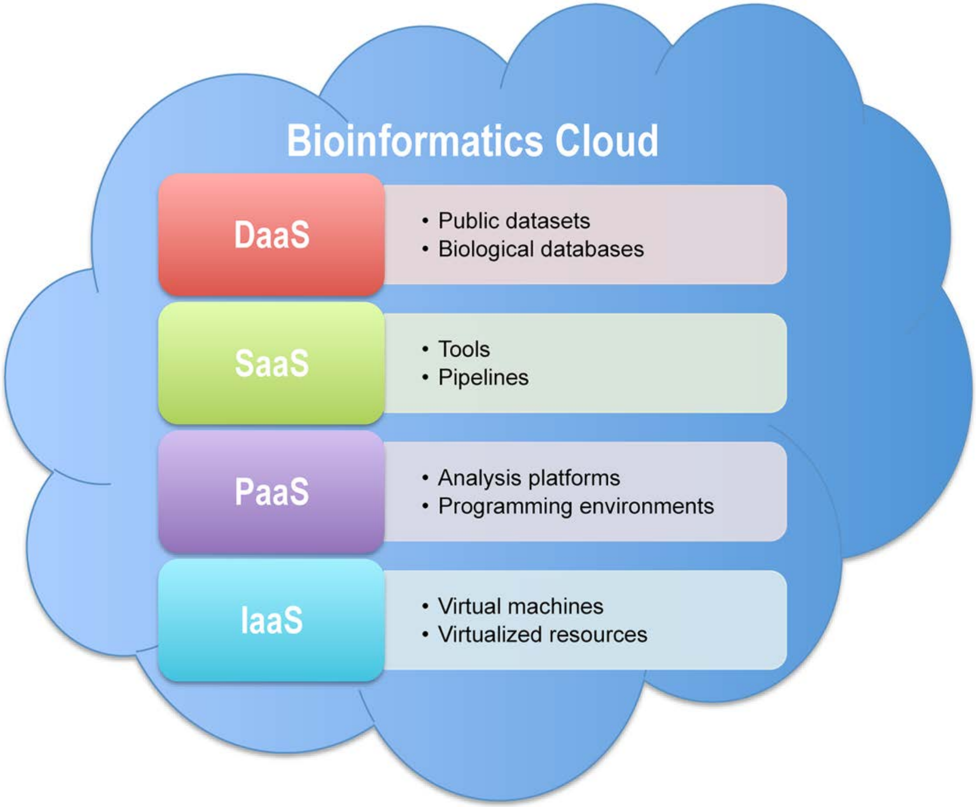
- Antivirus (OS-level):

  An integral security part of any system.

- VPN/IDS (Network-level):

  Network border-level access control using Virtual Private Network or Intrusion Detection System.

# Bioinformatics and Cloud Computing



## DaaS (Data as a Service)

- Example: AWS public data sets include GenBank, Ensembl, 1000 Genomes, etc.

## SaaS (Software as a Service)

- Sequence analysis tools (BGI easy genomics),
- Read mapping and SNP calling (Crossbow), etc.

## PaaS (Platform as a Service)

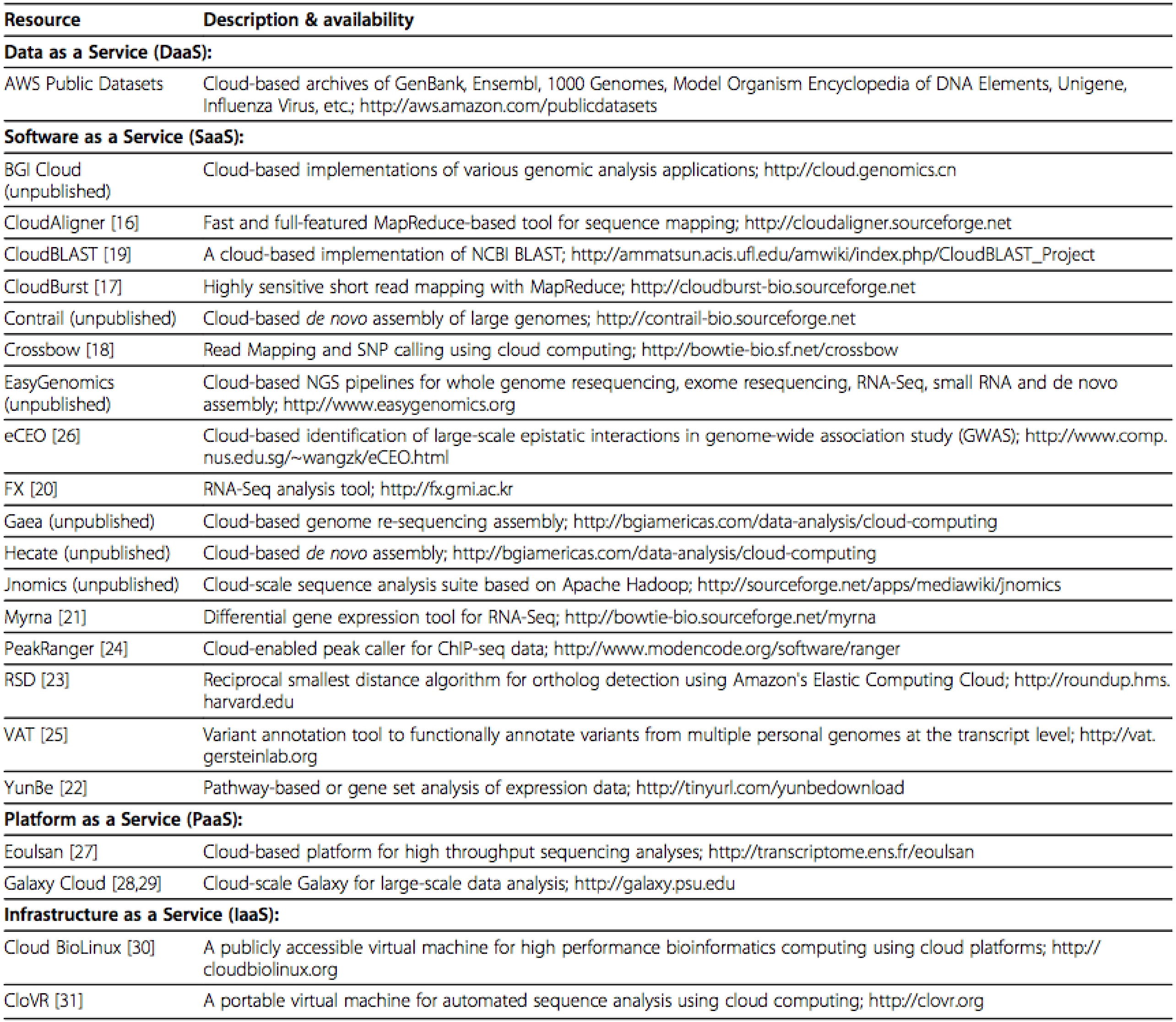
- Galaxy Cloud (Workflow platform)
- Eoulsan (RNAseq pipeline for diff. analysis)

## IaaS (Infrastructure as a Service)

- CloVR (Portable virtual machine),
- Cloud BioLinux

Dai et al. Bioinformatics clouds for big data manipulation, Biology Direct 2012

# Cloud Resources in Bioinformatics

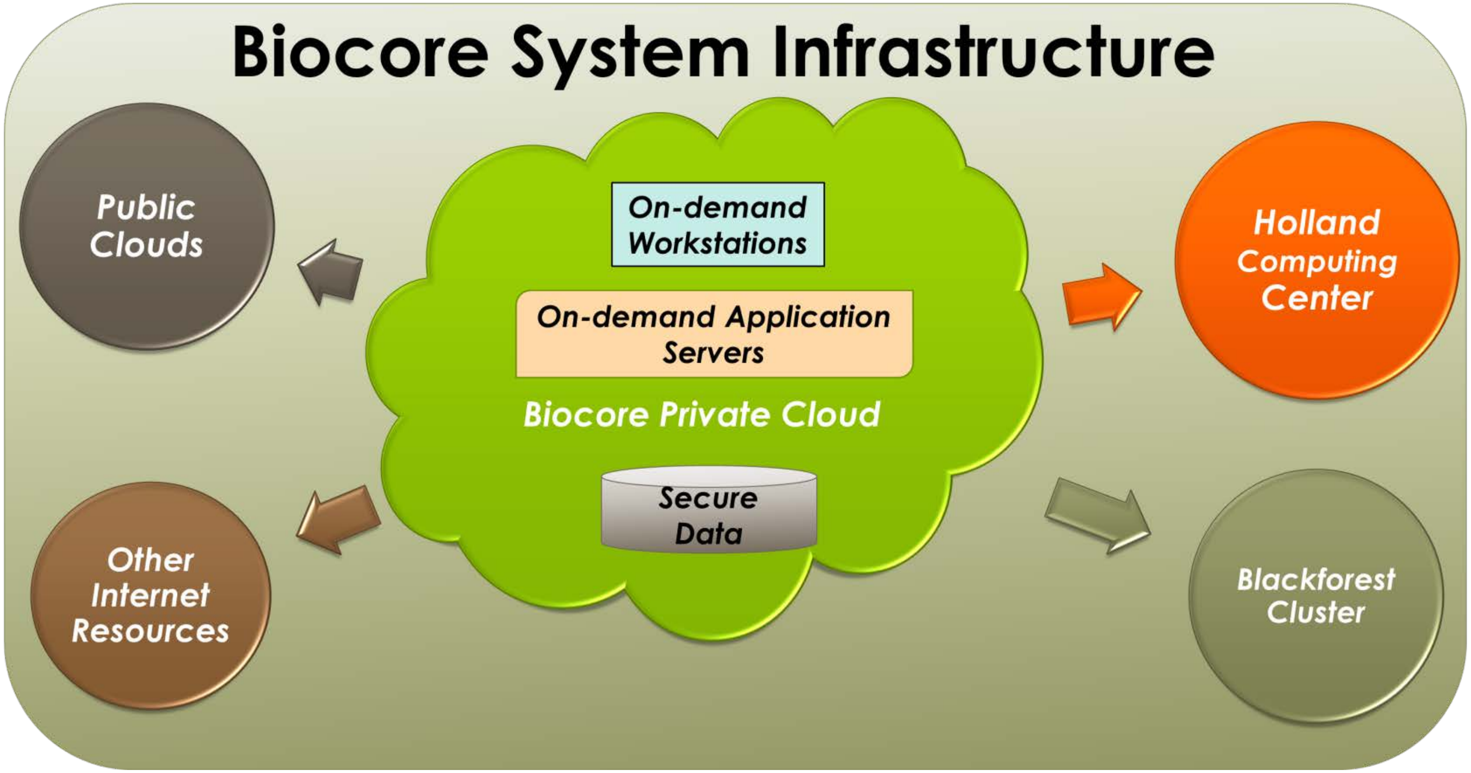


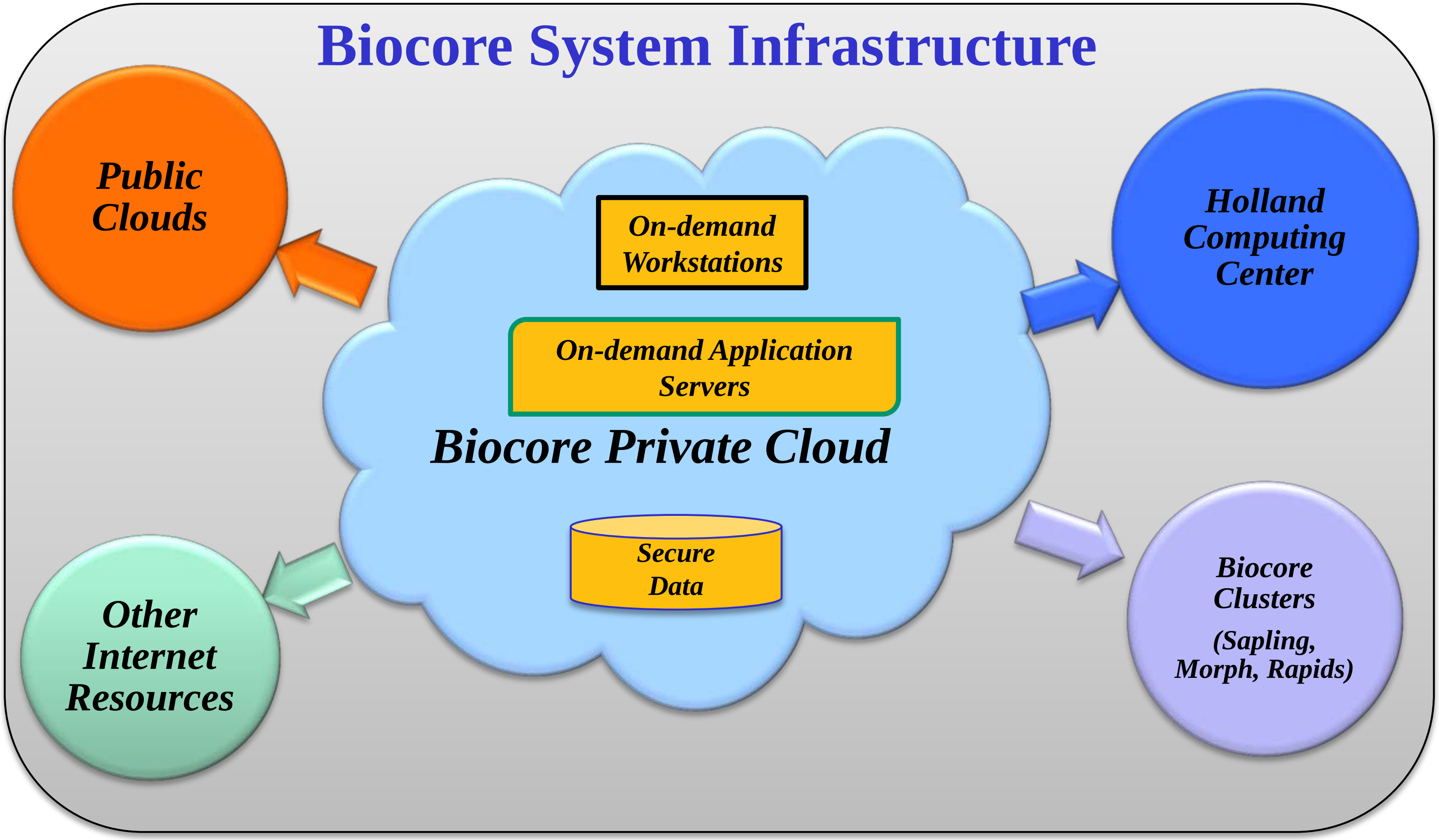
| Resource | Description & availability |
|---|---|
| **Data as a Service (DaaS):** | |
| AWS Public Datasets | Cloud-based archives of GenBank, Ensembl, 1000 Genomes, Model Organism Encyclopedia of DNA Elements, Unigene, Influenza Virus, etc.; http://aws.amazon.com/publicdatasets |
| **Software as a Service (SaaS):** | |
| BGI Cloud (unpublished) | Cloud-based implementations of various genomic analysis applications; http://cloud.genomics.cn |
| CloudAligner [16] | Fast and full-featured MapReduce-based tool for sequence mapping; http://cloudaligner.sourceforge.net |
| CloudBLAST [19] | A cloud-based implementation of NCBI BLAST; http://ammatsun.acis.ufl.edu/amwiki/index.php/CloudBLAST_Project |
| CloudBurst [17] | Highly sensitive short read mapping with MapReduce; http://cloudburst-bio.sourceforge.net |
| Contrail (unpublished) | Cloud-based *de novo* assembly of large genomes; http://contrail-bio.sourceforge.net |
| Crossbow [18] | Read Mapping and SNP calling using cloud computing; http://bowtie-bio.sf.net/crossbow |
| EasyGenomics (unpublished) | Cloud-based NGS pipelines for whole genome resequencing, exome resequencing, RNA-Seq, small RNA and de novo assembly; http://www.easygenomics.org |
| eCEO [26] | Cloud-based identification of large-scale epistatic interactions in genome-wide association study (GWAS); http://www.comp.nus.edu.sg/~wangzk/eCEO.html |
| FX [20] | RNA-Seq analysis tool; http://fx.gmi.ac.kr |
| Gaea (unpublished) | Cloud-based genome re-sequencing assembly; http://bgiamericas.com/data-analysis/cloud-computing |
| Hecate (unpublished) | Cloud-based *de novo* assembly; http://bgiamericas.com/data-analysis/cloud-computing |
| Jnomics (unpublished) | Cloud-scale sequence analysis suite based on Apache Hadoop; http://sourceforge.net/apps/mediawiki/jnomics |
| Myrna [21] | Differential gene expression tool for RNA-Seq; http://bowtie-bio.sourceforge.net/myrna |
| PeakRanger [24] | Cloud-enabled peak caller for ChIP-seq data; http://www.modencode.org/software/ranger |
| RSD [23] | Reciprocal smallest distance algorithm for ortholog detection using Amazon's Elastic Computing Cloud; http://roundup.hms.harvard.edu |
| VAT [25] | Variant annotation tool to functionally annotate variants from multiple personal genomes at the transcript level; http://vat.gersteinlab.org |
| YunBe [22] | Pathway-based or gene set analysis of expression data; http://tinyurl.com/yunbedownload |
| **Platform as a Service (PaaS):** | |
| Eoulsan [27] | Cloud-based platform for high throughput sequencing analyses; http://transcriptome.ens.fr/eoulsan |
| Galaxy Cloud [28,29] | Cloud-scale Galaxy for large-scale data analysis; http://galaxy.psu.edu |
| **Infrastructure as a Service (IaaS):** | |
| Cloud BioLinux [30] | A publicly accessible virtual machine for high performance bioinformatics computing using cloud platforms; http://cloudbiolinux.org |
| CloVR [31] | A portable virtual machine for automated sequence analysis using cloud computing; http://clovr.org |

# Private Clouds

- Core facilities need to acquire private infrastructure-level virtual cloud technology. Best vendor for such technology is VMware. The Bioinformatics Core facility at UNO uses *VMware vSphere Enterprise.*
- Public Clouds like Amazon EC2, RackSpace cannot be used in all cases due to various restrictions put forth by regulations (e.g. HIPAA data locality requirement). Such public clouds could only be used as a scalable platform for already anonymized data.
- Private Virtual Cloud is on-premise solution allowing all the benefits of virtualization technology both from an administrative and end-user perspective.

# Proposed Model

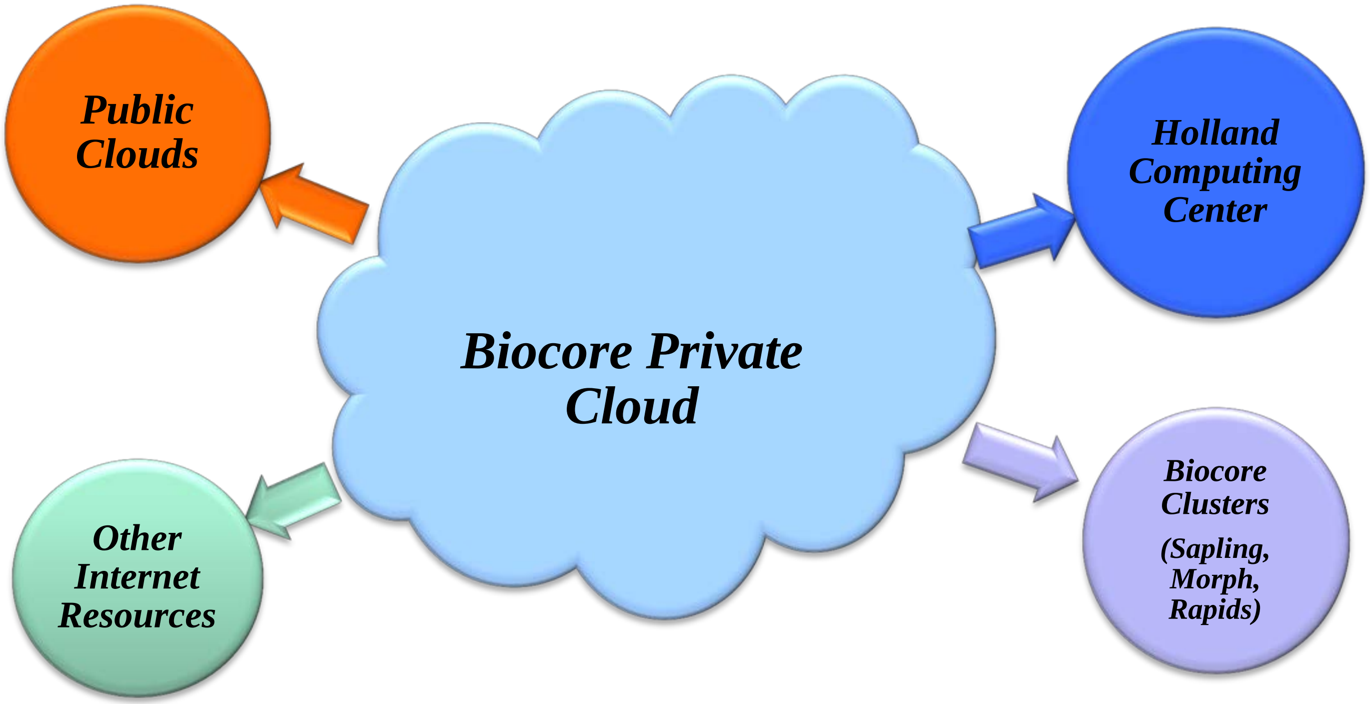




Infrastructure As A Service: Virtual workstations and servers provided to various core facilities from the Biocore private cloud

Data As A Service: Local mirrors of various public databases. (*GenBank*, *RefSeq* , *EMBL* and more…).Shared storage for various core facilities in the form of network drives

HIPAA compliance: (e.g. the **R**heumatoid **A**rthritis **I**nfrastructure **N**etwork website)

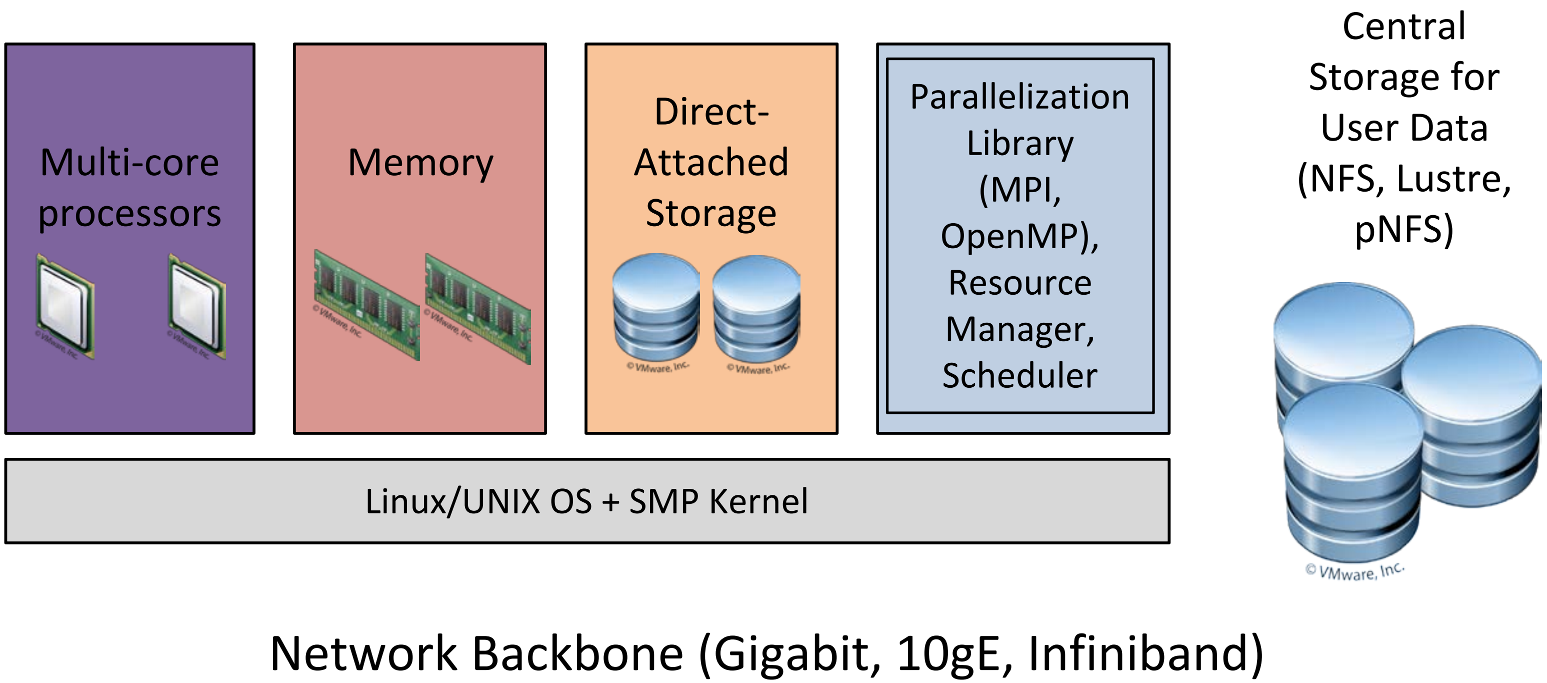


Software As A Service: SOAP web service for BioCatalogue plugin, other licensed software like VectorNTI, MATLAB

Platform As A Service: Galaxy Server, BIOCMS for media sharing

Middleware Service: Setup project servers in our private cloud that will interact with the HCC and Biocore clusters, or with public clouds (e.g. the Amazon EC2 Cloud) for scalability, and run custom-built high performance computing applications

# HPC – at UNO



Multi-core processors

Memory

Direct-Attached Storage

Parallelization Library (MPI, OpenMP), Resource Manager, Scheduler

Linux/UNIX OS + SMP Kernel

Central Storage for User Data (NFS, Lustre, pNFS)

Network Backbone (Gigabit, 10gE, Infiniband)

# Tutorial Outlines

- Biomedical Informatics - Challenges and Opportunities
- Next Generation Bioinformatics Tools – Focus on Data Analysis and Integration
- Systems Biology and Network Analysis
- Biomedical Informatics and the Cloud: Models and Security
- *Case Studies of Discoveries using Biological Networks*
- HPC in Network Analysis of High Throughput Biological Data
- Integration of different aspects of Biomedical Informatics
- Next Steps – where to go from here?

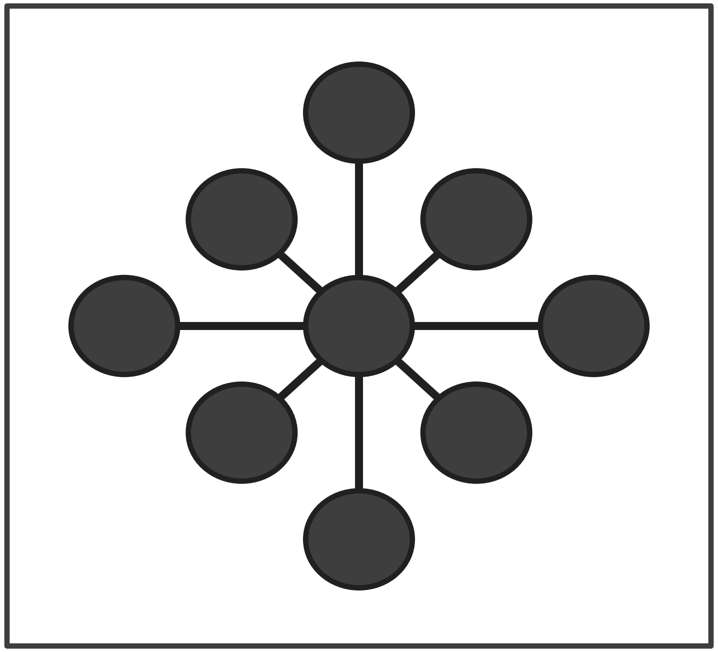
# Network Concepts

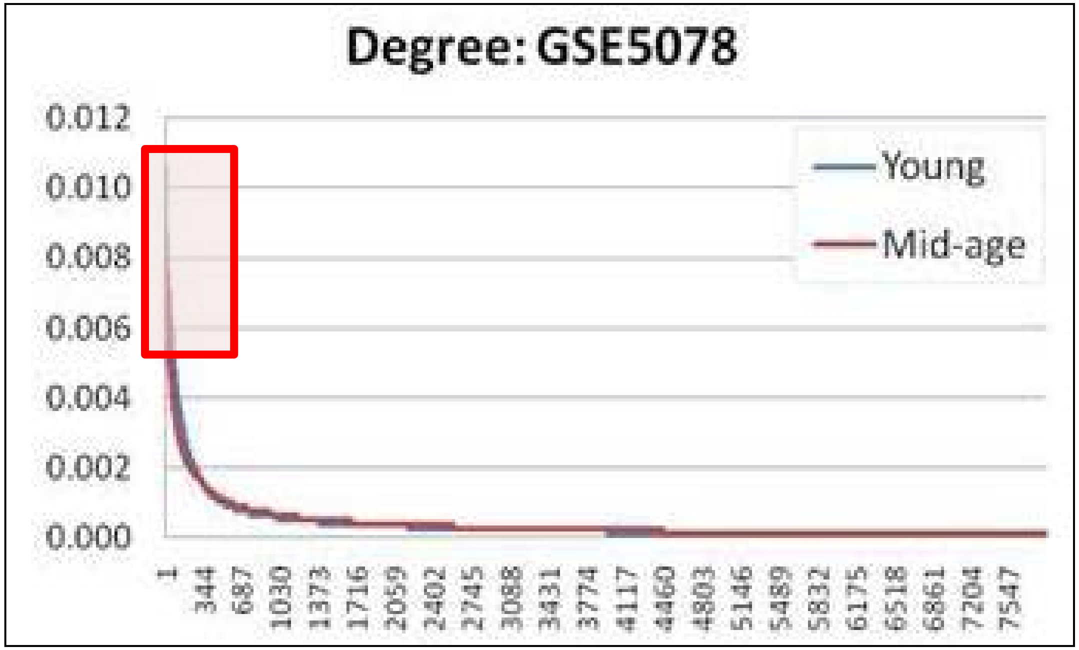
- **Biological networks have structural properties**
  - Can differ from one network to another

- **Specific structures/characteristics have biological meaning**
  - Degree can indicate essentiality
  - Cluster density can indicate relevance

- **Networks do not have to be static**
  - Most interesting discoveries coming from temporal or state-change network alignment & comparison

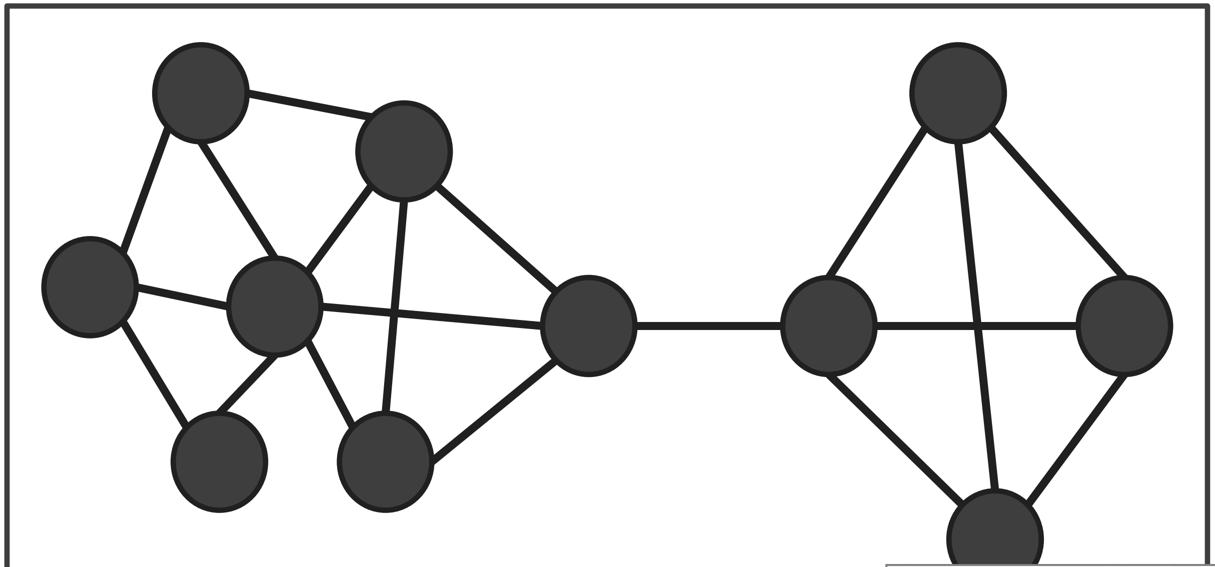
# Centrality Measures

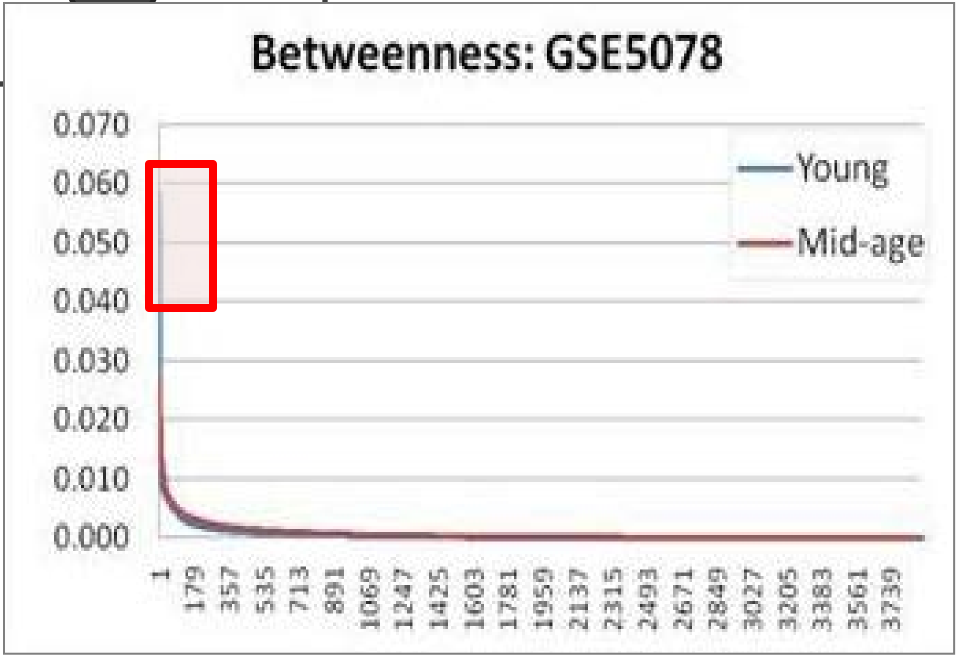




**Degree:** The number of edges a node has

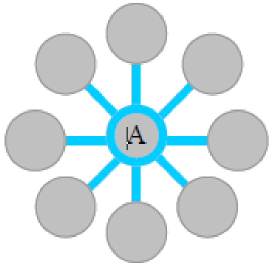
# Centrality Measures



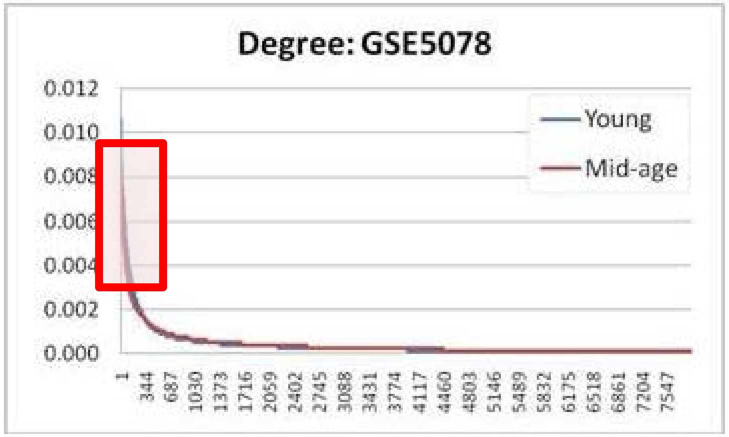


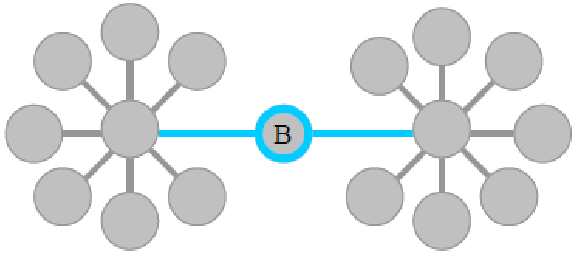
**Betweenness:**
How many shortest paths a node lies on

# Critical Nodes

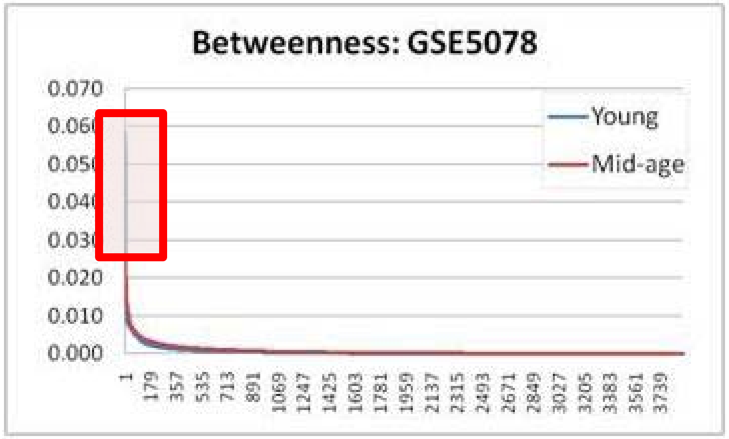


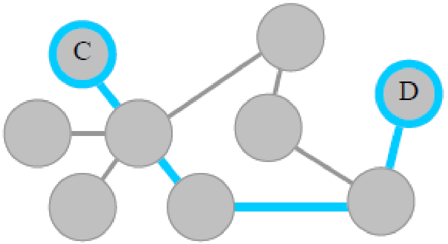
**Degree:** Number of neighbors



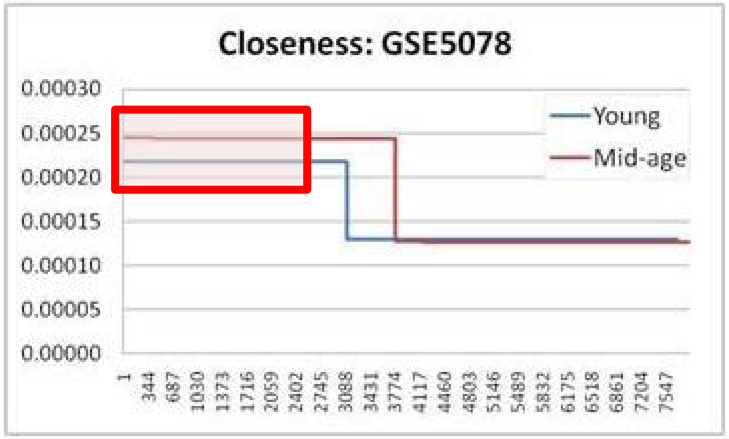


**Betweenness:** Number of shortest paths a node lies on





**Closeness:** Average shortest path from a node to all other nodes in the network



# Hypothesis

Correlation networks are an excellent tool for mining relationship rich knowledge from high-throughput data

Using systems biology approach, CN can help identify:

- *Critical Genes* that are essential for survival
- *Subsets of genes* that are responsible for biological functions

**Measures of centrality to identify key elements:**

**Proves existence of structure/fucntion relationship in correlation networks**

# Structures & their Functions

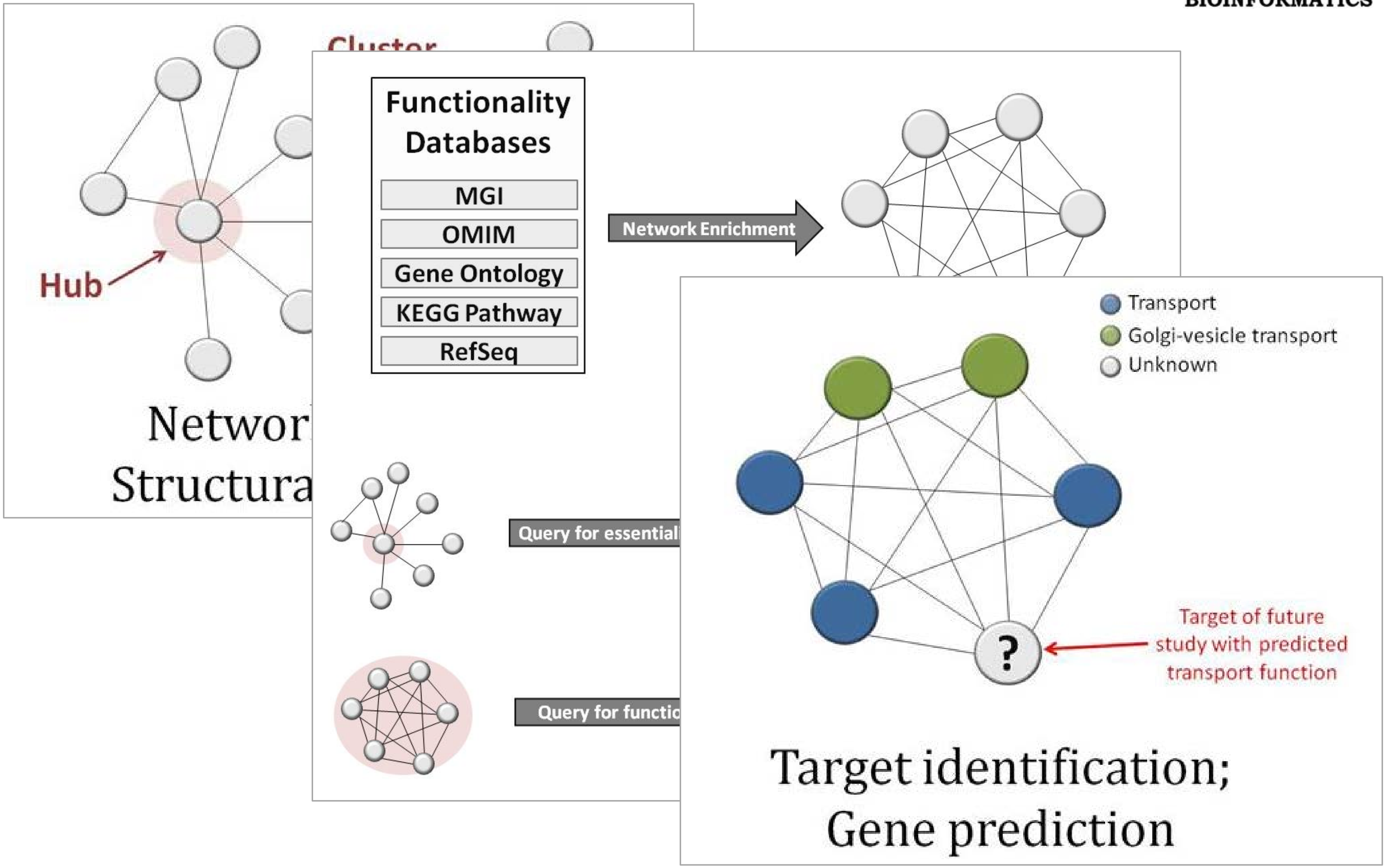
Network structures correspond to key cellular structures

# Objectives

- Confirm structure/ function relationships in integrated biological networks

- Uncover genetic drivers of aging and disease using application of graph theory

# Integrated Data Model



Cluster

Hub

Networ

Structura

Functionality Databases

MGI

OMIM

Gene Ontology

KEGG Pathway

RefSeq

Network Enrichment

Query for essential

Query for functio

● Transport

● Golgi-vesicle transport

○ Unknown

?

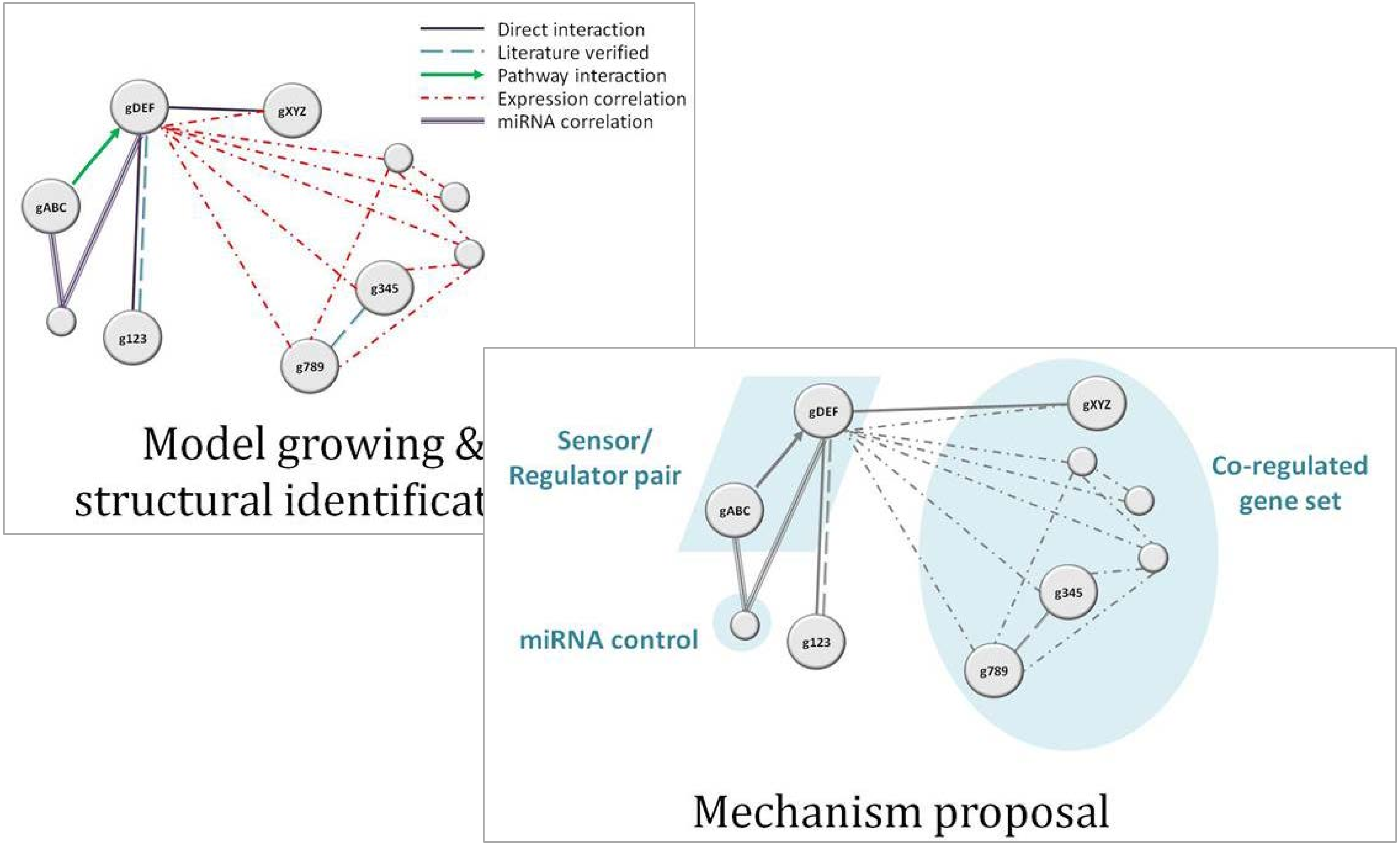
Target of future study with predicted transport function

Target identification; Gene prediction

# Network Integration

- **Network Alignment**
  - Homozygous (PPI aligned with PPI)
  - Heterozygous (Phenome aligned with transcriptome)
- **Network Combination**
  - Union, Intersection, Difference
- **Data Integration**
  - Knowledge-driven
  - Data-driven

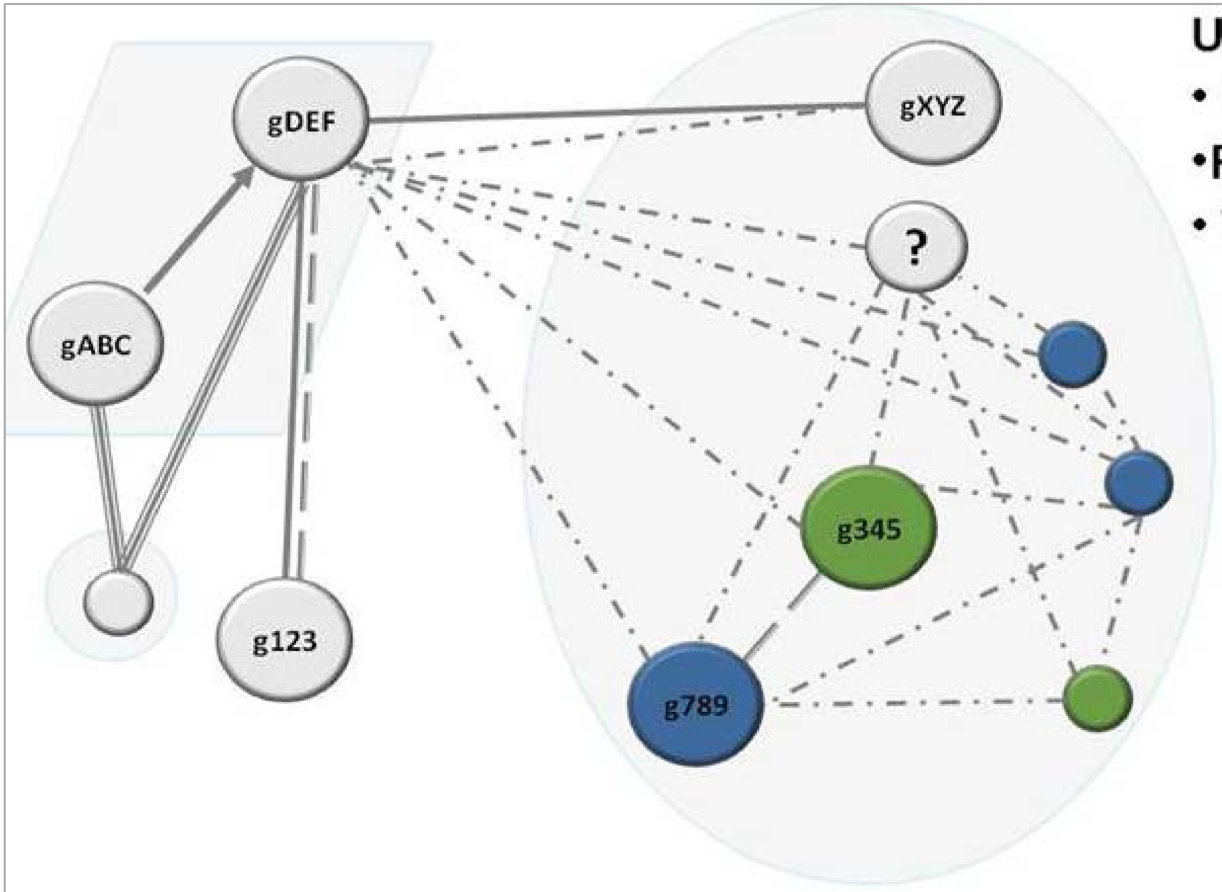
# Integration: Knowledge Model



Model growing &
structural identificat

Mechanism proposal

# Integration: Knowledge Model

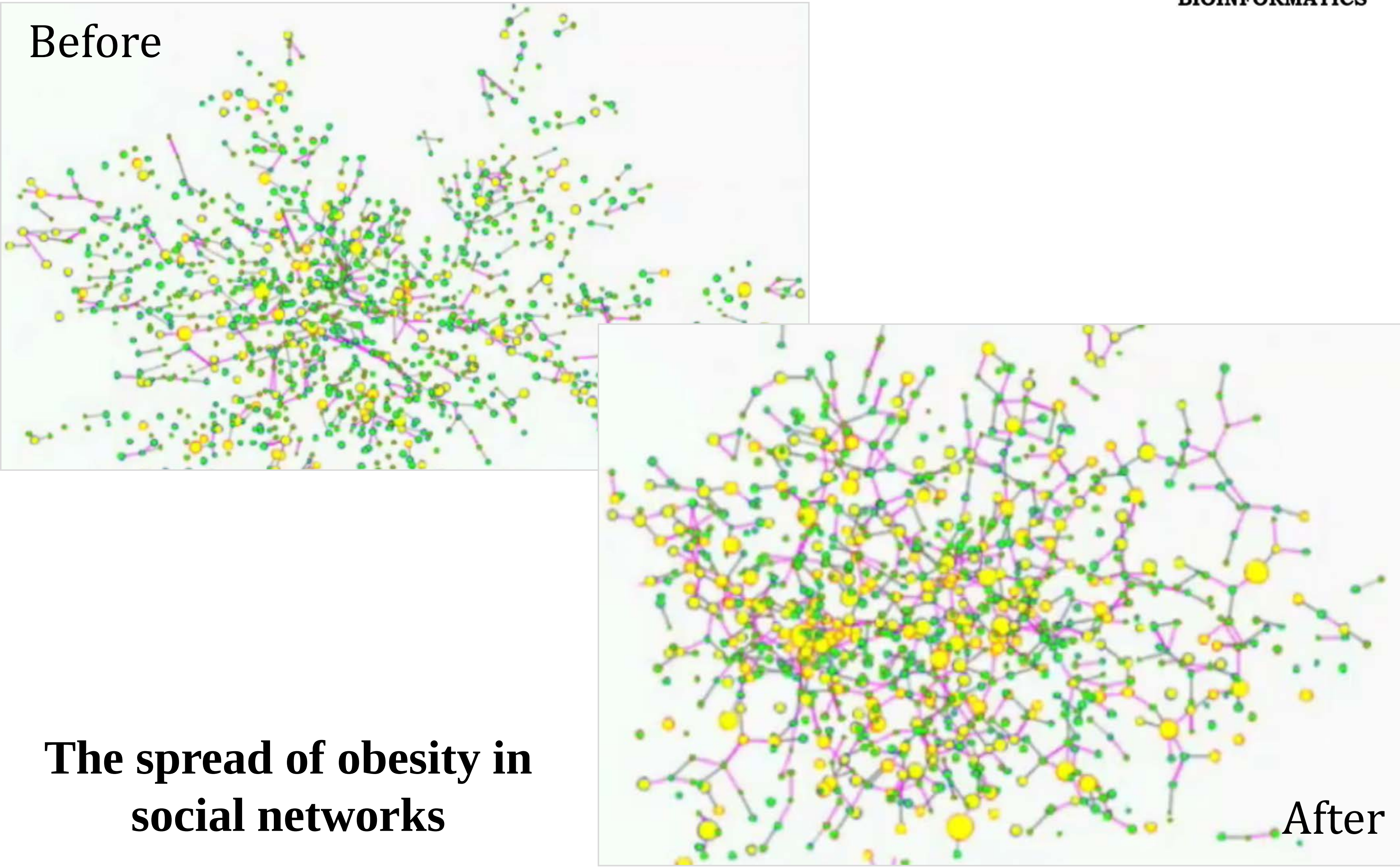


Unknown gene:

- Predicted transport function
- Regulated by gDEF/gABC
- Via potential miRNA control

Structure+Function
Integrated Model

# Other Applications: Health Informatics



**The spread of obesity in social networks**

# Aging Research: What is involved?

- Bioinformatics – Correlation Networks – Analysis of biological data – Bioinformatics Tools
- Computer Science: HPC - Wireless Networks – Database – Software development
- MIS: GUI – User Experience Factors
- Gerontology: Social Factors – Aging Research
- Medical Sciences: Impact of Medication – Impact of other Health Factors
- Engineering: Design and evaluation of Sensors
- IT Innovation: 3D Environments – Simulation and Modeling

# The Bioinformatics Angle

- 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 |

# Hub Lethality

- Young Male BalbC Mouse
  - 12/20 hubs tested for *in vivo* knockout
    - 8/12 lethal phenotype pre-/peri-natally
    - 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)

# Hub Lethality

- Young Male BalbC Mouse
  - 12/20 hubs tested for *in vivo* knockout
    - 8/12 lethal phenotype pre-/peri-natally
    - 4/12 non-lethal but system-affected:
      - Hspa1a: cellular, growth/size, homeostasis
      - Dapk1: cellular, renal/urinary
      - Ffar2: Increased susceptibility to colitis, asthma, arthritis

# Hub Lethality

- 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-affected:
      - Btn1a1: impaired lactation, impaired lipid accumulation in mammary gland
      - Bcl2l11: die later in life from auto-immune kidney disease
      - Rag2: arrested development of T and B cell maturation
    - 1/11 no observed phenotype (Aldh3a1)

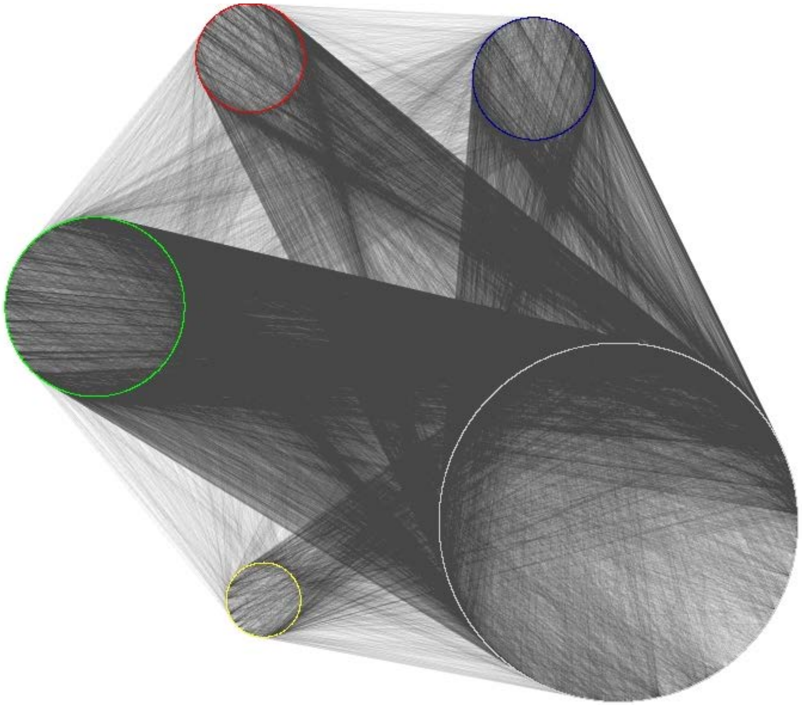
## Young Mice Top 20 Hubs

| Gene Name | Description | Lethal K/O? |
|---|---|---|
| Rad51l1 | RAD-51-like 1 | X |
| Akap13 | A kinase anchor protein | |
| Hspa1a | Heat shock protein 1A | |
| Prpf38b | P-mRNA processing factor | |
| Wdr51a | WD repeat domain 51a | |
| Pabpc4 | Poly(a) binding protein | |
| Dapk1 | Death assoc. protein kinase 1 | |
| Socs7 | Supp. of cytokine signaling 7 | X |
| Extl3 | Exostoses mulitiple-like 3 | X |
| Hspa1a | Heat shock protein 1A | |
| Fosl2 | Fos-like antigen 2 | X |
| Dcc | Deleted in colorectal carcinoma | X |
| Ins1 | Insulin 1 | X |
| Parp9 | Poly (ADP ribsose) polymerase 9 | X |
| Ffar2 | Free fatty acid receptor 2 | |
| Tex21 | Testis expressed gene 21 | |
| Gsta1 | Glutathione S-transferase alpha 1 | |
| Tg | thyroglobulin | |
| Ntf5 | Neurotrophin 5 | X |
| Dgcr8 | DiGeorge syndrome critical region | X |

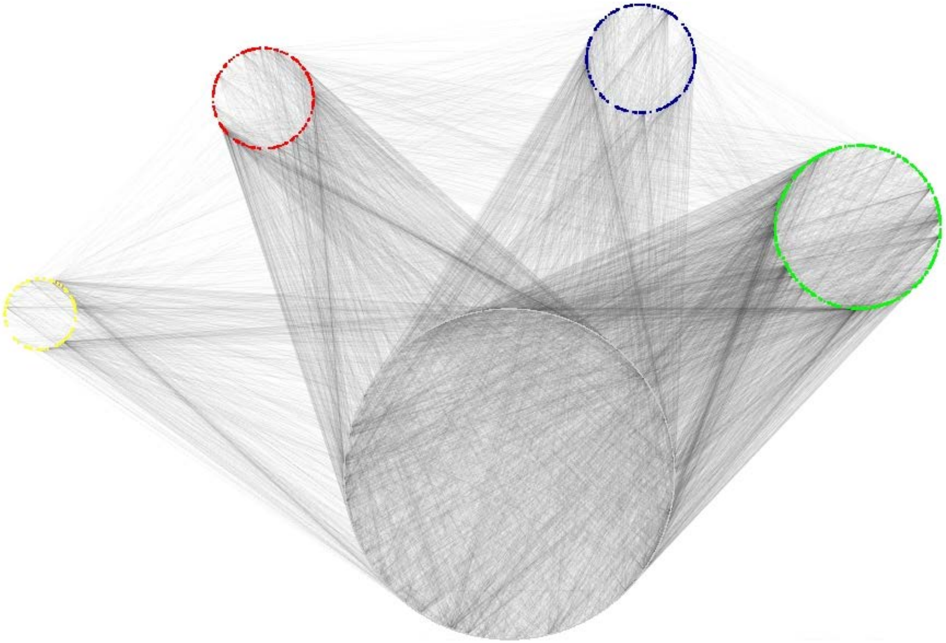
## Aged Mice Top 20 Hubs

| Gene Name | Description | Lethal K/O? |
|---|---|---|
| A930003O Ri | RIKEN cDNA A930003o13 gene | |
| Lgsn | Lengsin, lens protein | |
| Btn1a1 | butyrophilin | X |
| Bcl2l11 | BCL2-like 11 (apoptosis facilitator) | X |
| B230364Ri k | RIKEN cRNA B230369F24 gene | |
| Crkrs | CDC2-related kinase, RS rich | |
| Pkhd1l1 | Polycystic kidney disease like 1 | |
| Rag2 | Recombination activating gene2 | |
| Cyp1a2 | Cytochrome P450 polypept 2 | X |
| Ace | Angiotensin converting enzyme 1 | X |
| Htr4 | 5 hydroxytryptamine receptor 4 | X |
| Ttc17 | Tetratric opeptide repeat domain | |
| Aldh3a1 | Aldehyde dehydrogenase fam 3 | |
| Tfpi | Tissue factor pathway inhibitor | X |
| Trem1 | Triggering receptor (myeloid cells) | |
| Tbx5 | T-box 5 | X |
| Trp63 | Transformation relatied protein 63 | X |
| Klra9 | Killer cell lectin-like receptor | |

# Aging and Biological Networks

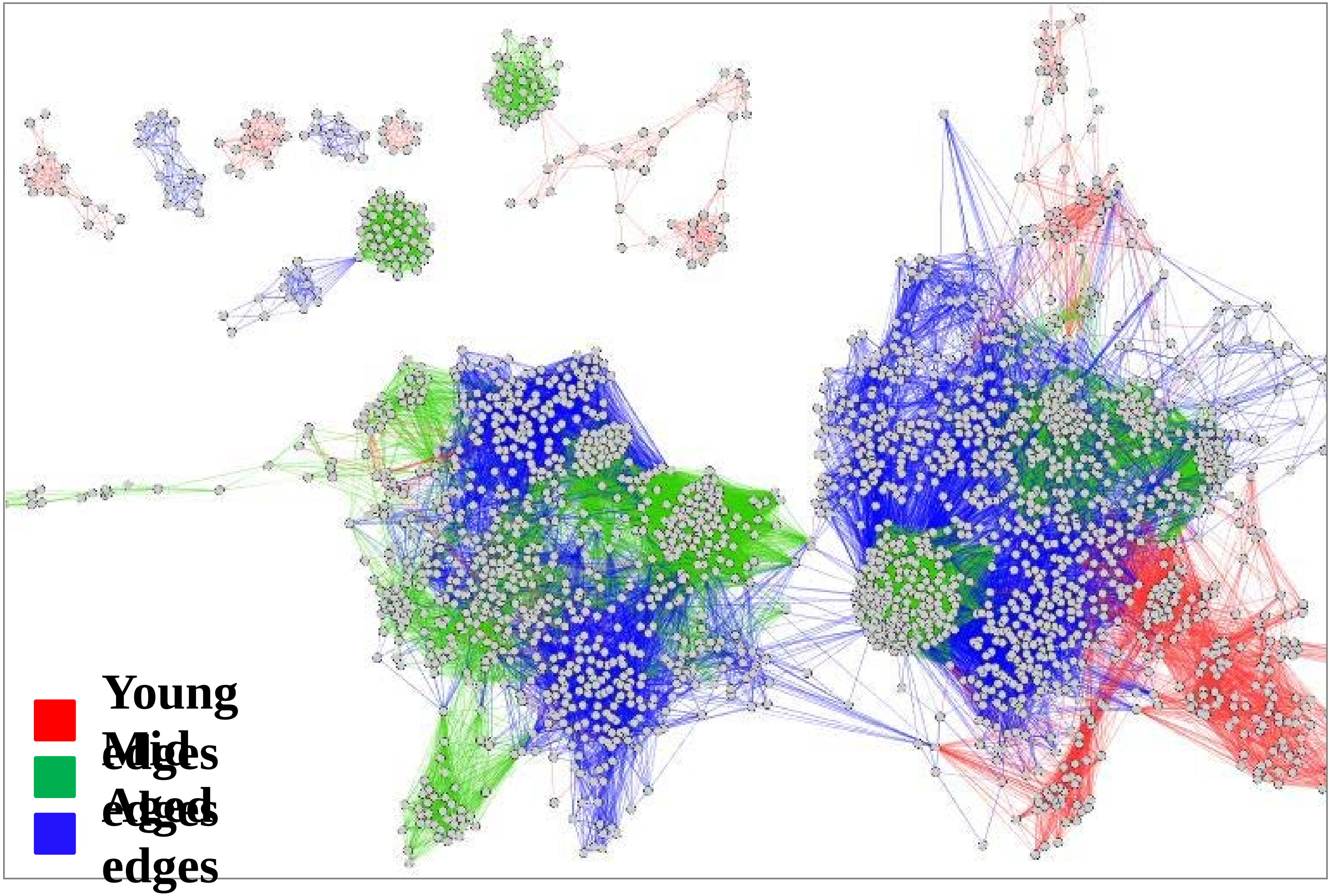


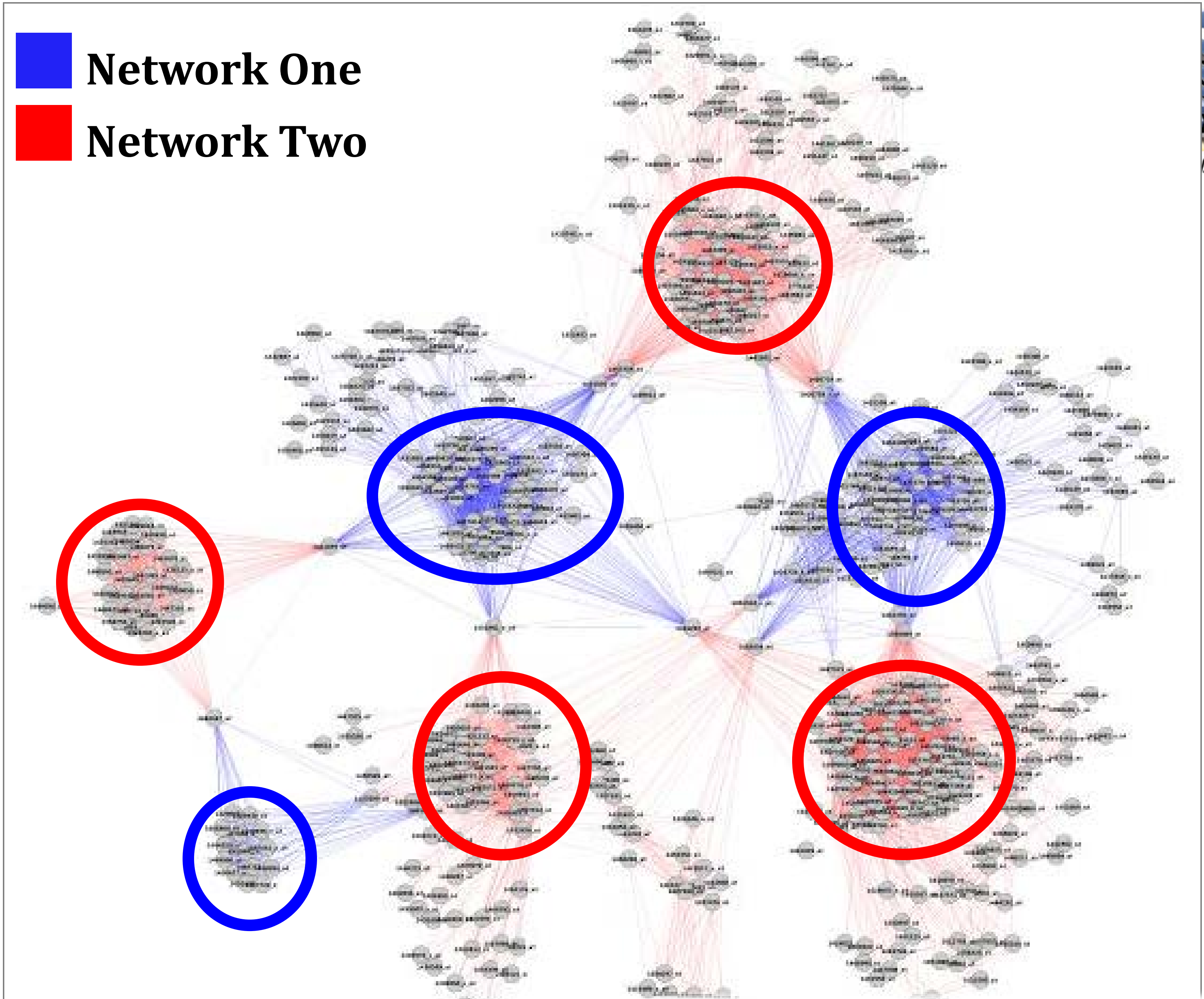
[young]



[aged]

# Aging Networks





Bender *et al.* 2008 . Neuro Biol Aging 29(9):1404-11

# 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

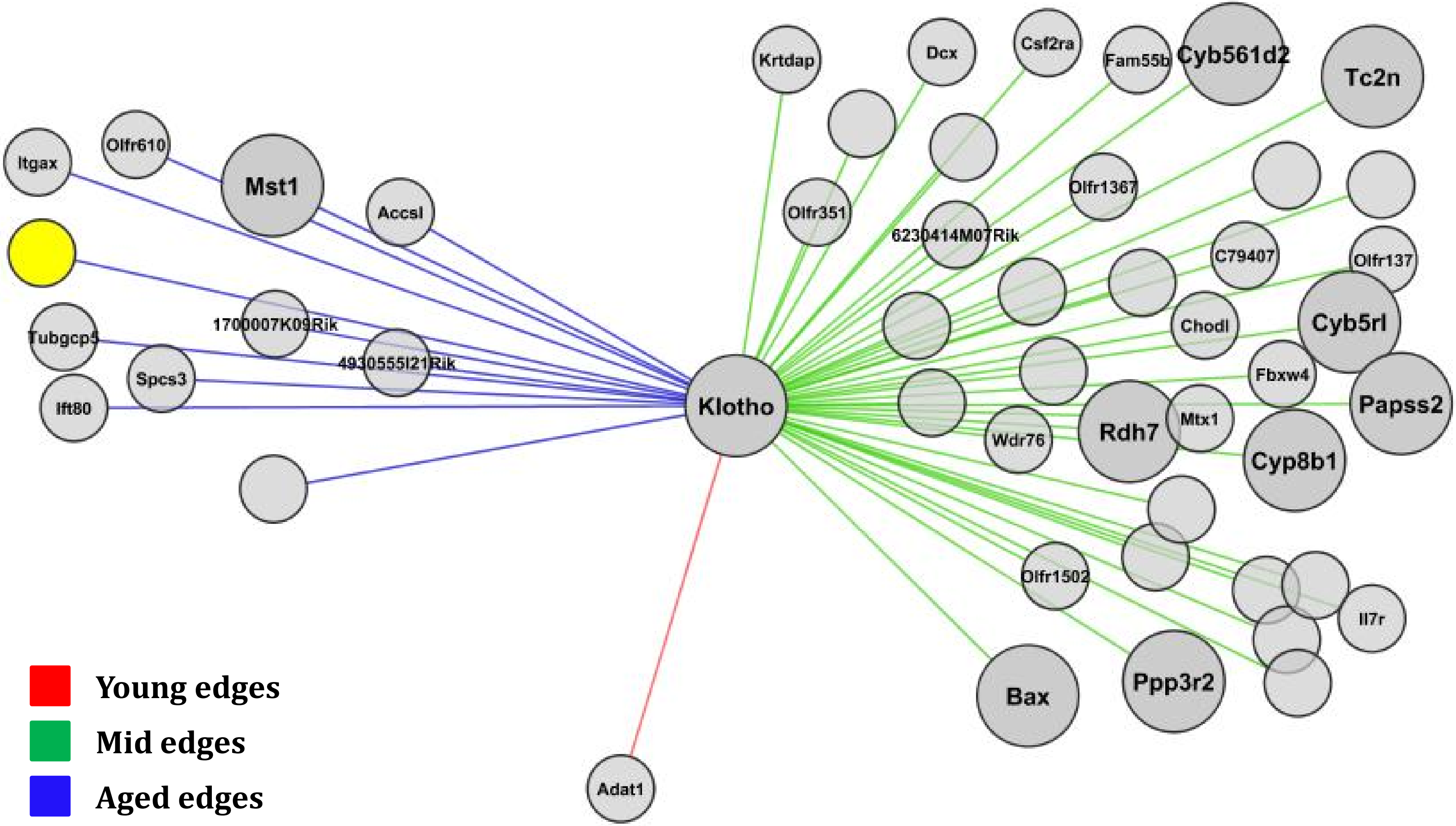
- Define subset of nodes $S = V(X) \cap V(Y)$

- 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:

$$gatewayness_s = \frac{E_{(s)}}{(E1(X) + (E2(Y))}$$

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


Klotho
Itgax
Olfr610
Mst1
Accsl
1700007K09Rik
Tubgcp5
4930555I21Rik
Spcs3
Ift80
Adat1
Krtdap
Dcx
Csf2ra
Fam55b
Cyb561d2
Tc2n
Olfr351
6230414M07Rik
Olfr1367
C79407
Olfr137
Chodl
Cyb5rl
Fbxw4
Papss2
Wdr76
Rdh7
Mtx1
Cyp8b1
Olfr1502
Il7r
Bax
Ppp3r2
🟥 Young edges
🟩 Mid edges
🟦 Aged edges

# Centrality: Integrated Networks

- Networks representing multiple types/states

- Does centrality identify interesting nodes?

- Case study: aging

# Structure Types

**Elements (Nodes):**

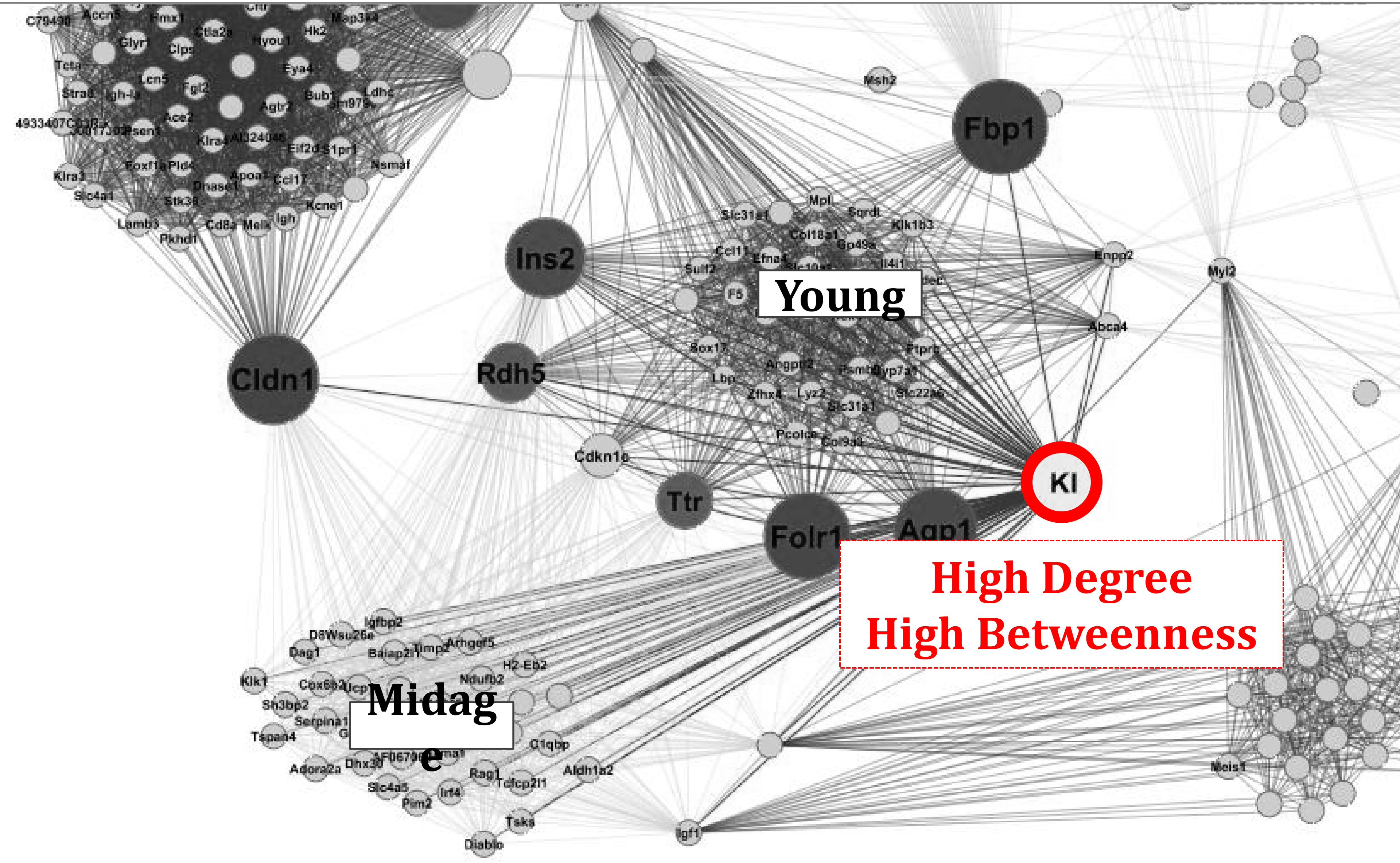
- Betweenness
- Closeness
- Degree
- BC: Highest betweenness + closeness
- CD: Highest degree + closeness
- BD: Highest betweenness + degree
- BCD: Betweenness + closeness + degree

**Our Focus**

**Subsystems (Relationships, groups) :**

- Clusters, cliques
- Pathways
- Loops/cycles

# High BD Node: Klotho

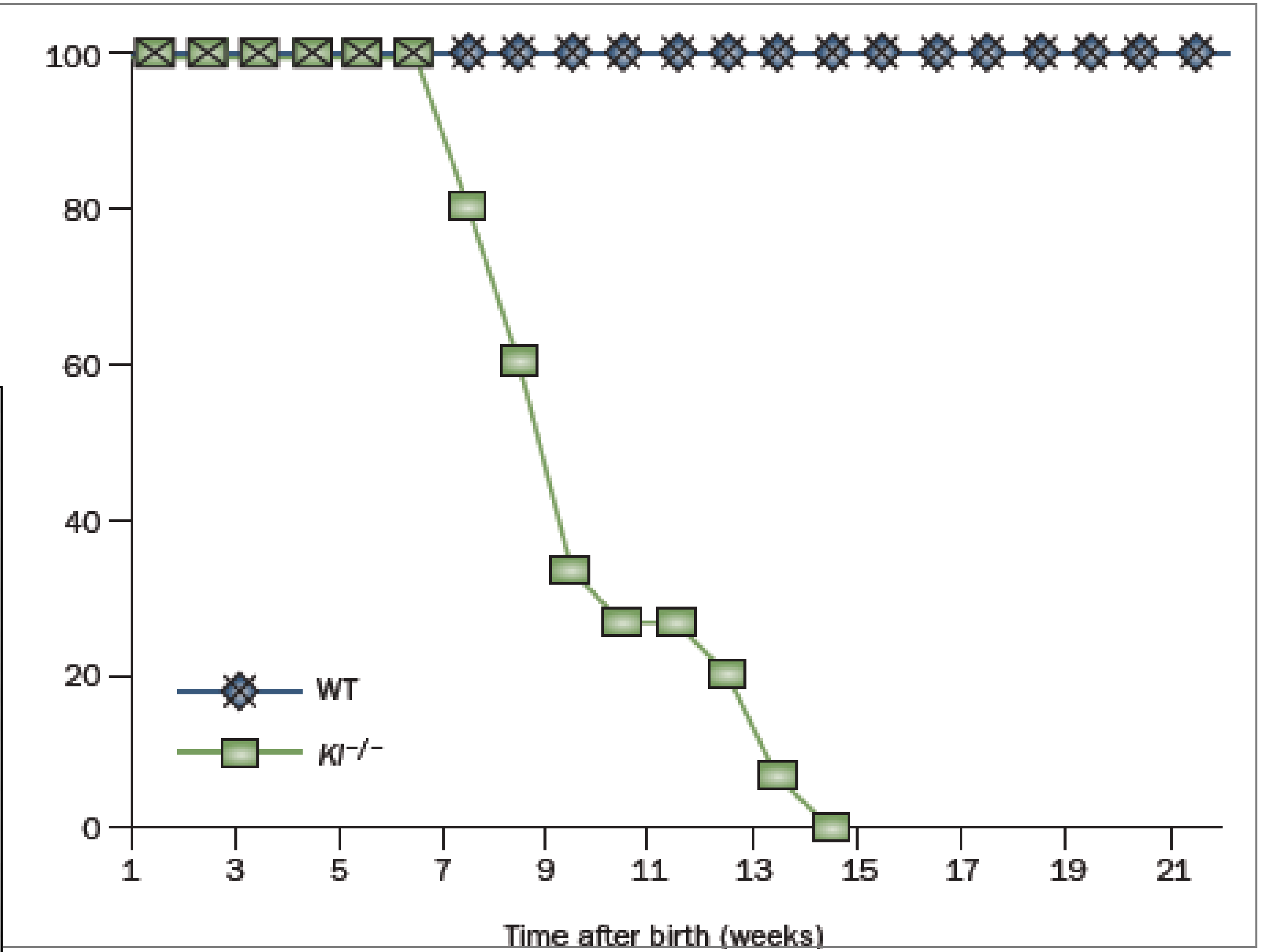


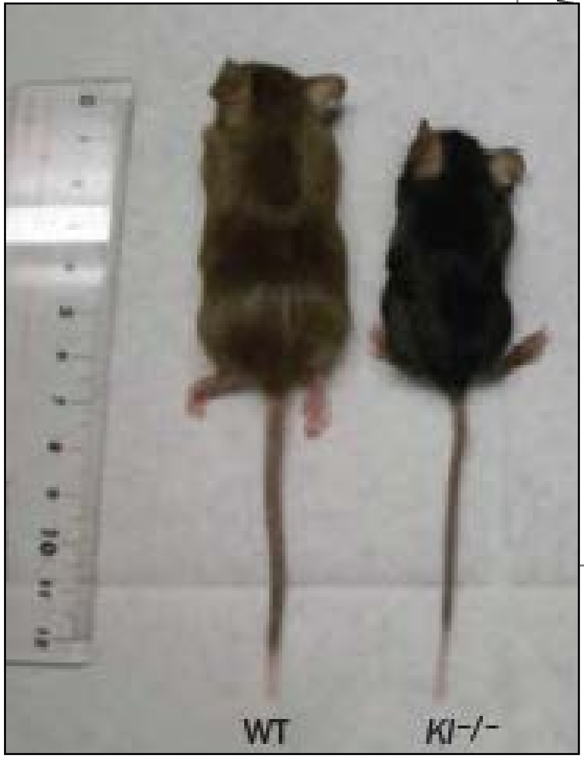
Young

Midag
e

High Degree
High Betweenness

# High BD Node: Validation



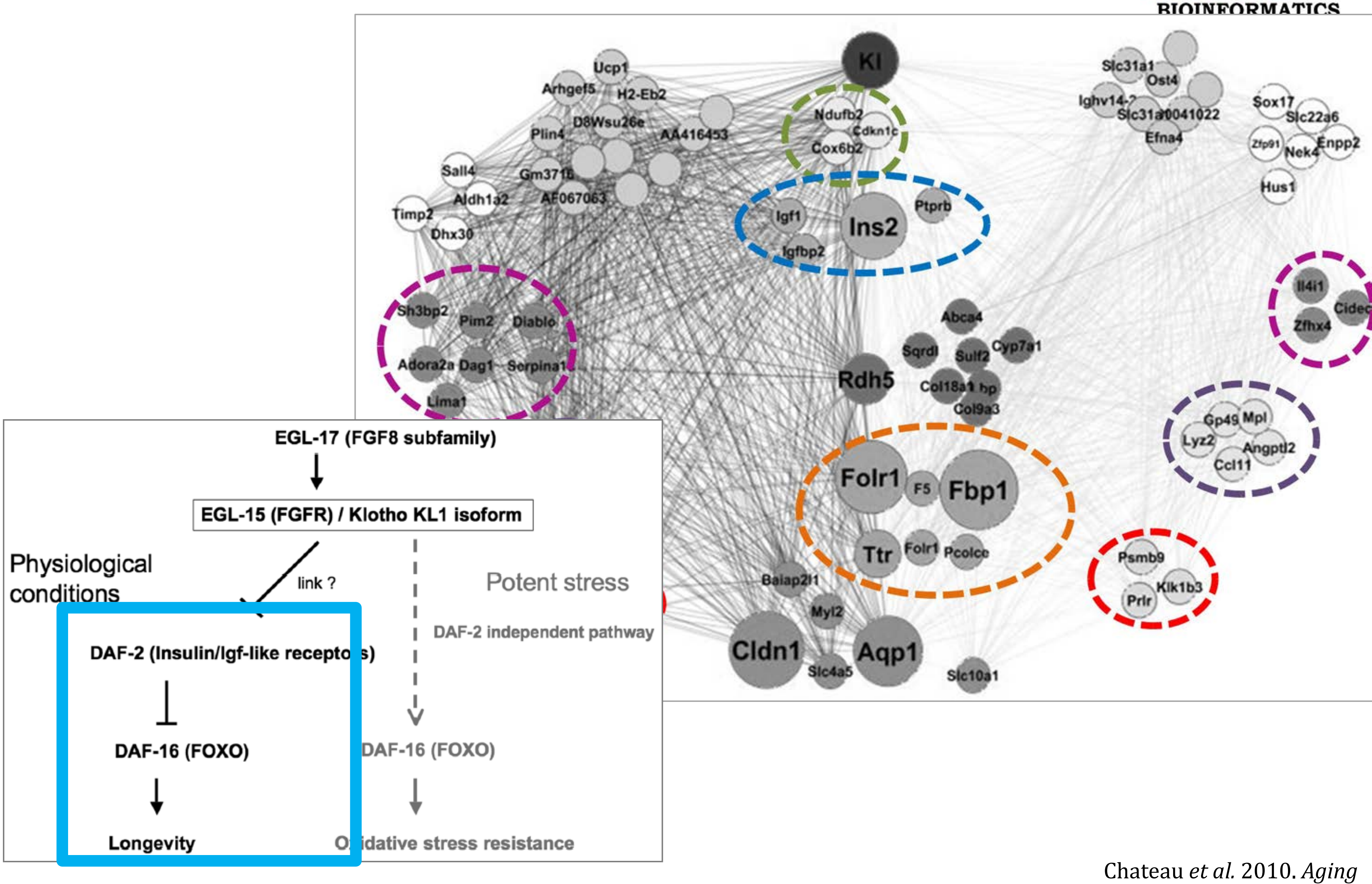


Razzaque,2009. Nat. Rev. Endocrin. 5:611-619

| | | | DEGREE | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ID | GENE | X | Y | TOTAL | GATEWAY SCORE | PERCENT | RANK | TARGETED KNOCKOUT LITERATURE |
| ***X,Y* Gateway Nodes** | 96599_at | Slc4a5 | 55 | 19 | 74 | 0.0999 | 9.9865% | 7 | None available. |
| | 100956_at | Kl | 67 | 43 | 110 | 0.1484 | 14.8448% | 1 | Kl-/- mice are growth retarded with shortened lifespan, mouse model of aging [Kuro-o 1997] |
| | 95471_at | Cdkn1c | 49 | 29 | 78 | 0.1053 | 10.5263% | 6 | Mutations can result in growth retardation and lethality [Yan 1997] |
| | 162302_f_at | Folr1 | 53 | 34 | 87 | 0.1174 | 11.7409% | 2 | Folr1-/- mice have cardiovascular development abnormalities [Zhu 2007] |
| | 95350_at | Ttr | 45 | 37 | 82 | 0.1107 | 11.0661% | 5 | RNAi in *C. elegans* increases lifespan by 14+% [Hansen 2005] |
| | 101431_at | Rdh5 | 52 | 33 | 85 | 0.1147 | 11.4710% | 3 | Mutations in Rdh5 result in visually impaired mouse models [Driessen 2000] |
| | 96123_at | Lbp | 32 | 23 | 55 | 0.0742 | 7.4224% | 8 | Lbp-/- mice have increased susceptibility to bacteria [Wurfel 1997] |
| | 103845_at | Slc31a1 | 19 | 27 | 46 | 0.0621 | 6.2078% | 9 | Homozygous mutants exhibit pre-natal lethality [Nose 2006] |
| | 100150_f_at | Ins2 | 52 | 32 | 84 | 0.1134 | 11.3360% | 4 | Gene disruptions result in post-natal lethality [Duville 1997], insulin secretion related to aging observed in human studies [Chang 2003] |
| | 101877_at | Slc31a1 | 10 | 30 | 40 | 0.0540 | 5.3981% | 10 | 4 |
| | **Cluster Total** | | **434** | **307** | **741** | | | | |

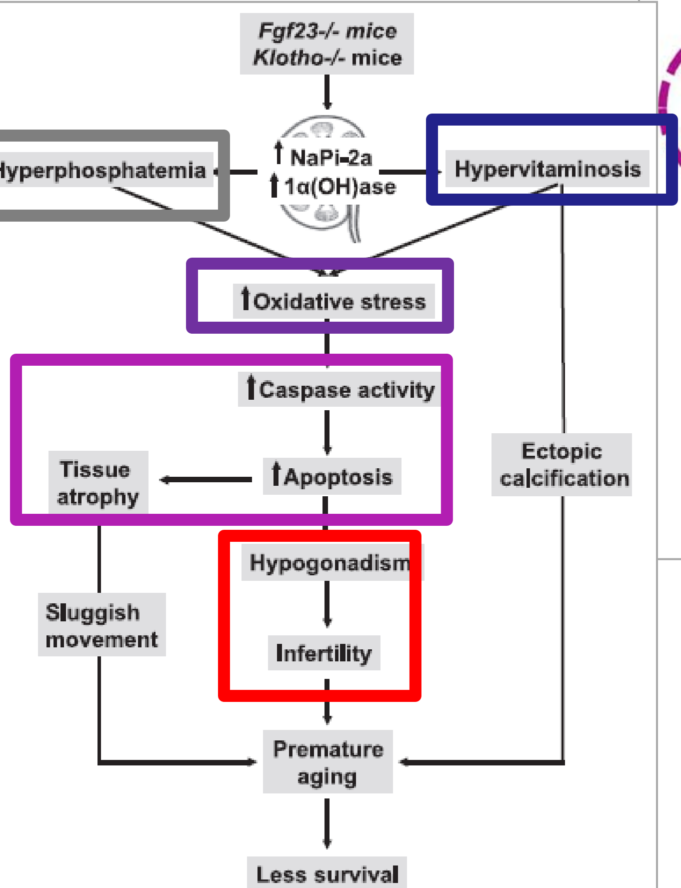
| | | | DEGREE | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ID | GENE | X | Y | TOTAL | GATEWAY SCORE | PERCENT | RANK | TARGETED KNOCKOUT LITERATURE |
| ***U,V* Gateway Nodes** | 93293_at | Calm1\|Calm2\|Calm3 | 9 | 6 | 15 | 0.0806 | 8.0600% | 9 | Loses calcium binding ability with age [Tarcsa 2000] |
| | 101016_at | Arf1 | 12 | 4 | 16 | 0.0860 | 8.6000% | 8 | Expression results in age-correlated change in proliferation [Kim 2006] |
| | 98085_f_at | Apoa1 | 16 | 7 | 23 | 0.1237 | 12.3700% | 4 | Difficulty managing homeostasis, numerous cholesterol regulation issues [Plump 1997] |
| | 160546_at | Aldoc | 12 | 7 | 19 | 0.1022 | 10.2200% | 5 | Associated with age-dependent cellular decline and apoptosis [MGI] |
| | 97696_r_at | Rps24 | 16 | 9 | 25 | 0.1344 | 13.4400% | 2 | Identified as up-regulated in late stages of cognitive aging [Kadish 2009] |
| | 160455_s_at | Zwint | 17 | 9 | 26 | 0.1398 | 13.9800% | 1 | Negatively regulates cell proliferation [Endo 2011] |
| | 96307_s_at | Rpl34 | 16 | 9 | 25 | 0.1344 | 13.4400% | 2 | - |
| | **Cluster Total** | | **121** | **65** | **186** | | | | |

# Validation



Chateau *et al.* 2010. *Aging*

# Subsystems Validation


*Fgf23-/- mice*
*Klotho-/- mice*

↑ NaPi-2a
↑ 1α(OH)ase

yperphosphatemia

Hypervitaminosis

↑Oxidative stress

↑Caspase activity

↑Apoptosis

Tissue atrophy

Ectopic calcification

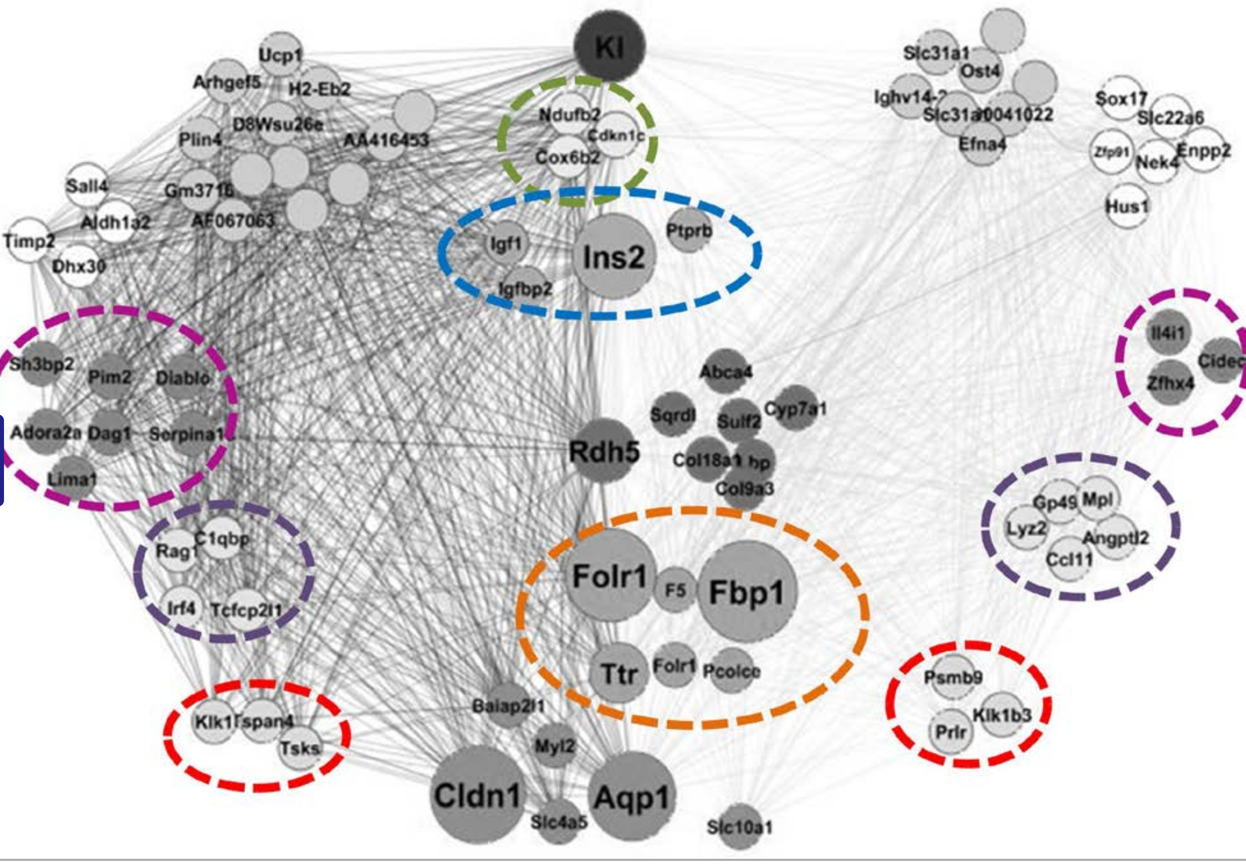
Hypogonadism

Infertility

Sluggish movement

Premature aging

Less survival



Razzaque , 2009. Am J Physiol Renal Physiol. 296(3):F470-6.

# Results Validation

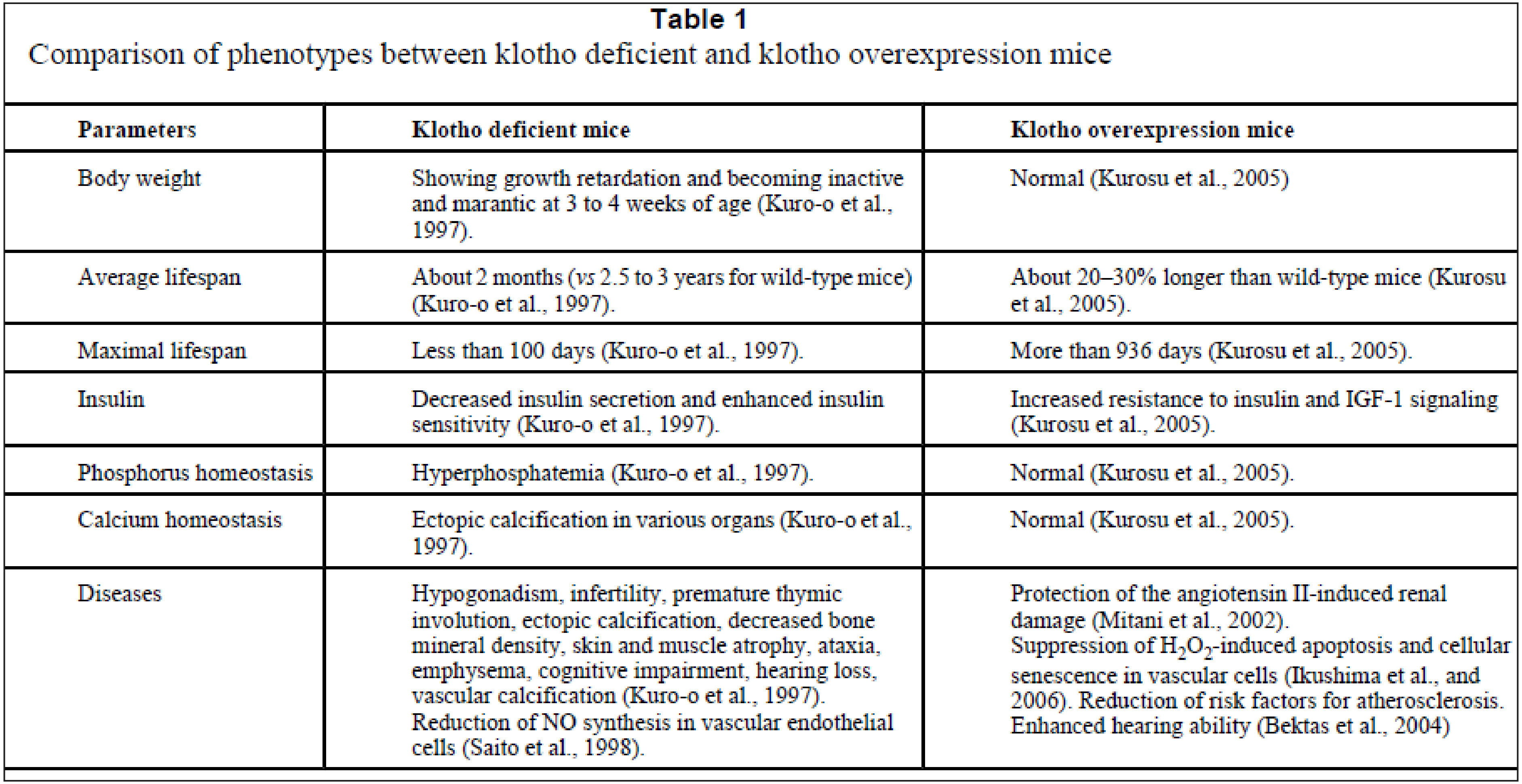


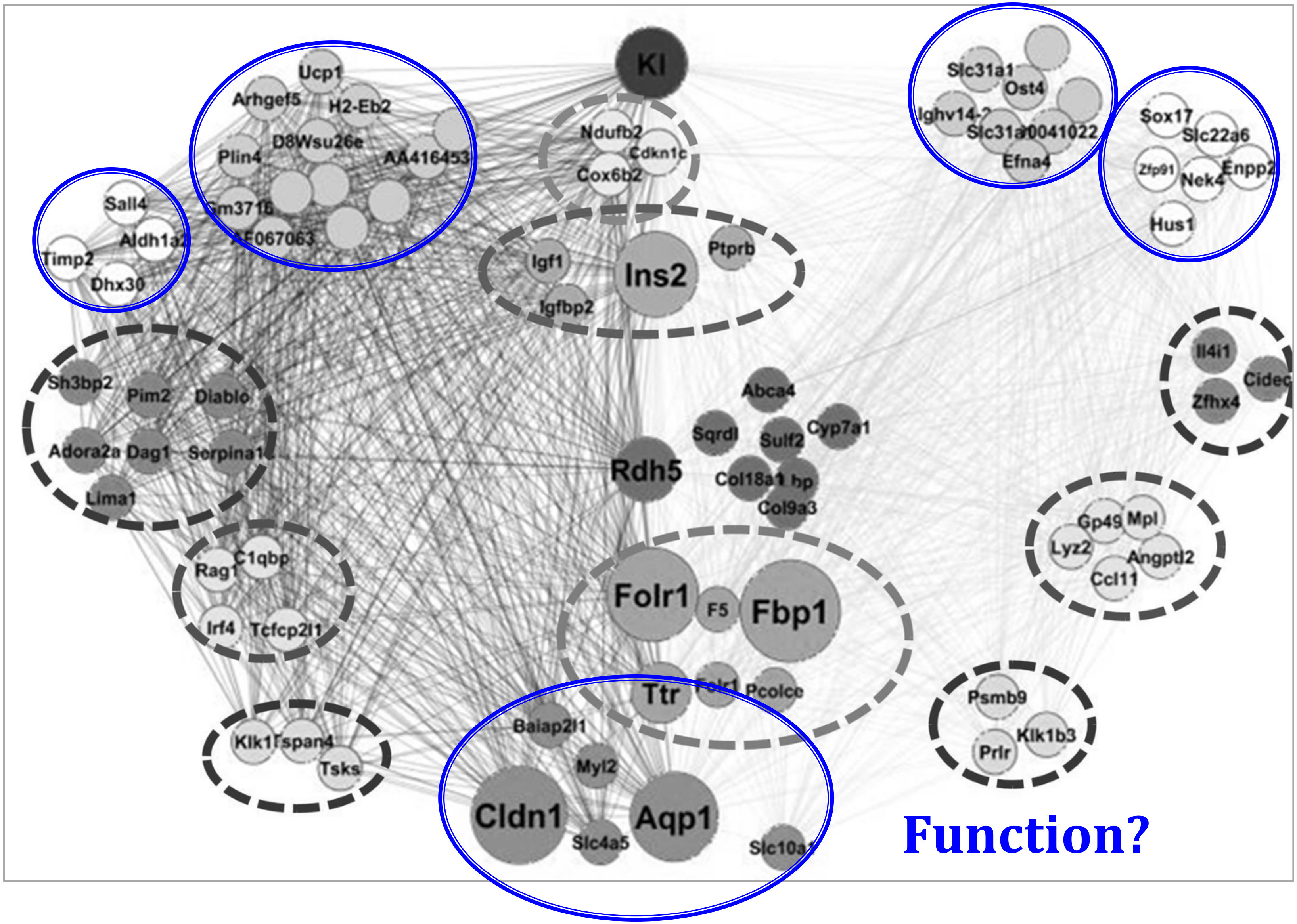
**Table 1**

Comparison of phenotypes between klotho deficient and klotho overexpression mice

| Parameters | Klotho deficient mice | Klotho overexpression mice |
|---|---|---|
| Body weight | Showing growth retardation and becoming inactive and marantic at 3 to 4 weeks of age (Kuro-o et al., 1997). | Normal (Kurosu et al., 2005) |
| Average lifespan | About 2 months (*vs* 2.5 to 3 years for wild-type mice) (Kuro-o et al., 1997). | About 20–30% longer than wild-type mice (Kurosu et al., 2005). |
| Maximal lifespan | Less than 100 days (Kuro-o et al., 1997). | More than 936 days (Kurosu et al., 2005). |
| Insulin | Decreased insulin secretion and enhanced insulin sensitivity (Kuro-o et al., 1997). | Increased resistance to insulin and IGF-1 signaling (Kurosu et al., 2005). |
| Phosphorus homeostasis | Hyperphosphatemia (Kuro-o et al., 1997). | Normal (Kurosu et al., 2005). |
| Calcium homeostasis | Ectopic calcification in various organs (Kuro-o et al., 1997). | Normal (Kurosu et al., 2005). |
| Diseases | Hypogonadism, infertility, premature thymic involution, ectopic calcification, decreased bone mineral density, skin and muscle atrophy, ataxia, emphysema, cognitive impairment, hearing loss, vascular calcification (Kuro-o et al., 1997).<br>Reduction of NO synthesis in vascular endothelial cells (Saito et al., 1998). | Protection of the angiotensin II-induced renal damage (Mitani et al., 2002).<br>Suppression of $H_2O_2$-induced apoptosis and cellular senescence in vascular cells (Ikushima et al., and 2006). Reduction of risk factors for atherosclerosis.<br>Enhanced hearing ability (Bektas et al., 2004) |

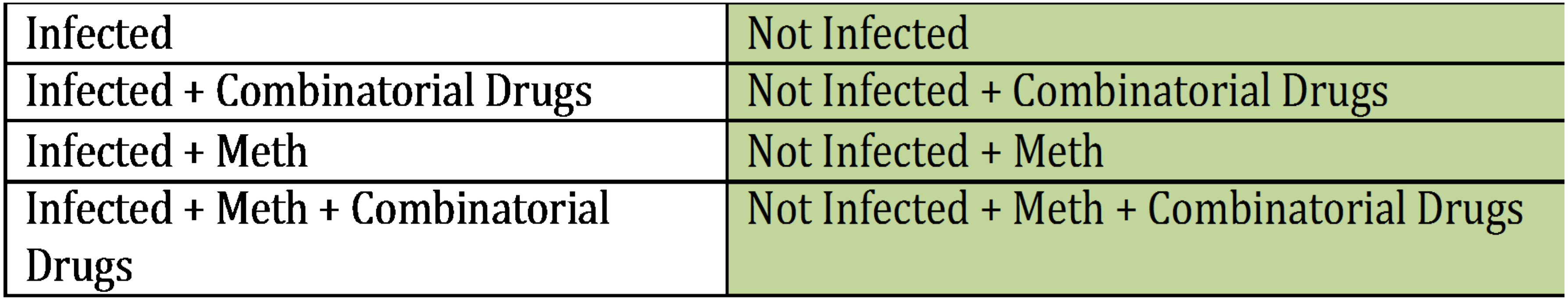
Wang, Sun *et al*. 2009 *Ageing Res Rev. 8(1): 43–51.*

# Discovery



Function?

# Case Study: HIV and Drug Addiction



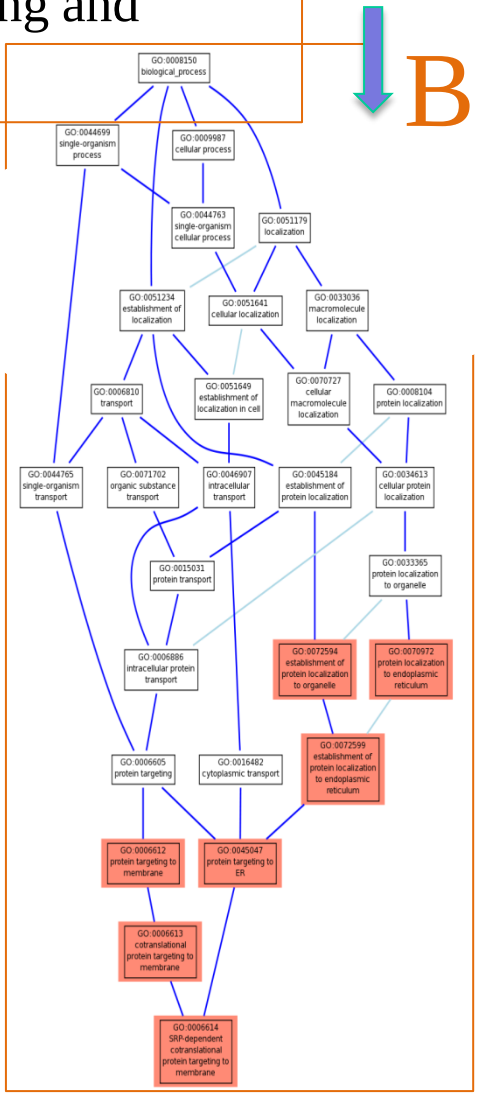
| | |
|---|---|
| Infected | Not Infected |
| Infected + Combinatorial Drugs | Not Infected + Combinatorial Drugs |
| Infected + Meth | Not Infected + Meth |
| Infected + Meth + Combinatorial Drugs | Not Infected + Meth + Combinatorial Drugs |

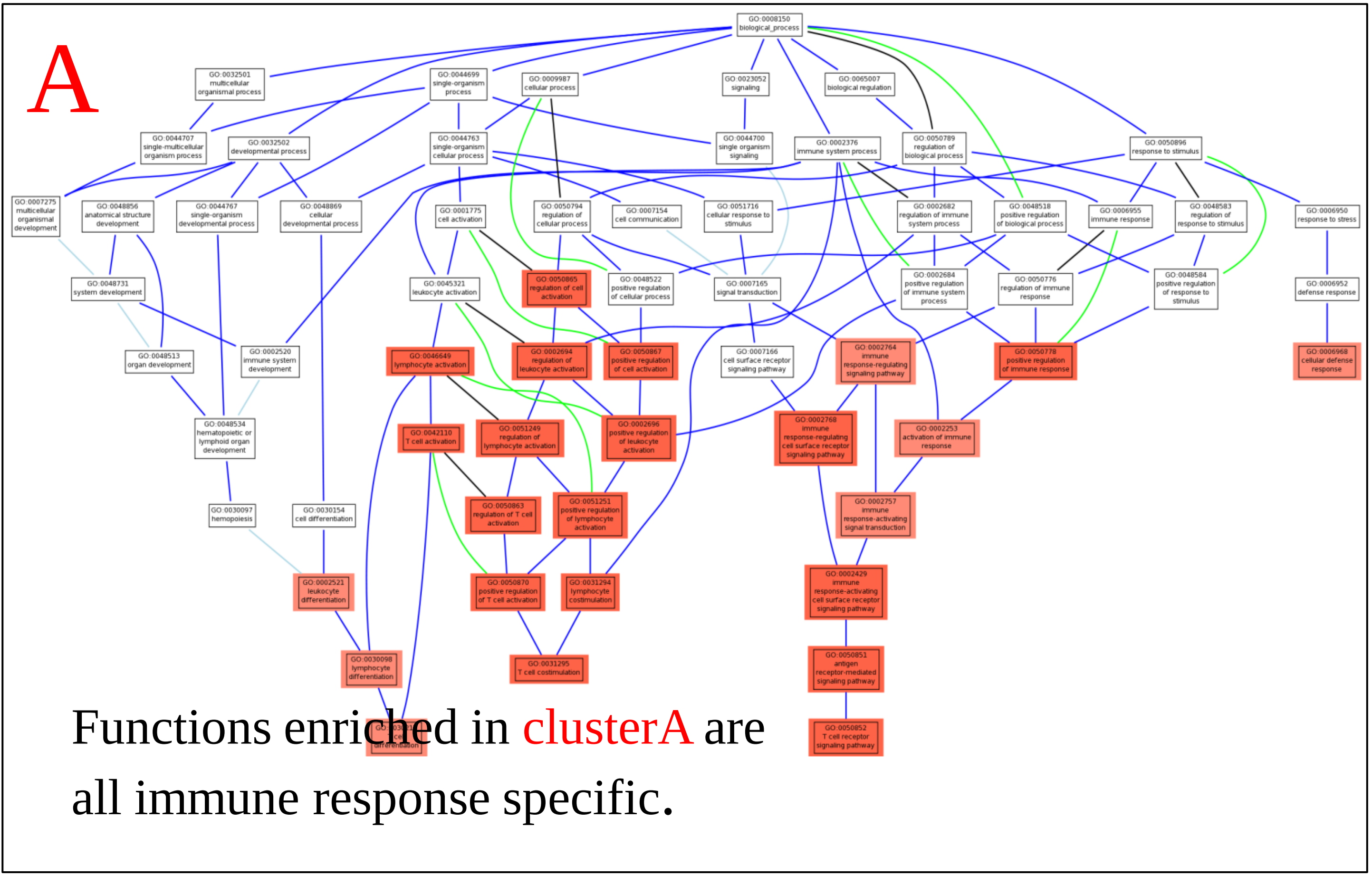
# Role of Methamphetamine

- 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 Enriched in Specific Functions

Functions enriched in clusterB are protein targeting and localization.

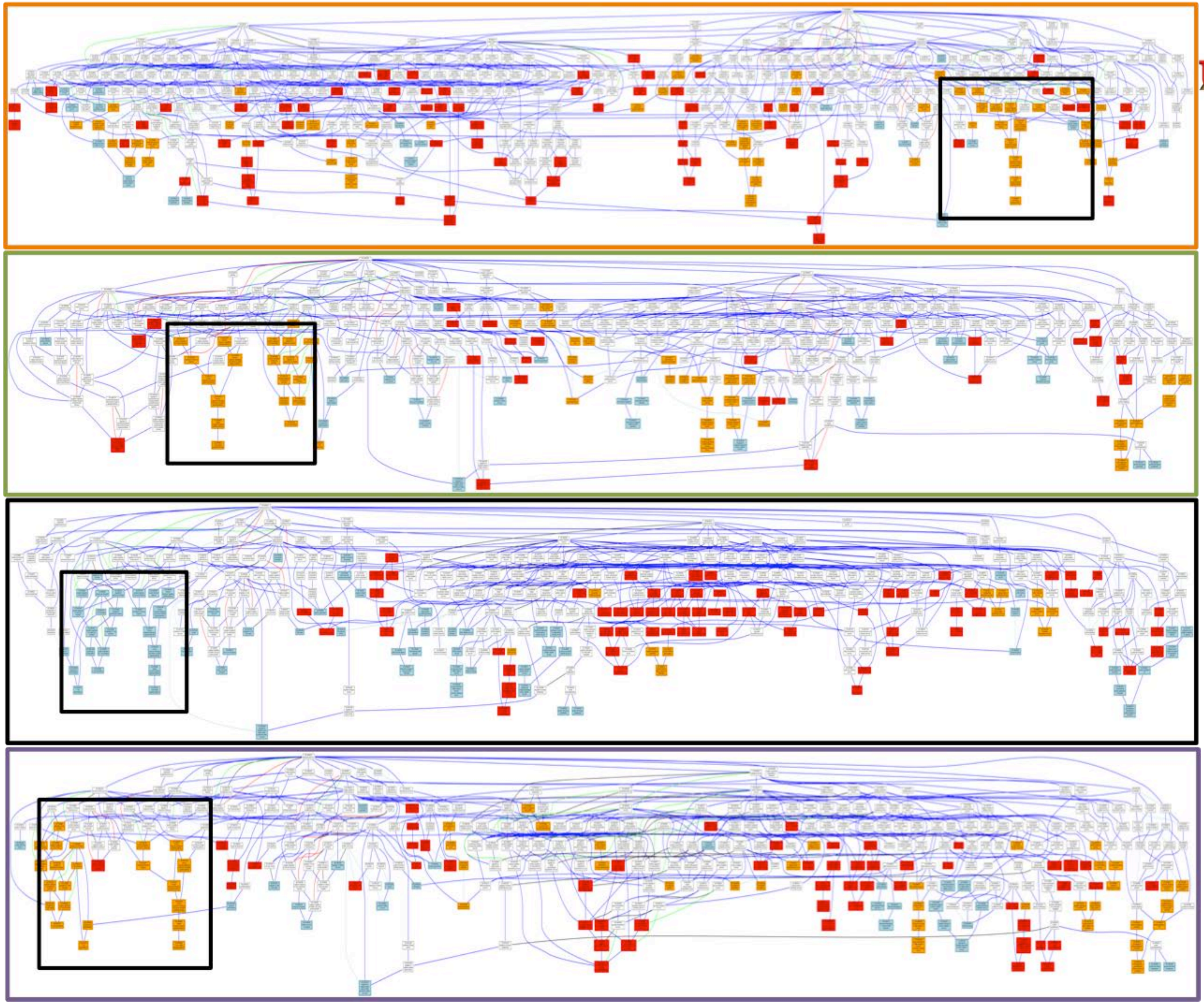

B
GO:0008150 biological_process
GO:0044699 single-organism process
GO:0009987 cellular process
GO:0044763 single-organism cellular process
GO:0051179 localization
GO:0051234 establishment of localization
GO:0051641 cellular localization
GO:0033036 macromolecule localization
GO:0006810 transport
GO:0051649 establishment of localization in cell
GO:0070727 cellular macromolecule localization
GO:0008104 protein localization
GO:0044765 single-organism transport
GO:0071702 organic substance transport
GO:0046907 intracellular transport
GO:0045184 establishment of protein localization
GO:0034613 cellular protein localization
GO:0015031 protein transport
GO:0033365 protein localization to organelle
GO:0006886 intracellular protein transport
GO:0072594 establishment of protein localization to organelle
GO:0070972 protein localization to endoplasmic reticulum
GO:0072599 establishment of protein localization to endoplasmic reticulum
GO:0006605 protein targeting
GO:0016482 cytoplasmic transport
GO:0006612 protein targeting to membrane
GO:0045047 protein targeting to ER
GO:0006613 cotranslational protein targeting to membrane
GO:0006614 SRP-dependent cotranslational protein targeting to membrane


A
GO:0008150 biological_process
GO:0032501 multicellular organismal process
GO:0044699 single-organism process
GO:0009987 cellular process
GO:0023052 signaling
GO:0065007 biological regulation
GO:0044707 single-multicellular organism process
GO:0032502 developmental process
GO:0044763 single-organism cellular process
GO:0044700 single organism signaling
GO:0002376 immune system process
GO:0050789 regulation of biological process
GO:0050896 response to stimulus
GO:0007275 multicellular organismal development
GO:0048856 anatomical structure development
GO:0044767 single-organism developmental process
GO:0048869 cellular developmental process
GO:0001775 cell activation
GO:0050794 regulation of cellular process
GO:0007154 cell communication
GO:0051716 cellular response to stimulus
GO:0002682 regulation of immune system process
GO:0048518 positive regulation of biological process
GO:0006955 immune response
GO:0048583 regulation of response to stimulus
GO:0006950 response to stress
GO:0048731 system development
GO:0045321 leukocyte activation
GO:0050865 regulation of cell activation
GO:0048522 positive regulation of cellular process
GO:0007165 signal transduction
GO:0002684 positive regulation of immune system process
GO:0050776 regulation of immune response
GO:0048584 positive regulation of response to stimulus
GO:0006952 defense response
GO:0048513 organ development
GO:0002520 immune system development
GO:0046649 lymphocyte activation
GO:0002694 regulation of leukocyte activation
GO:0050867 positive regulation of cell activation
GO:0007166 cell surface receptor signaling pathway
GO:0002764 immune response-regulating signaling pathway
GO:0050778 positive regulation of immune response
GO:0006968 cellular defense response
GO:0048534 hematopoietic or lymphoid organ development
GO:0042110 T cell activation
GO:0051249 regulation of lymphocyte activation
GO:0002696 positive regulation of leukocyte activation
GO:0002768 immune response-regulating cell surface receptor signaling pathway
GO:0002253 activation of immune response
GO:0030097 hemopoiesis
GO:0030154 cell differentiation
GO:0050863 regulation of T cell activation
GO:0051251 positive regulation of lymphocyte activation
GO:0002757 immune response-activating signal transduction
GO:0002521 leukocyte differentiation
GO:0050870 positive regulation of T cell activation
GO:0031294 lymphocyte costimulation
GO:0002429 immune response-activating cell surface receptor signaling pathway
GO:0030098 lymphocyte differentiation
GO:0031295 T cell costimulation
GO:0050851 antigen receptor-mediated signaling pathway
GO:0050852 T cell receptor signaling pathway
differentiation

Functions enriched in clusterA are all immune response specific.

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

UNO BIOINFORMATICS



Infected vs. Uninfected

HIV treatment vs. Uninfected

Infected + Meth vs. Uninfected

Infected + Meth+ Treatment vs. Uninfected

# Obtained Results

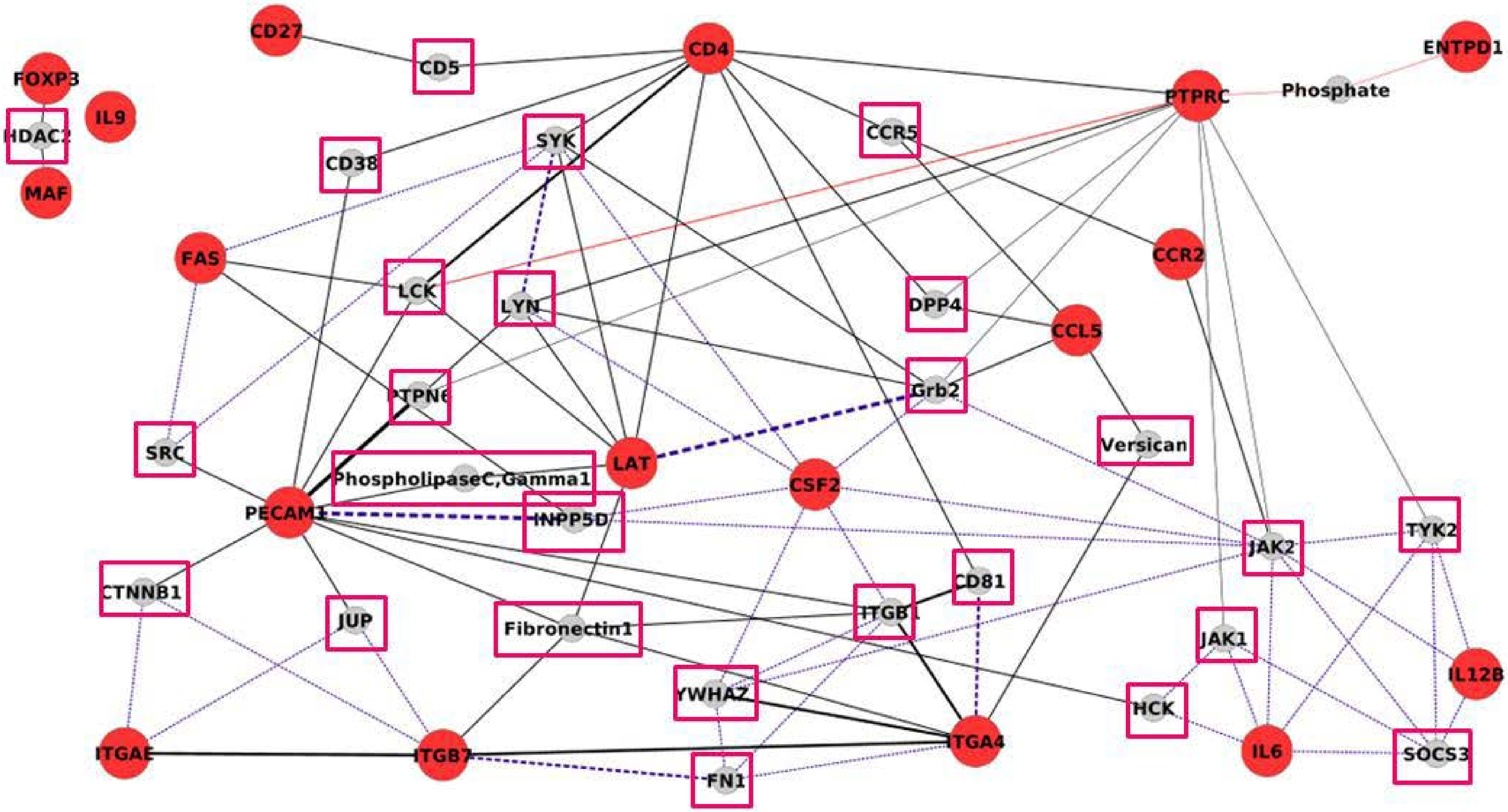
- Large number of nodes are enriched in only one network in Infected + Meth network.
  - Many functions enriched in other conditions have been dropped out in Infected + Meth network.
- Most of the lost functions reappear in Infected + Treated
- Some of these lost functions reappear in Infected + Meth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

# Case Study: Parkinson's Disease

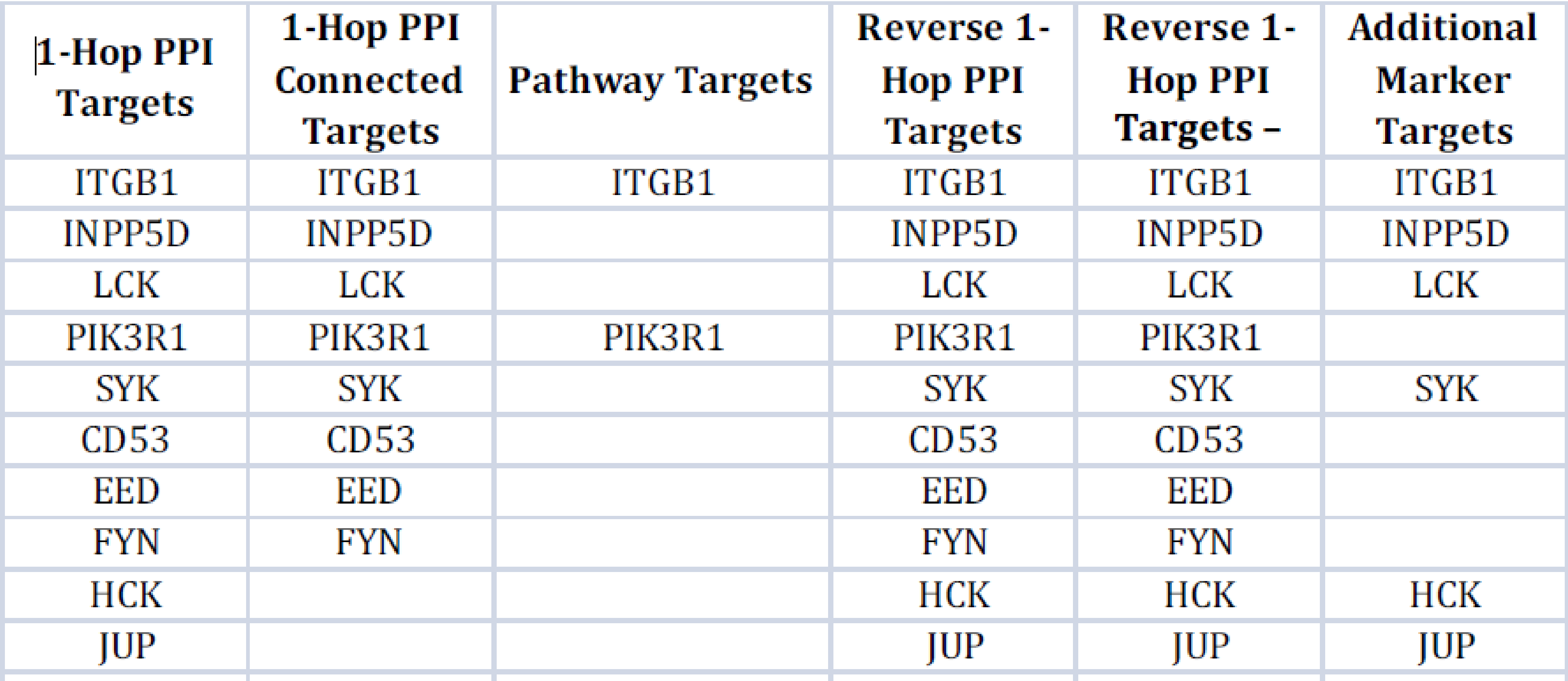
- Data: Flow cytometry markers
  - Parkinsons Disease patients
  - Caretakers (non-Parkinsons)

- 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 |



| 1-Hop PPI Targets | 1-Hop PPI Connected Targets | Pathway Targets | Reverse 1-Hop PPI Targets | Reverse 1-Hop PPI Targets – | Additional Marker Targets |
|---|---|---|---|---|---|
| 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 |

# Case Study III: Diabetes

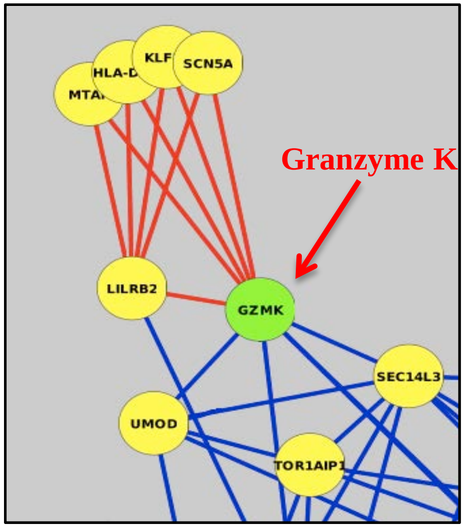
- Decrease in insulin can be a cause of Type II diabetes
- Insulin made by pancreatic beta cells in the islets of langerhans
- Isolate these cells in diabetic patients and compare to normal cells

# Case Study III: Diabetes

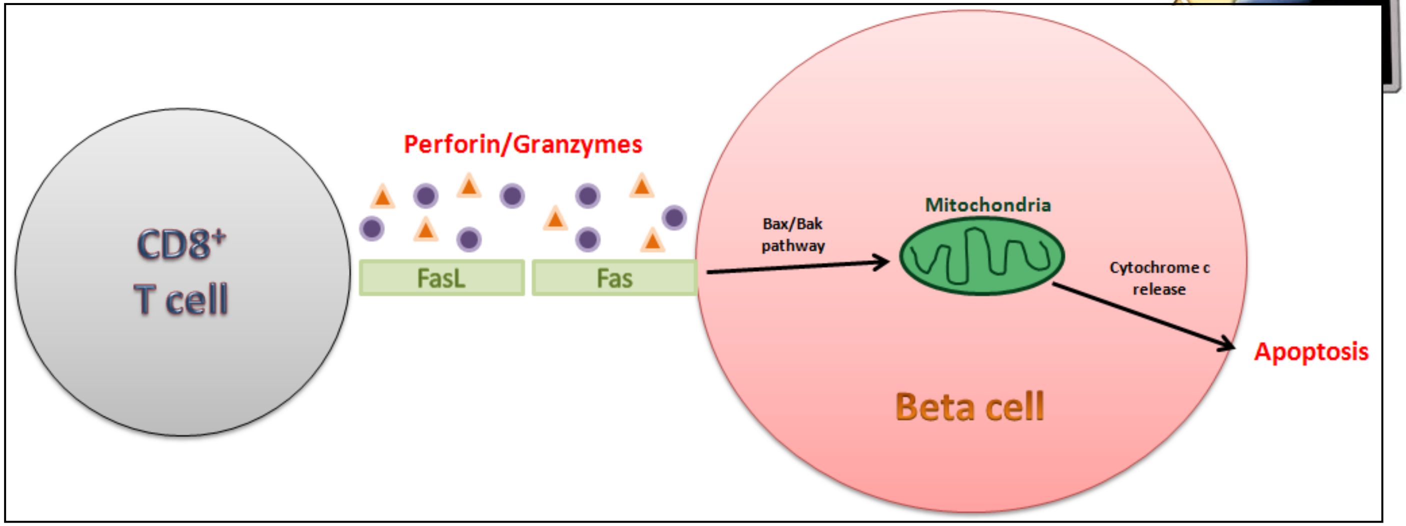
- Data: Pancreatic beta cells (GSE25724)
  - Healthy adults
  - Diabetic adults (Type II, adult onset)
  - Case-matched

- Method:
  - Identify gateway nodes in 2-state correlation network
  - Normal vs. diseased

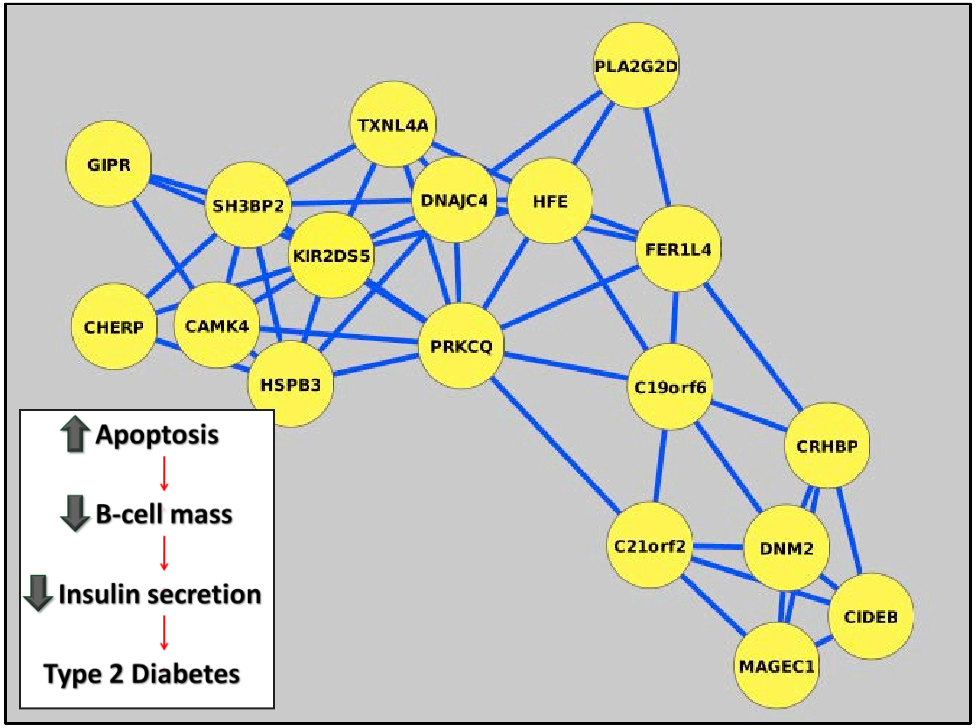
- Outcome:
  - Identification of top gateway nodes
  - Notable: Granzyme K
    - Facilitates apoptosis in pancreatic beta cells
    - Apoptosis is an upstream step in the development of adult onset Type II diabetes



Granzyme K



Apoptosis



⬆ Apoptosis

⬇ B-cell mass

⬇ Insulin secretion

Type 2 Diabetes

🟦 Normal pancreatic β cells

🟥 Diabetic pancreatic β cells

# Case Study III: Summary

- Identified an enzyme (Granzyme K) that is involved in apoptosis (cell death)
- Identified a cluster of genes involved in apoptosis that was present in normal cells but absent in diabetic cells
- Increase in pancreatic cell death leads to lower insulin production → cause of Type II diabetes
- Suggests that disregulated apoptosis could be a cause of type II diabetes

# Summary: Correlation Networks

- Networks → very efficient modeling system
  - Basis of next generation data analysis tools in systems biology
- Structure/function relationship exists
  - Integrated networks to identify gene drivers
- Future: Model will play a role in aging/disease *prevention, diagnosis,* and *treatment*

# Tutorial Outlines

- Biomedical Informatics - Challenges and Opportunities
- Next Generation Bioinformatics Tools – Focus on Data Analysis and Integration
- Systems Biology and Network Analysis
- Biomedical Informatics and the Cloud: Models and Security
- Case Studies of Discoveries using Biological Networks
- *HPC in Network Analysis of High Throughput Biological Data*
- Integration of different aspects of Biomedical Informatics
- Next Steps – where to go from here?

# HPC and Biological Networks

- Network creation: 2 weeks on PC
    - 10 hours in parallel, 50 nodes
    - 40,000 nodes = 800 million edges
- Network analysis: Best in parallel
    - Only 3% of entire genome forms complexes
- UNO Sapling Cluster
- Holland Computing Center: Firefly

# Challenges

(1) Biological networks can be massive in size

Supercomputing access may be limited
Biological network knowledge may be limited.

(2) Noise within the network is likely

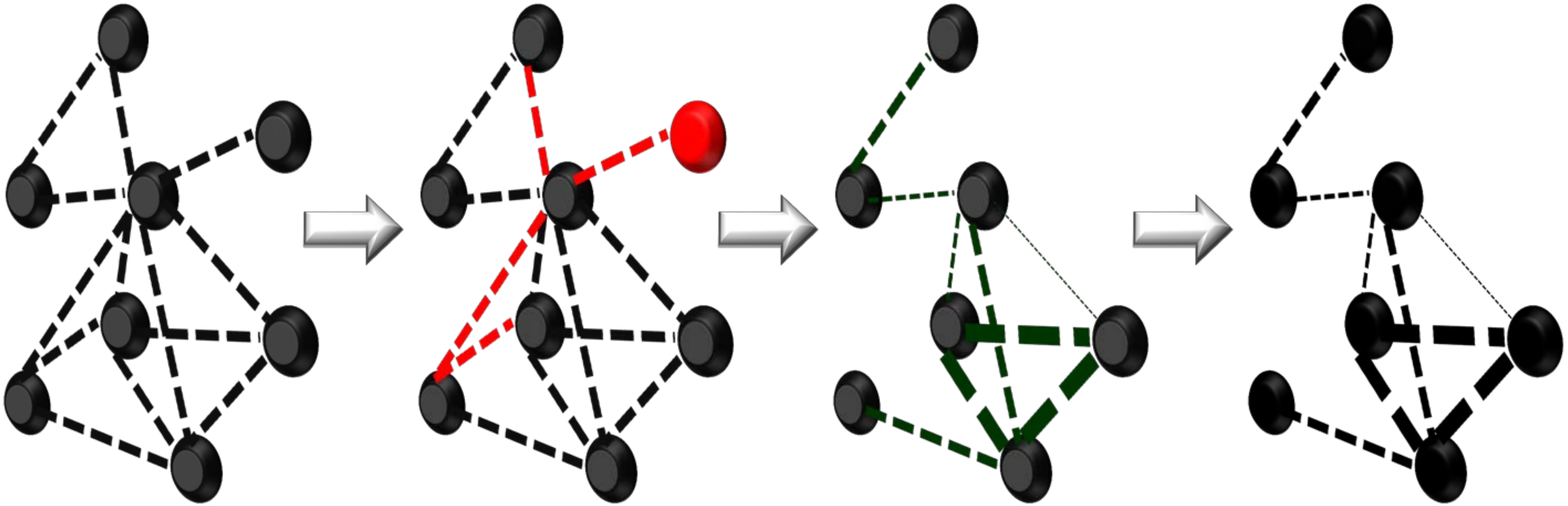
Noise within the network cannot be ignored.

How to address these issues: Network filters

- Reduce network size
- Maintain biological signal
- Improve upon biological signal?

# HPC and Network Analysis

- Network sizes tend to be large
- Signal-to-noise ratio can be high
    - ID biologically relevant relationships?
    - Remove irrelevant nodes/edges?



Original network

ID noisy edges

Weight with literature

Enriched network

# Large-Scale Networks and Data Analysis

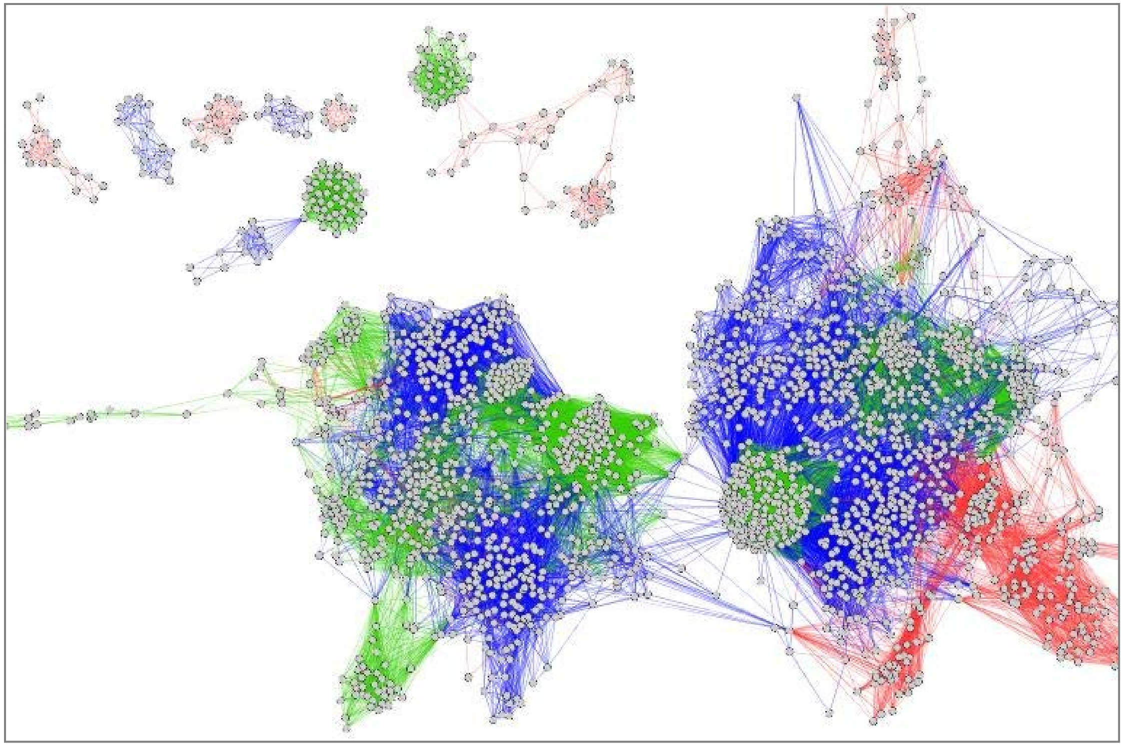
Many application domain rely on creating networks to model and analyze key relationships among data elements in the domain

Examples: Biological networks – social networks – inference networks - scheduling networks – Transportation networks

- ✓ Modeling versus data mining
  - ➢ Such networks are normally very large
  - ➢ They are susceptive to significant noise related problems
- ✓ Sampling sub-networks (sub-graphs)
  - ➢ Reduce network size
  - ➢ Reduce noise impact
  - ➢ Questions: does it preserve integrality of the original network

# Back to Correlation Networks

- Model for handling high-throughput biological data
- Network contains biologically relevant subgraphs:
  - – Hubs
  - – Clusters
  - – Motifs
  - – Bottlenecks



# Size and Noise

- Network made from average gene expression experiment will have:
  - 40,000 nodes
  - 800 million edges

- Only 3% of genes in entire genome work together to form complexes

- Even with parallel computing resources, unfiltered networks are too noisy for biological discovery

# Network Filters

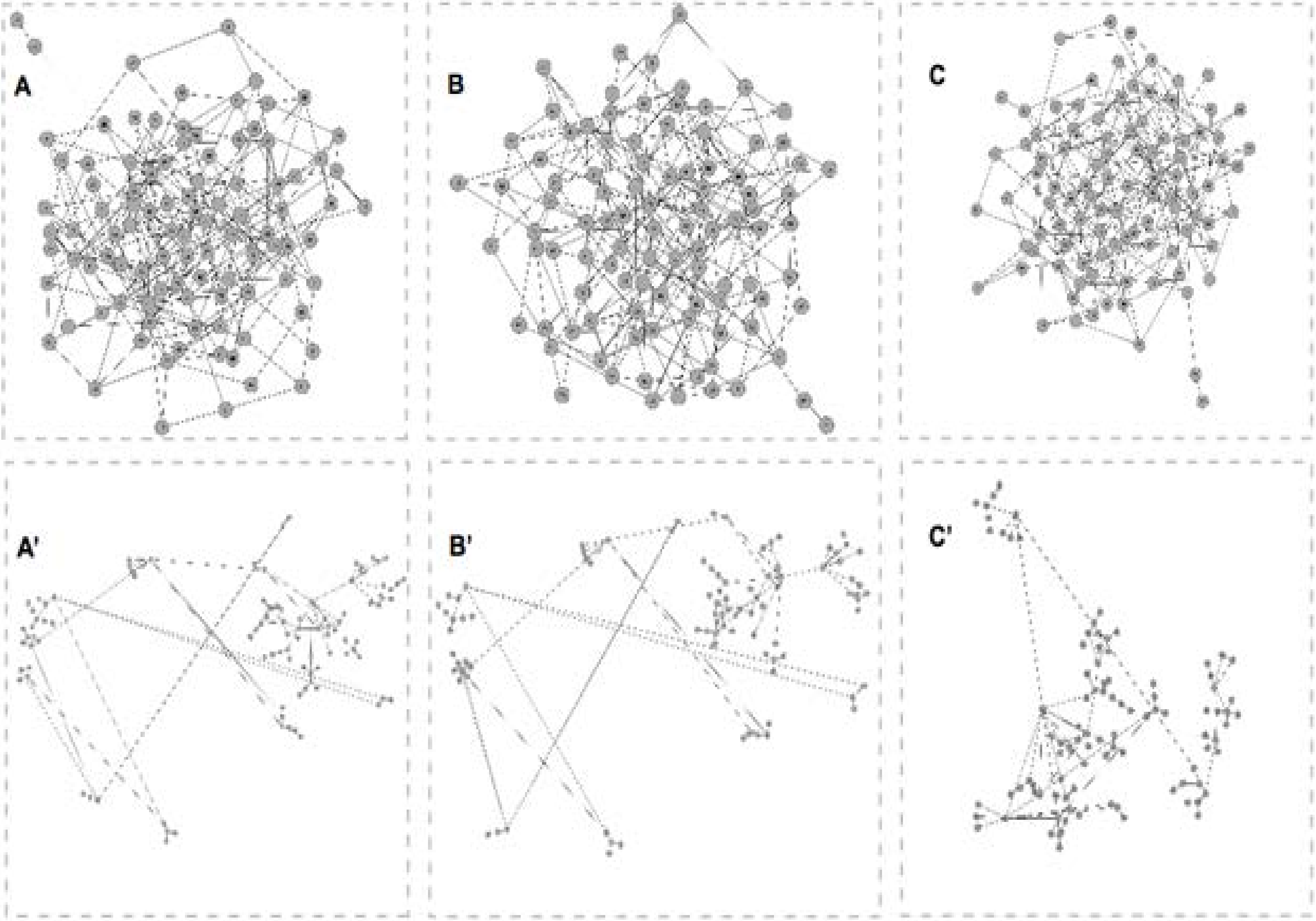
Design a network filter and obtain a sub-network of the original network such that:

- It maintains the important stuff – signal
- Remove unimportant stuff – noise
- Maintain network elements of biological relevance
- Uncover new ones

# Network Filters

- Chordal graph sampling
    - Keep triangles in expression graphs
    - Remove large cycles, extra edges
    - Keep clusters, identify new clusters
- Spanning tree sampling
    - Keep high degree nodes (maybe?)
    - Remove up to 50% of edges
    - Enhance identification of lethal nodes
- Hybrid chordal-spanning tree method
    - Keep high degree nodes
    - Keep clusters
    - Remove 40-50% of edges
    - Proactively distort/enlarge network structures

# Need to Maintain Key Structures



# Chordal Graph Sampling

**Goal:** Develop a parallel network sampling technique that *filters noise*, while *preserving the important characteristics of the network*.

✓Maximal Chordal Subgraph

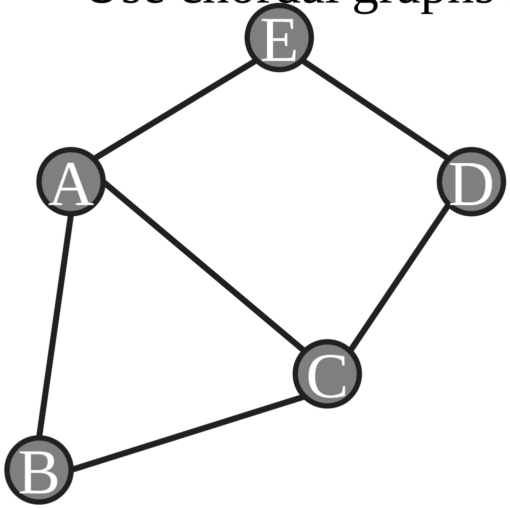
- ➢Spanning subgraph of the network w
  - ➢No cycles of length larger than three

✓Properties of Chordal Graph

- ➢Preserves most cliques and highly connected regions of the network
- ➢Most NP hard problems can be solved in polynomial time
- ➢Complexity of finding maximal chordal subgraphs: **O(|E|*max_deg)**

# Why chordal graphs?

- Chordal graphs are triangulated
  - We want to preserve $K_3$ subgraphs (triangle)
  - $K_3$ graphs/motifs are known to represent co-regulated genes
  - Use chordal graphs as a filter for finding co-regulated structures

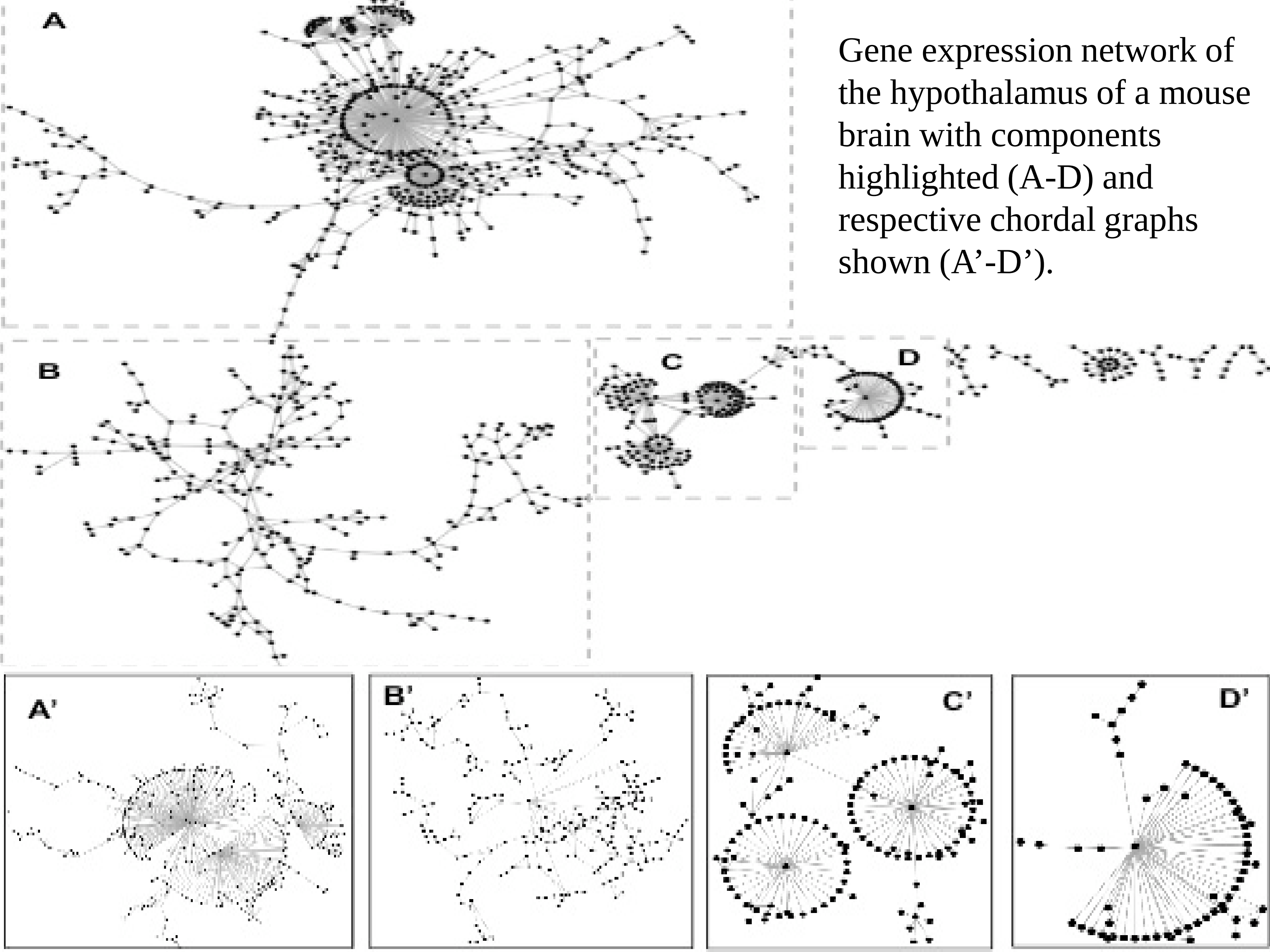


Subgraph formed by A,B,C is more likely to be biologically relevant.

If gene A and gene B are co-regulated, and if gene A and gene C are co-regulated, then genes B and C will be co-regulated.

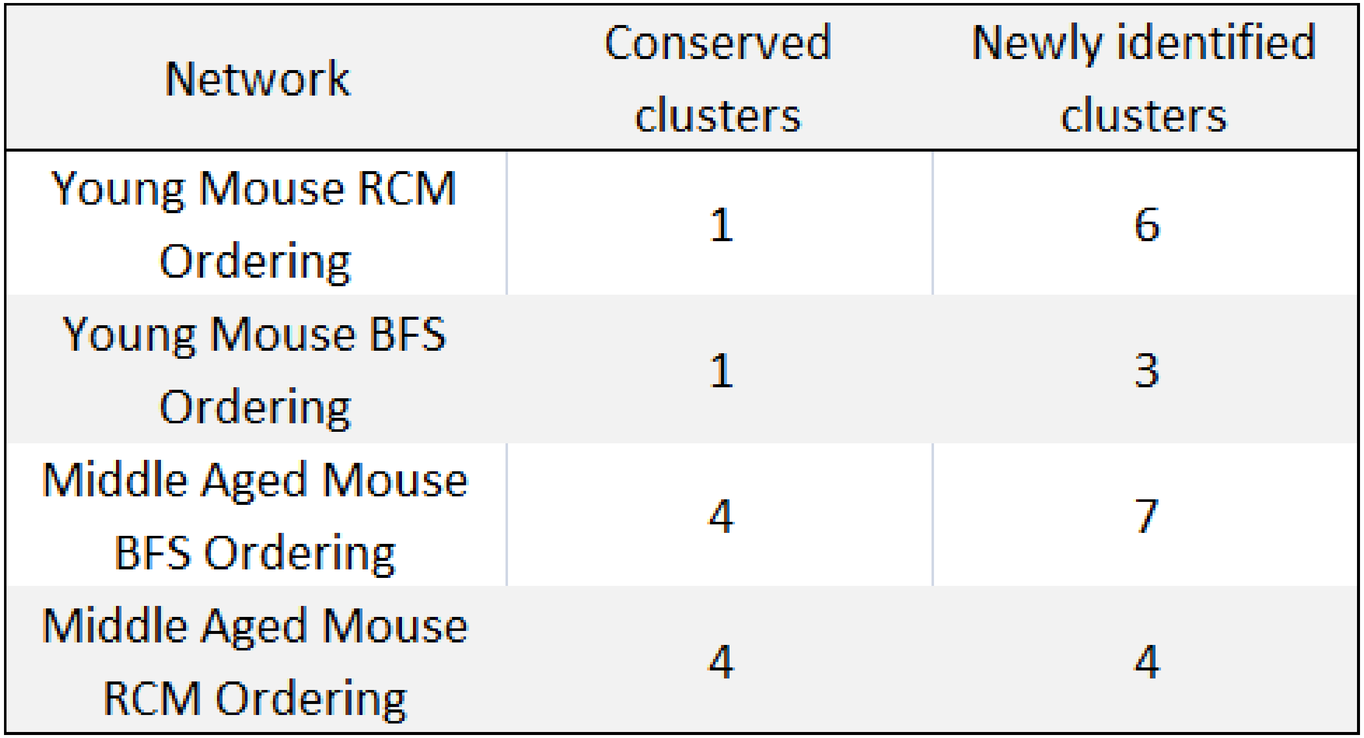
# Hypothesis

- Hypothesis $H_0$: Given a graph G representing a correlation network, maximal chordal subgraph $G_1$ will maintain most of the highly dense subgraphs of G while excluding edges representing noise-related relationships in the network.
  - $H_{0a}$ - Key functional properties found in the clusters of unfiltered networks G are maintained in the sampled networks $G_1$
  - $H_{0b}$ - New clusters with biological function are uncovered. Functional attributes previously lost in noise can now be identified.



Gene expression network of the hypothalamus of a mouse brain with components highlighted (A-D) and respective chordal graphs shown (A’-D’).

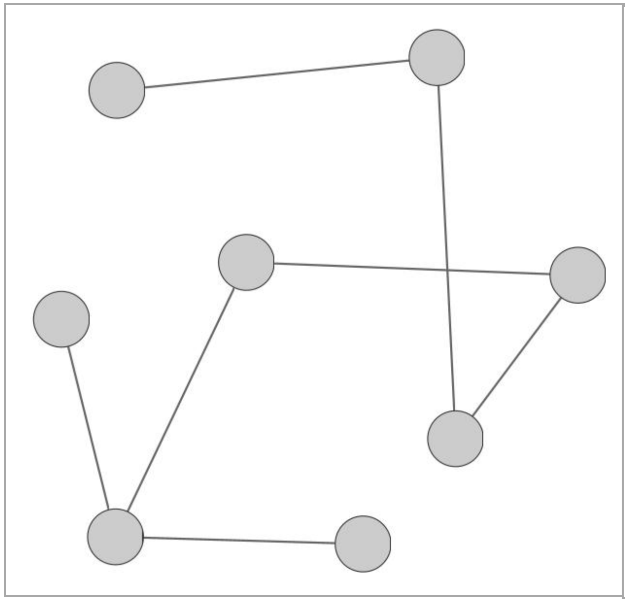
# Identification of New Clusters



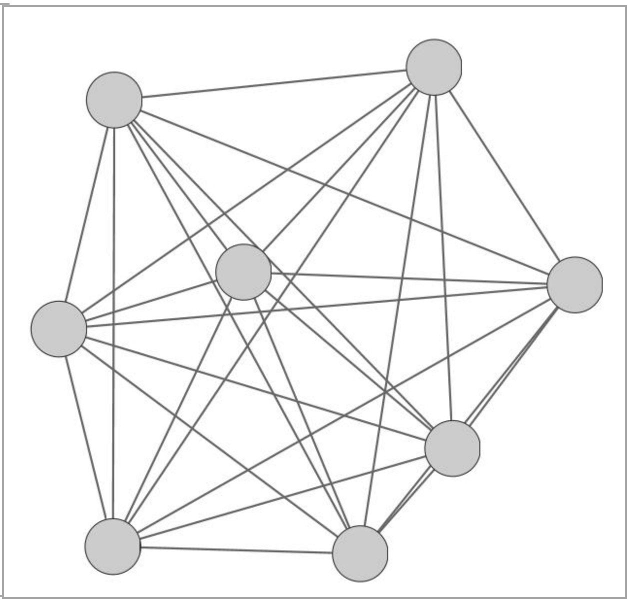
| Network | Conserved clusters | Newly identified clusters |
|---|---|---|
| Young Mouse RCM Ordering | 1 | 6 |
| Young Mouse BFS Ordering | 1 | 3 |
| Middle Aged Mouse BFS Ordering | 4 | 7 |
| Middle Aged Mouse RCM Ordering | 4 | 4 |

# Dynamic Filters

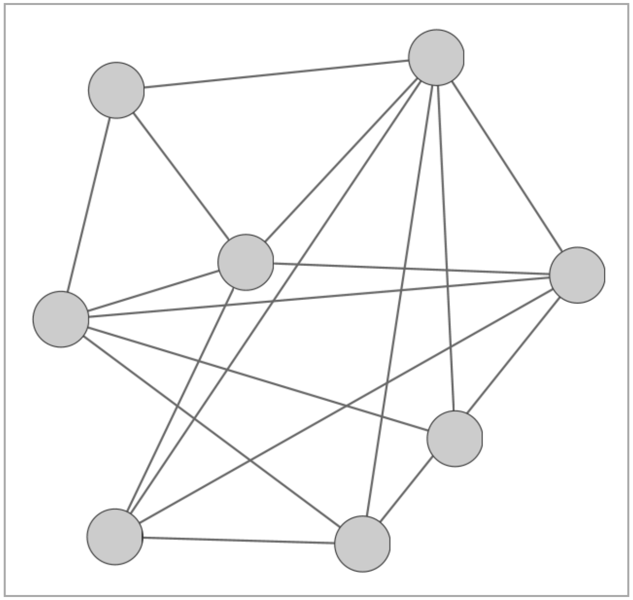
| Chordal-based filters | | | Tree-based filters | | |
|---|---|---|---|---|---|
| ***Filter*** | ***Name*** | ***Description*** | ***Filter*** | ***Name*** | ***Description*** |
| HD | High Degree | Traversal based on ascending order of vertices | ST | Spanning Tree | Tree determined by Prims Algorithm |
| LD | Low Degree | Traversal based on descending order of vertices | | | |



Spanning Tree



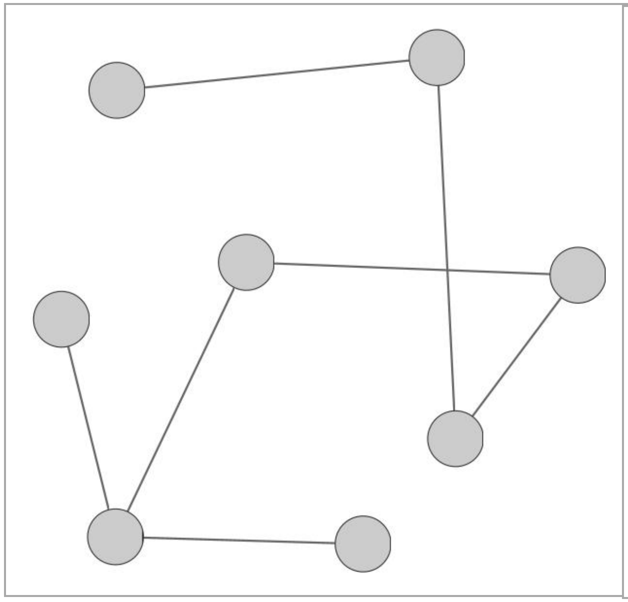
Original



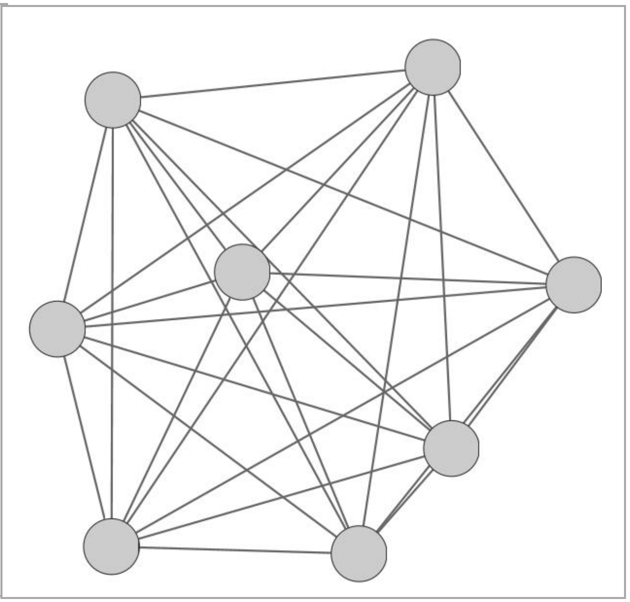
Chordal

# Other Filters

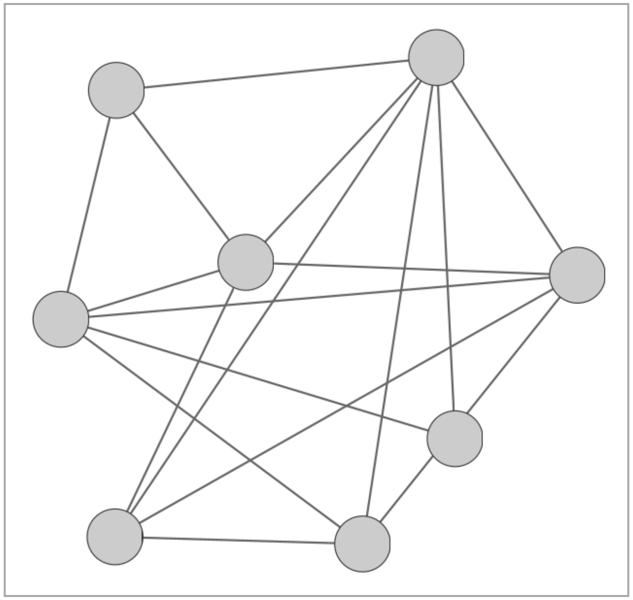
| Chordal-based filters | | | Tree-based filters | | |
|---|---|---|---|---|---|
| ***Filter*** | ***Name*** | ***Description*** | ***Filter*** | ***Name*** | ***Description*** |
| HD | High Degree | Traversal based on ascending order of vertices | ST | Spanning Tree | Tree determined by Prims Algorithm |
| LD | Low Degree | Traversal based on descending order of vertices | | | |



Spanning Tree

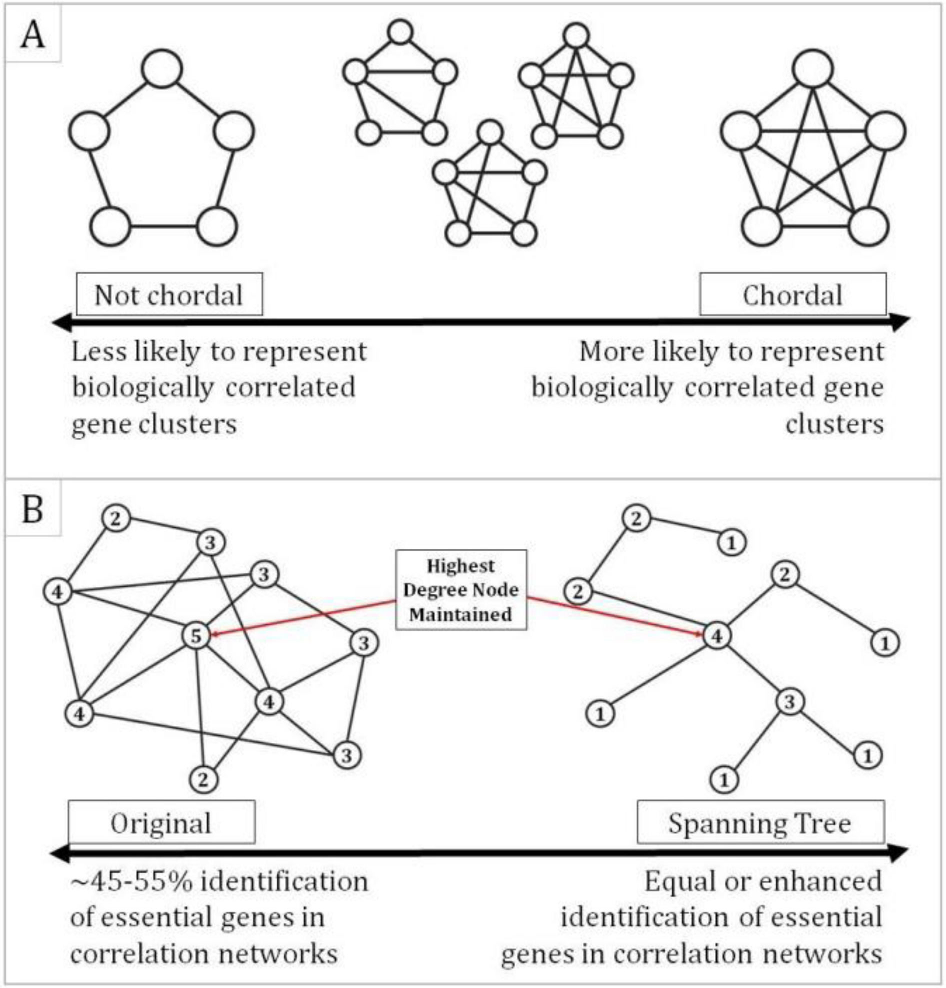


Original



Chordal

# A Combined Chordal Spanning Tree Model



Less likely to represent biologically correlated gene clusters

More likely to represent biologically correlated gene clusters

~45-55% identification of essential genes in correlation networks

Equal or enhanced identification of essential genes in correlation networks

# Tutorial Outlines

- Biomedical Informatics - Challenges and Opportunities
- Next Generation Bioinformatics Tools – Focus on Data Analysis and Integration
- Systems Biology and Network Analysis
- Biomedical Informatics and the Cloud: Models and Security
- Case Studies of Discoveries using Biological Networks
- HPC in Network Analysis of High Throughput Biological Data
- *Integration of different aspects of Biomedical Informatics*
- Next Steps – where to go from here?

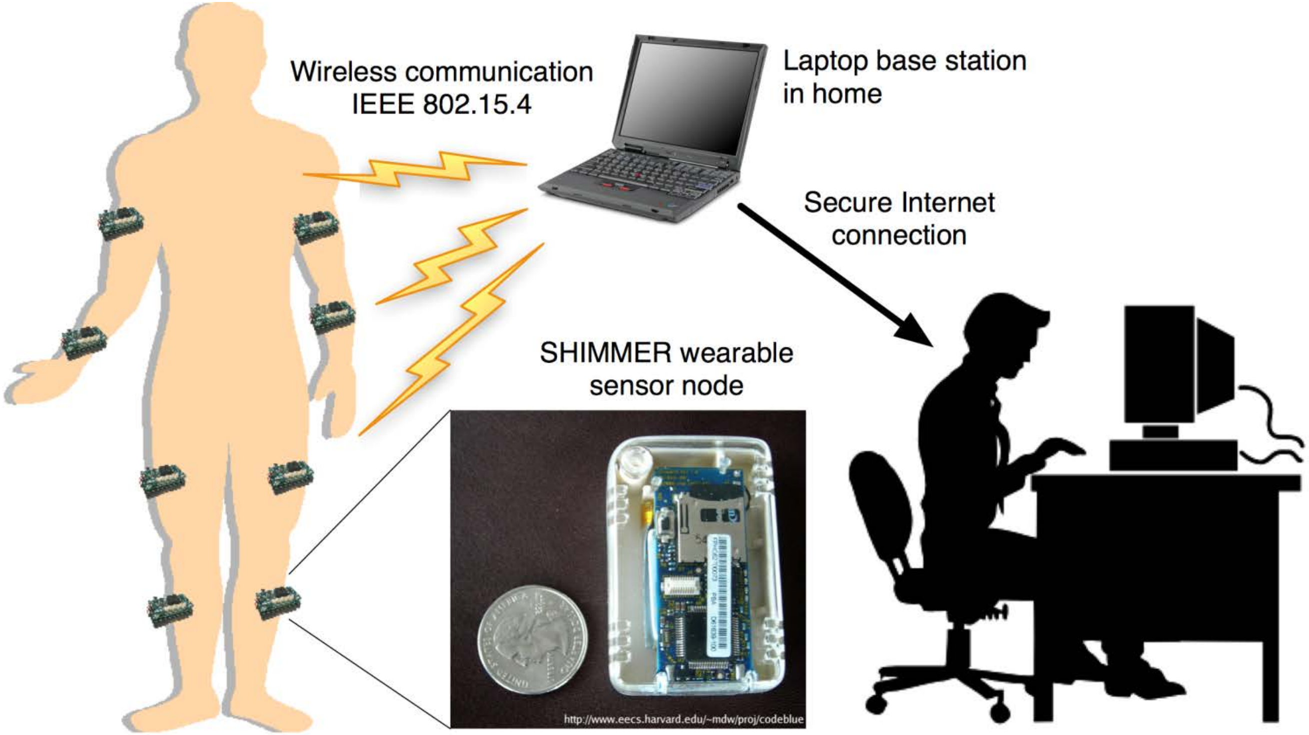
# Wireless Networks in Aging and Mobility

- Correlation between mobility and health level
- Monitoring mobility levels
- Aging of cells and aging of systems
- Collaboration between Bioinformatics group, Wireless Networks group and Decision Support Systems group

# Wireless Sensor Based Mobility Monitoring

- Inexpensive
- Comfortable

- High mobility
- Simple



http://fiji.eecs.harvard.edu/Mercury

# Goals of the Project

- Mobility Profile
  - Patient wearing a 3D-accelerometer will be monitored 24/7.
  - A complete mobility profile will be available for patients and care providers.

- Fall Prediction using Mobility Profiles
  - The system will identify anomalous movement and patterns that usually result in a fall or injury,
  - We would be able to take preemptive measures when such a pattern is detected, in order to reduce the occurrence of falls and prevent fall-related injuries.
  - We will develop an index that enables health care providers to determine how likely people are to fall.

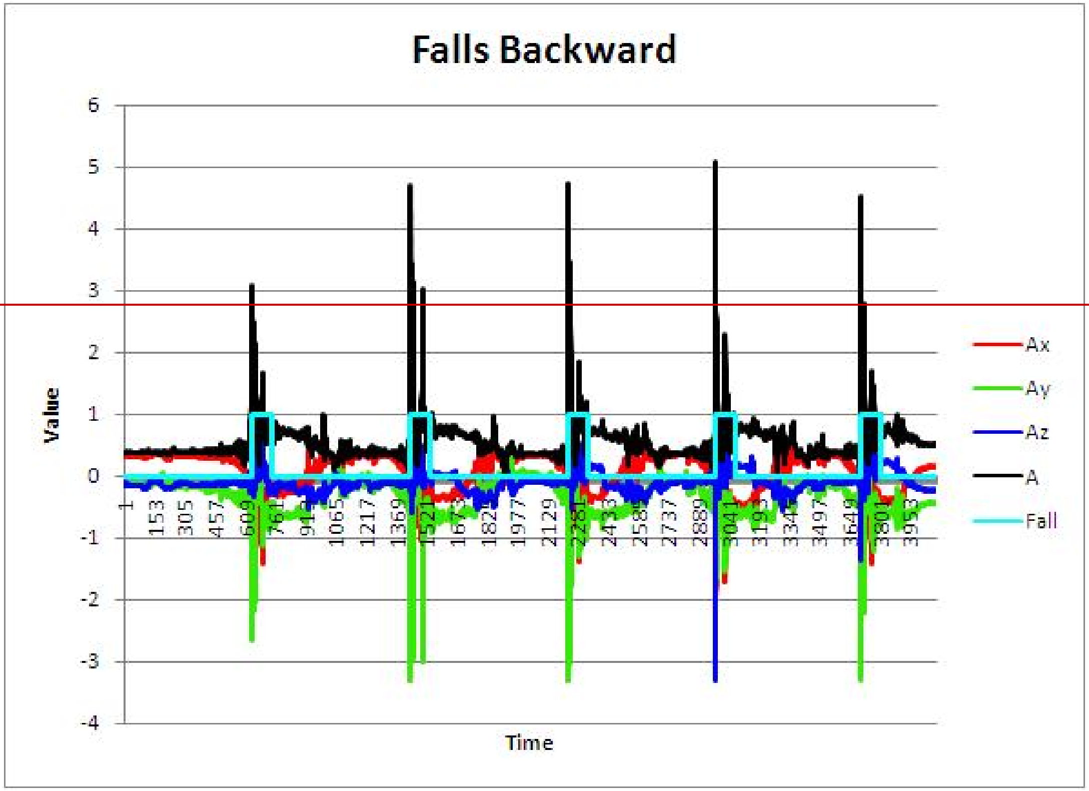
# Four Project Phases

- Four Phases:
  - Phase I: Fall Detection (completed)
    - achieved over 95% of fall detection rate
  - Phase II: Classification of ADLs (Activities of Daily Living, completed)
    - Running, Walking, Jumping, Stair Climbing, Standing, Sitting, and Lying.
  - Phase III: Construction of Mobility Profiles and Gait monitoring (in progress)
  - Phase IV: Fall Prediction based on mobility profiles (in progress by Bioinformatics group at UNO)

# Phase I: Fall Detection

## Accelerometer-based fall detection

- Determine an acceleration threshold.
- Detect fall.



Threshold

# Phase II: Classification of ADLs

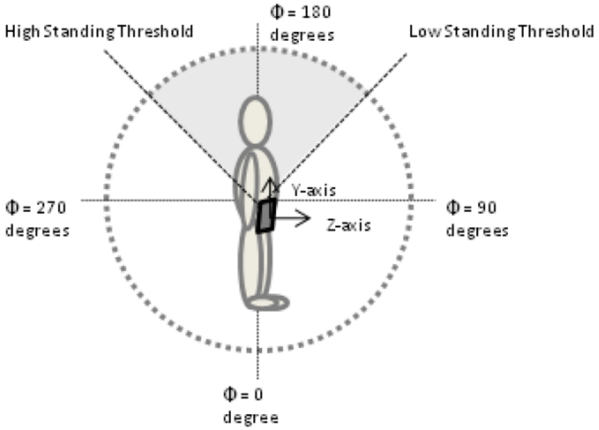
- Many Activities of Daily Living (ADLs) can be classified by analyzing the real-time acceleration data collected from sensors.

- Key Metrics
  - Inclination Angle
  - Standard deviation
  - Skewness
  - Signal Magnitude Area

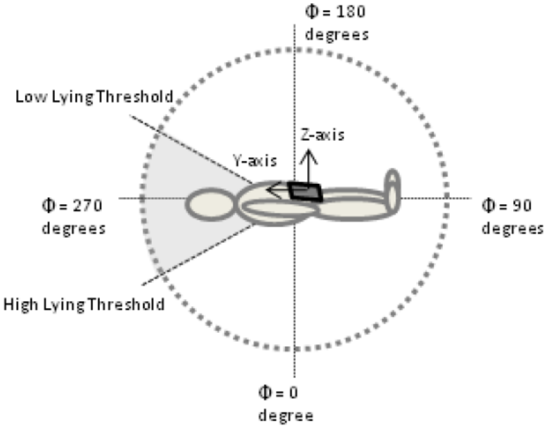
# Phase II: Classification of ADLs

- Many Activities of Daily Living (ADLs) can be classified by analyzing the real-time acceleration data collected from sensors.

- Data Processing: acceleration to metrics
  - Key Metrics
    - Inclination Angle
    - Standard deviation



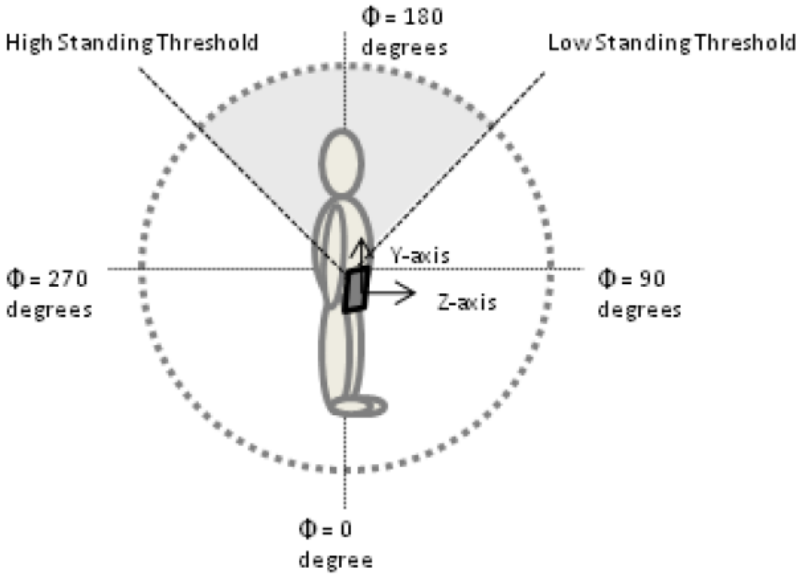
Standing Position



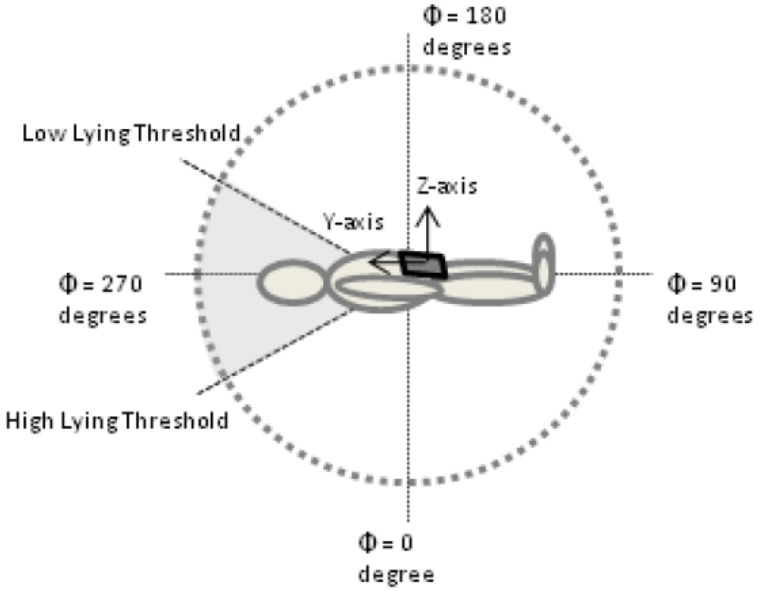
Lying Position

# Inclination Angle

- Inclination angle helps in determining posture.
  - Inclination angle is the measure between the x and y axis. It is assumed that if this value is around 180°, then the person would be standing.



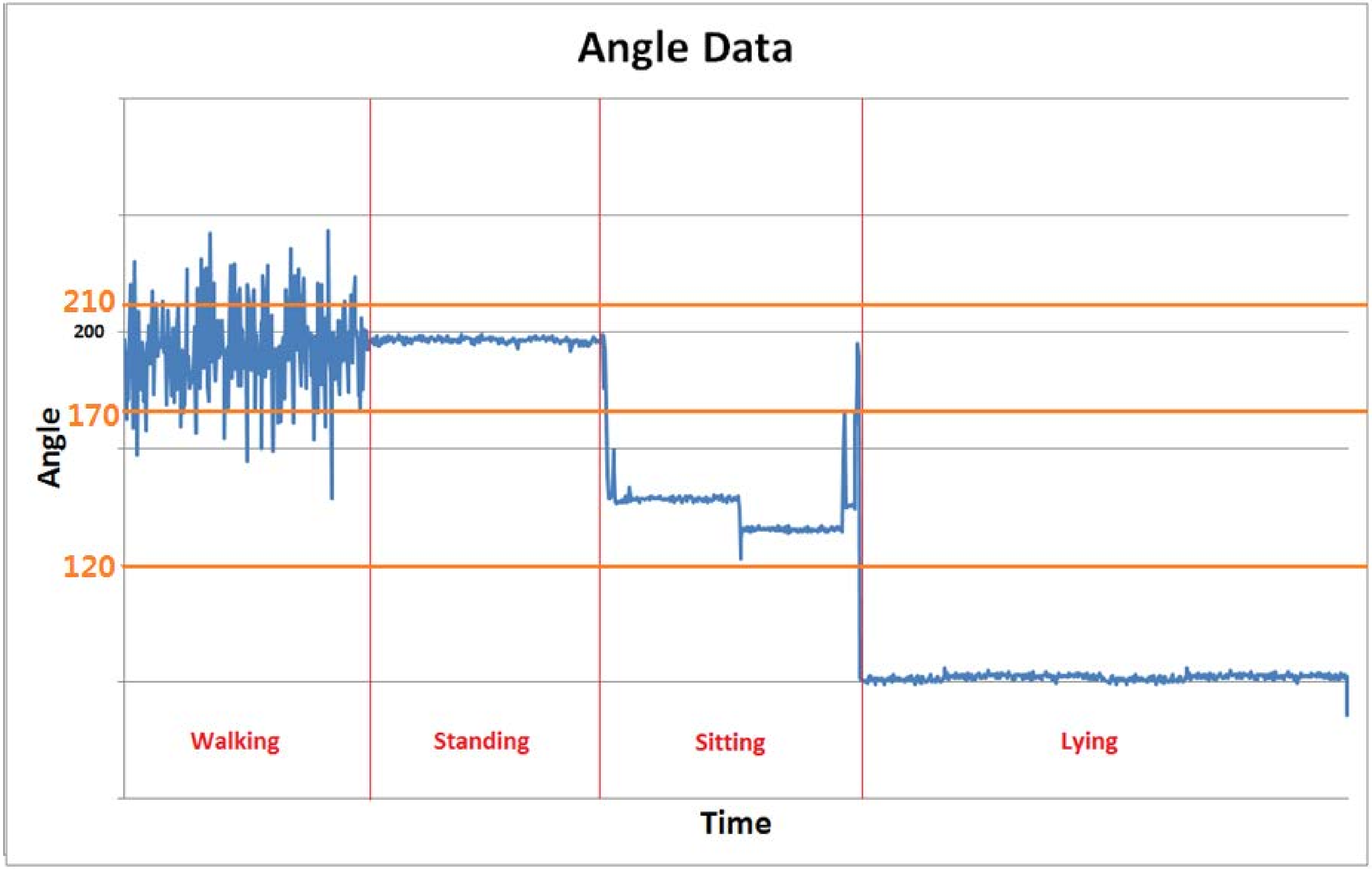
Standing Position



Lying Position

- Can now differentiate standing, sitting, and lying.

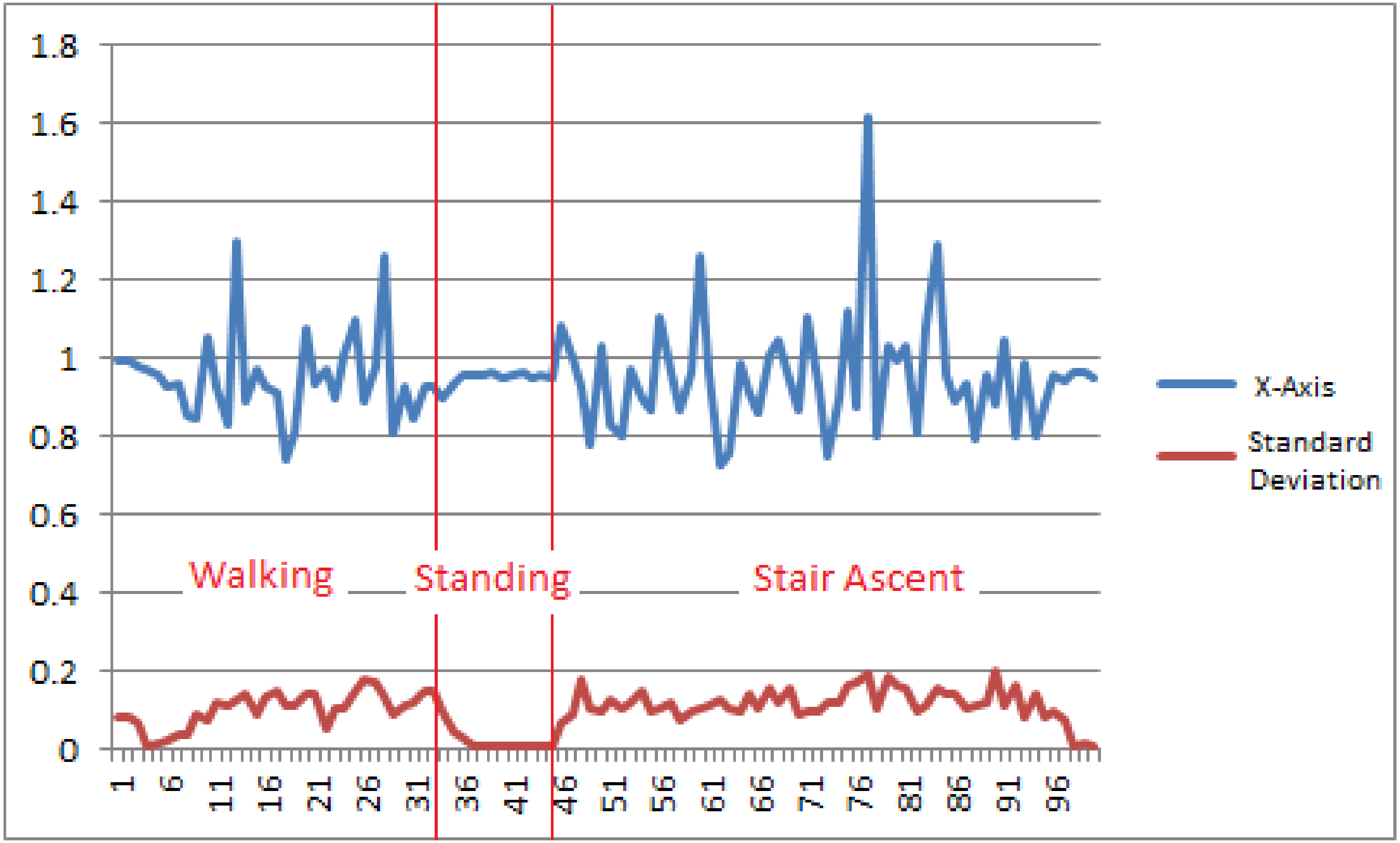
# Inclination Angle



# Standard Deviation

- Standard deviation of the x-axis acceleration helps in determining if the current mobility state is dynamic or static
  - Standard deviation measures the variability of data from the mean. Dynamic data will have measurably more variability than static data.
- When used with angle measurement, can now differentiate standing from walking/running.

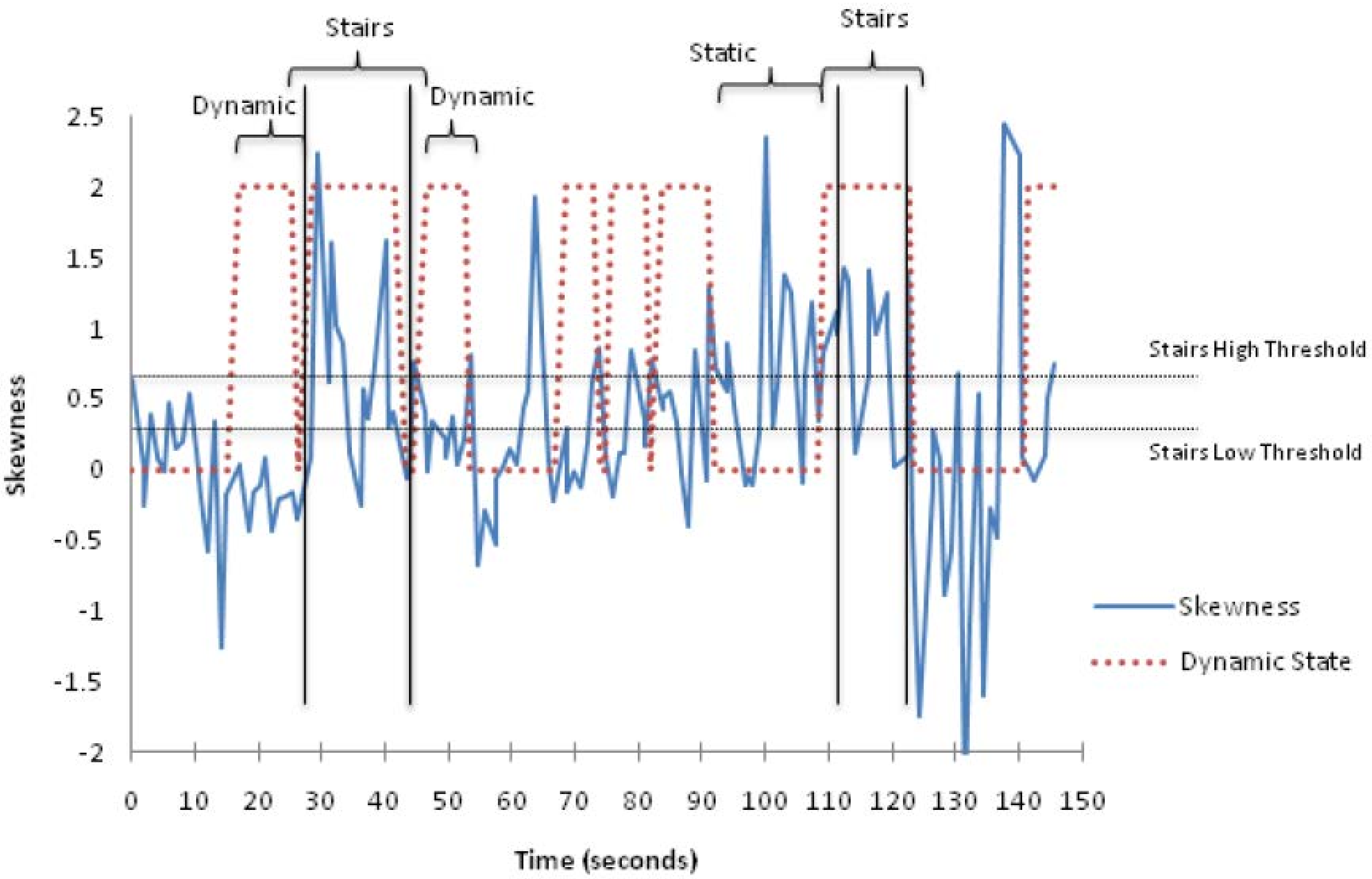
# Standard Deviation



# Skewness

- Skewness helps in determining if the current mobility state is going up or down stairs.
  - Skewness measures the asymmetry of the distribution of x-axis acceleration values. It is assumed that going up/down stairs will produce data that has greater asymmetry than walking.
- When used with angle and standard deviation, can now differentiate walking/running from going up/down stairs.

# Skewness

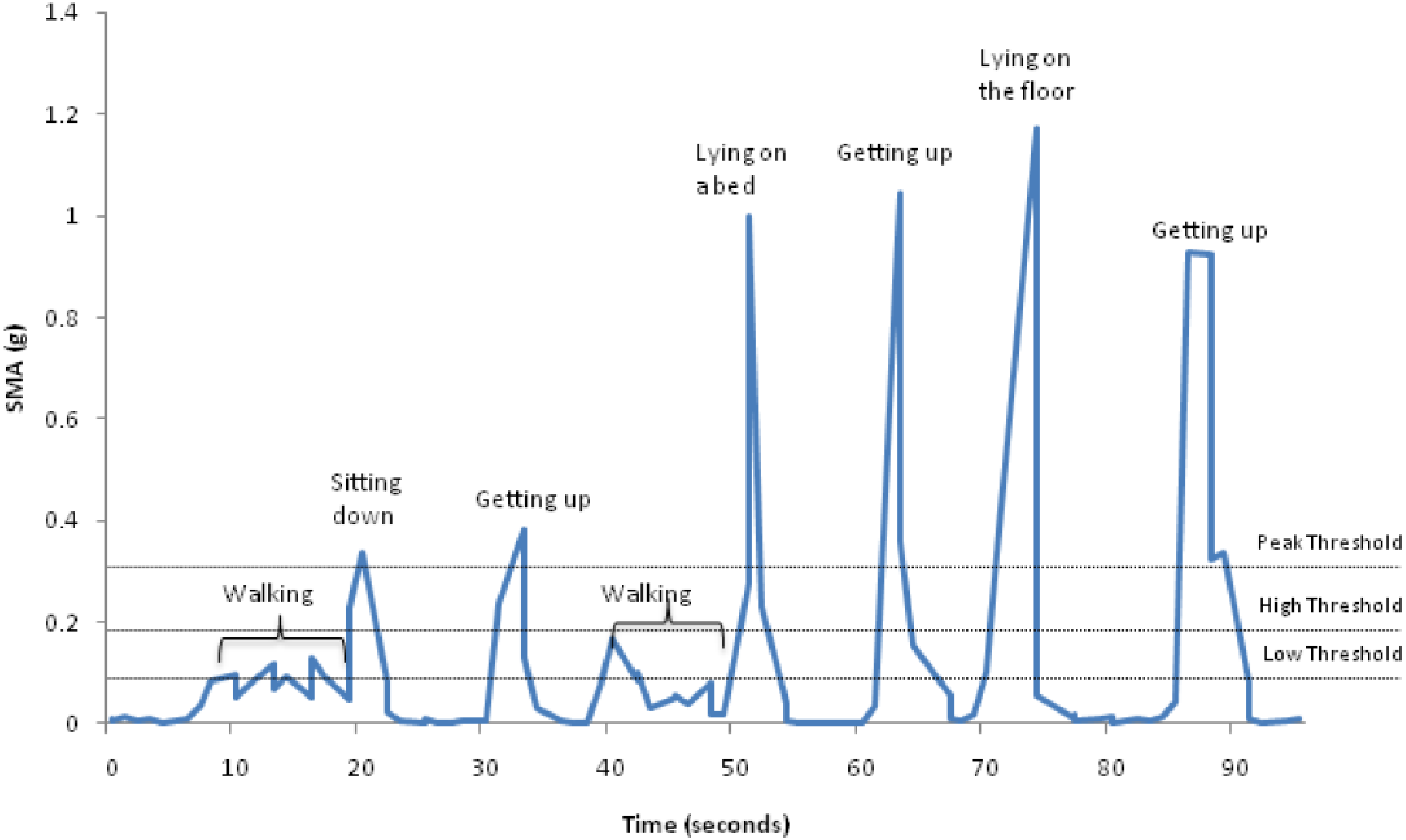


G. Hache, E. Lemaire, N.Baddour, “Development of wearable mobility Monitoring System, “ in Proc. Can. Med. Biological Eng. Conf., Calgary, Canada, May 2009.

# Signal Magnitude Area

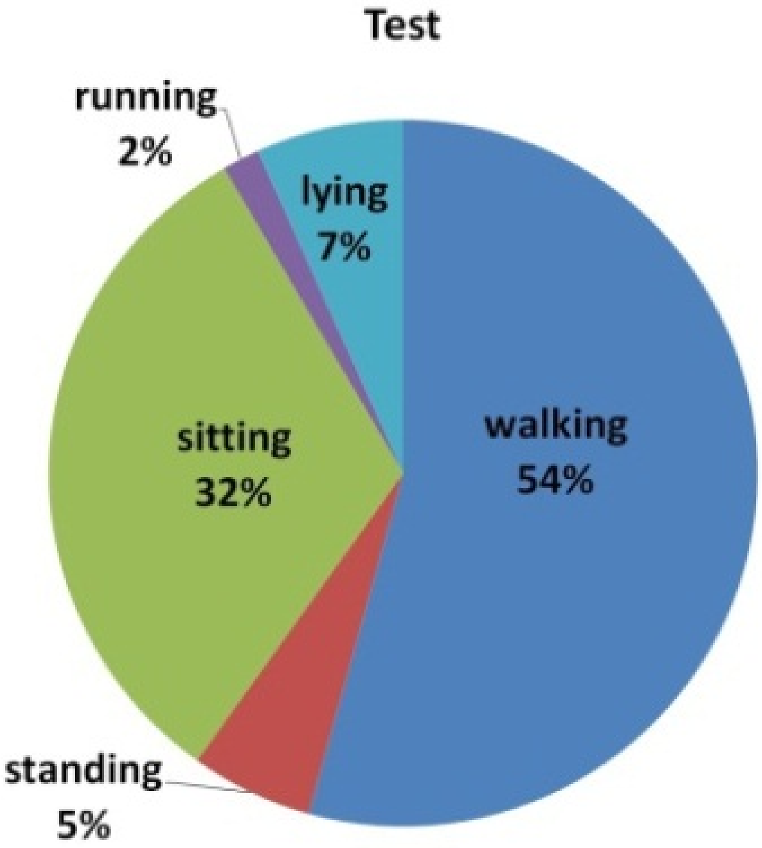
- Measures amplitude and duration variation in the acceleration signal.
  - Assumed that amplitude and duration variation will be greater when the intensity of the activity changes. As such, SMA values will be greater when changing state as opposed to not changing state. Ex: getting out of bed vs. walking
- When used with the prior four measurements, SMA will help differentiate transitions from other dynamic mobility states.

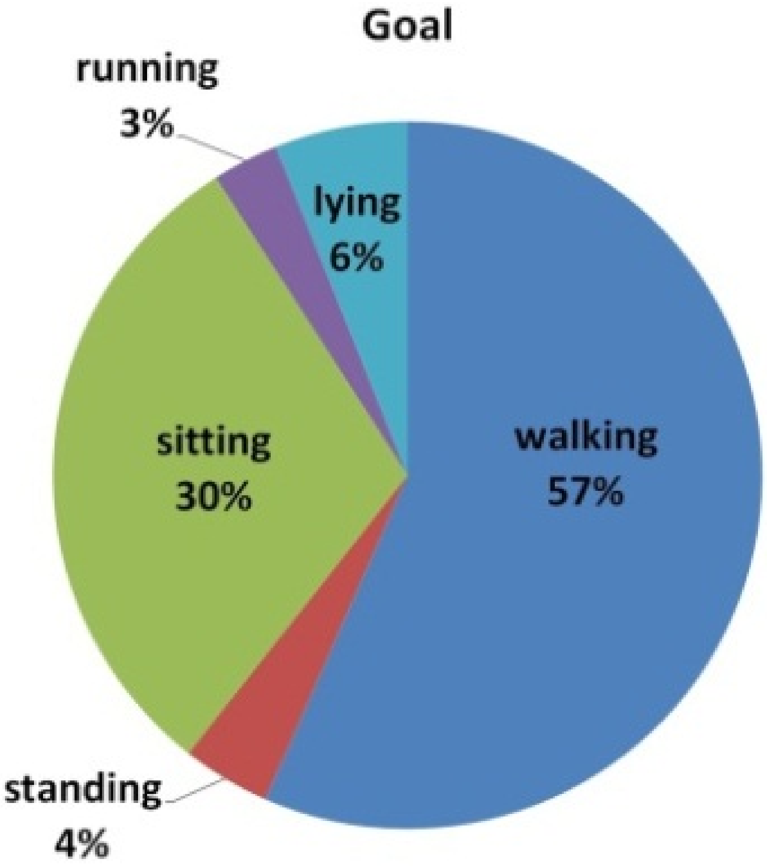
## Signal Magnitude Area (SMA) of acceleration signals versus Time



G. Hache, E. Lemaire, N.Baddour, “Development of wearable mobility Monitoring System, “ in Proc. Can. Med. Biological Eng. Conf., Calgary, Canada, May 2009.

# Phase III: Profiles





# Fall Prediction

Current system can detect falls, next step is to predict falls

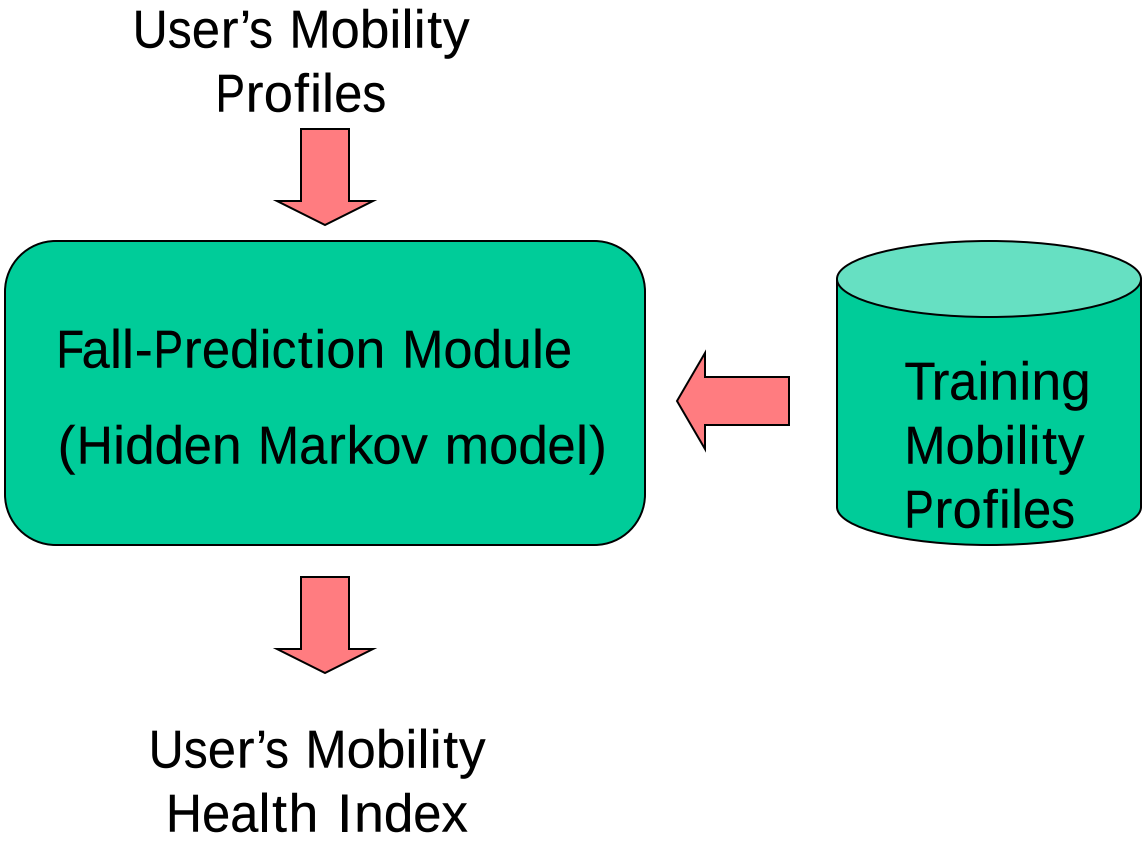
## Fall Detection:

Approach: Using GSM, detect certain gait patterns associated with people who fall.

–Is a sudden change in gait predictive of a likely fall?

–Can the system be used to detect early signs of deterioration and/or improvement?

# Phase IV: Fall Prediction



User’s Mobility Profiles

Fall-Prediction Module

(Hidden Markov model)

Training Mobility Profiles

User’s Mobility Health Index

# Tutorial Outlines

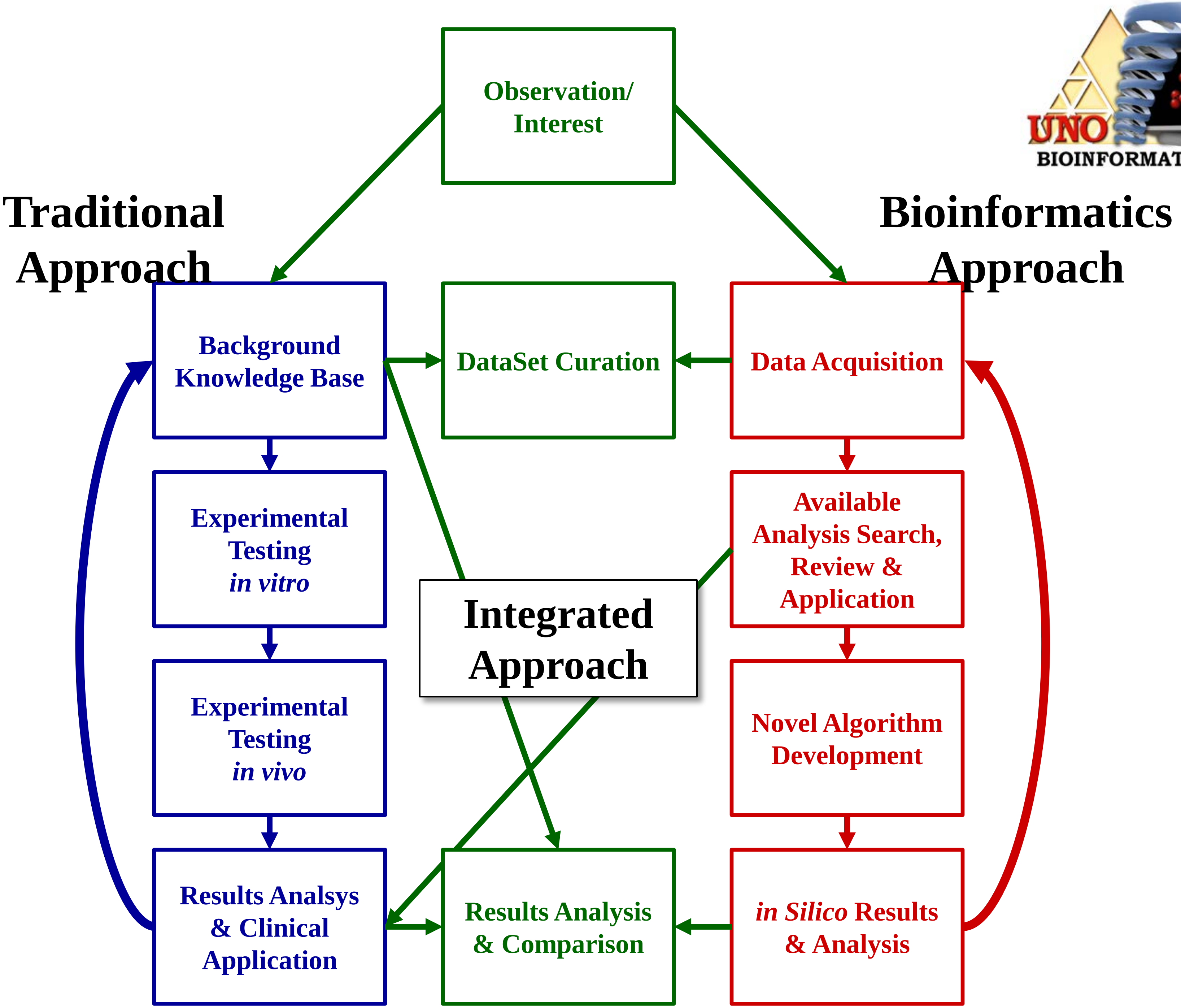
- Biomedical Informatics - Challenges and Opportunities
- Next Generation Bioinformatics Tools – Focus on Data Analysis and Integration
- Systems Biology and Network Analysis
- Biomedical Informatics and the Cloud: Models and Security
- Case Studies of Discoveries using Biological Networks
- HPC in Network Analysis of High Throughput Biological Data
- Integration of different aspects of Biomedical Informatics
- *Next Steps – where to go from here?*

# Data-Driven Decisions

- 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

Next Generation Tools: ICD Tools

UNO BIOINFORMATICS



Traditional Approach

Bioinformatics Approach

Observation/ Interest

Background Knowledge Base

DataSet Curation

Data Acquisition

Experimental Testing *in vitro*

Available Analysis Search, Review & Application

Integrated Approach

Experimental Testing *in vivo*

Novel Algorithm Development

Results Analsys & Clinical Application

Results Analysis & Comparison

*in Silico* Results & Analysis

# Next Generation Tools

- Next Generation Bioinformatics Tools need to be Intelligent, Collaborative, and Dynamic
- Biomedical scientists, Bioinformatics researchers and computer scientists need to work together to best utilize the combination of tools development and domain expertise
- HPC is critical to the success of the next phase of Biomedical research but again the integration needs to happen at a deeper level
- The outcome of collaboration has the potential of achieving explosive results with significant impact on human health and overall understanding of biological mysteries

# BMI at Crossroads

- Many Scientific disciplines are now at crossroads
- The proper penetration of IT represent tremendous challenges and great opportunities
- Discovery are likely to take place at many places
- The importance of interdisciplinary approach to problem solving
- This may lead to scientific revolution

# 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