Biological Gene and Protein Networks and Their Place in Downstream Analysis

WHY WOULD WE USE NETWORKS?

BOTTOM LINE UP-FRONT:

PRIOR BIOLOGICAL KNOWLEDGE GREATLY FACILITATES THE MEANINGFUL INTERPRETATION OF GENE-EXPRESSION DATA. CAUSAL NETWORKS CONSTRUCTED FROM INDIVIDUAL RELATIONSHIPS CURATED FROM THE LITERATURE ARE PARTICULARLY SUITED FOR THIS TASK, SINCE THEY CREATE MECHANISTIC HYPOTHESES THAT EXPLAIN THE EXPRESSION CHANGES OBSERVED IN DATASETS.

Biological Networks

Gene regulatory network: two genes are connected if the expression of one gene modulates expression of another one by either activation or inhibition

Protein interaction network: proteins that are connected in physical interactions or metabolic and signaling pathways of the cell;

Metabolic network: metabolic products and substrates that participate in one reaction;

Background Knowledge

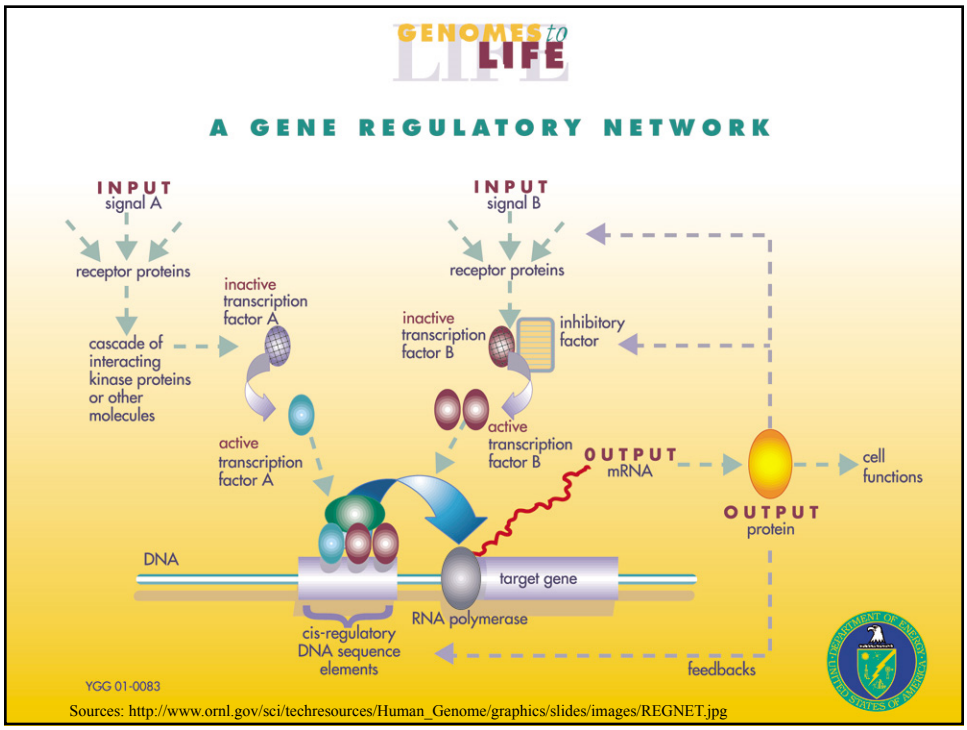
- Cell reproduction, metabolism, and responses to the environment are all controlled by proteins;
- Each gene is responsible for constructing a single protein;
- Some genes manufacture proteins which control the rate at which other genes manufacture proteins (either promoting or suppressing);
- Hence some genes regulate other genes (via the proteins they create) ;

What is Gene Regulatory Network?

Gene regulatory networks (GRNs) are the on-off switches of a cell operating at the gene level.

Two genes are connected if the expression of one gene modulates expression of another one by either activation or inhibition

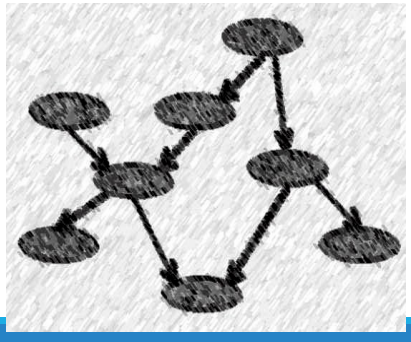
An example.



Sources: http://www.ornl.gov/sci/techresources/Human_Genome/graphics/slides/images/REGNET.jpg

Simplified Representation of GRN

A gene regulatory network can be represented by a directed graph;



- **Node** represents a gene;
- **Directed edge** stands for the modulation (regulation) of one node by another:
  - e.g. arrow from gene X to gene Y means gene X affects expression of gene Y

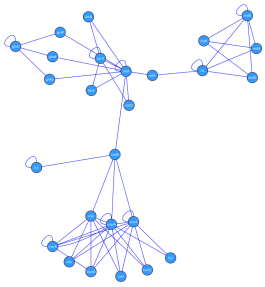
Directed vs. Undirected

Nodes 

- DNA / RNA / Protein / Metabolite / Ontology

Edges 

- Directed
- Distinction between source and target
  - Activation (direct/indirect) 
  - Repression (direct/indirect)
- Undirected
  - No distinction between source and target
    - Co-expression (indirect)
    - Binding (direct)
    - Similarity/strength



Basic Features

Degree

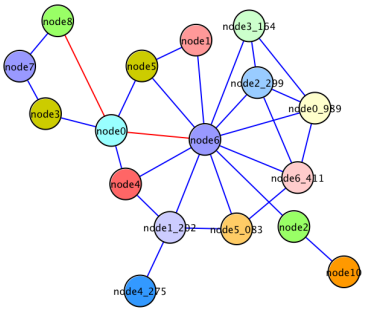
- Number of connections that a node has

Distance

- Number of connections between two nodes,

Path

- A sequence of connections
- Is there a path (reachability)
- Mean Shortest Path distance (closeness)
- In how many shortest paths (betweenness)

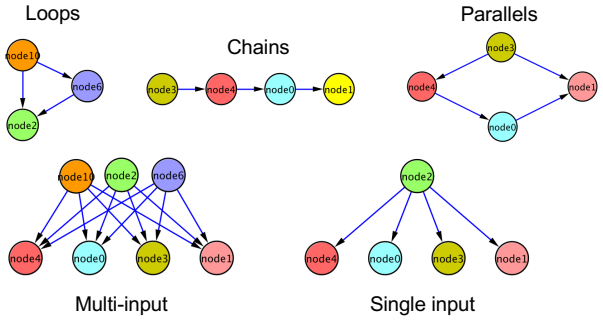


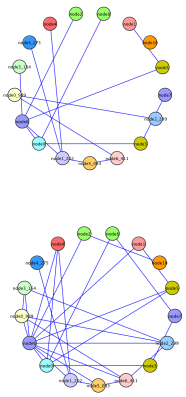
Basic Features

Size of a network (Number of nodes)

Density of a network (Proportion of the connections)

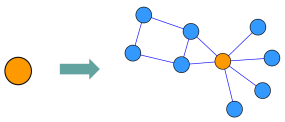
Motifs/Cliques/Clusters/Sub-networks



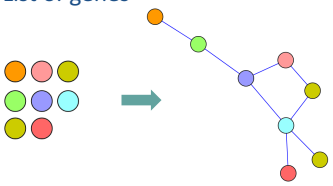


Basic Features

Single gene



List of genes



- Regulators/Co-regulators
- Upstream/Downstream elements in the network
- Global connectivity/interconnectivity
- Functional features
- Differentially expressed subnetworks
- One gene – one disease : bunch of genes – pathways
- Nextgen sequencing data
- Meta-analysis

Why Study GRN?

Genes are **not independent**;

- They regulate each other and act collectively;
- This collective behavior can be observed using microarray;

Some genes **control the response** of the cell to changes in the environment by regulating other genes;

Potential discovery of **triggering mechanism** and **treatments for disease**;

Modeling Gene Regulatory Networks

- Linear Model;
- Bayesian Networks;
- Differential Equations;
- **Boolean Network**
  - Originally introduced by Kauffman (1969)
  - Boolean network is a kind of **Graph**
    - G(V, F) – V is a set of **nodes** (genes) as $x_1, x_2, \ldots, x_n$ F is a list of **Boolean functions** $f(x_1, x_2, \ldots, x_n)$
  - Gene expression is quantized to only two level:
    - 1 (**On**) and 0 (**OFF**);
  - Every function has the result value of each node;

Boolean Network Example

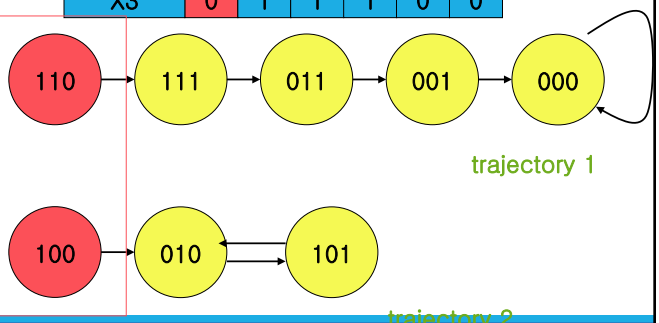
Nodes (genes)		
x1	0	1
x2	0	1
x3	0	1

$f_1(x_1, x_2, x_3) = x_2 \wedge \overline{x_3}$

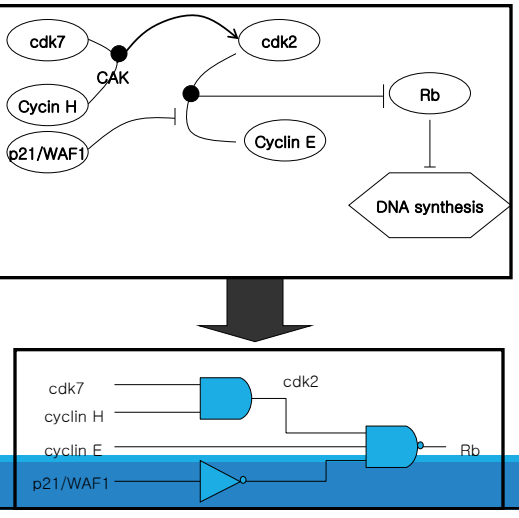
$f_2(x_1, x_2, x_3) = x_1$

$f_3(x_1, x_2, x_3) = x_2$

Iteration	1	2	3	4	5	6
X1	1	1	0	0	0	0
X2	1	1	1	0	0	0
X3	0	1	1	1	0	0



Boolean Network as Models of Gene Regulatory Networks



- **Cyclin E and cdk2 work together to phosphorylate the Rb protein and inactivate it**
- **Cdk2/Cyclin E is regulated by two switches:**
  - **Positive switch complex called CAK;**
  - **Negative switch P21/WAF1;**
- **The CAK complex can be composed of two gene products:**
  - **Cyclin H;**
  - **Cdk7**
- **When cyclin H and cdk7 are present, the complex can activate cdk2/cyclin E.**

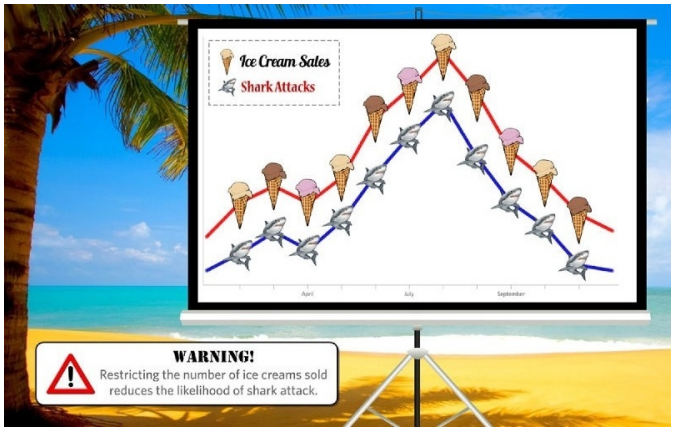
Learning Causal Relationships

High-throughput genetic technologies empowers to study **how genes interact with each other**;

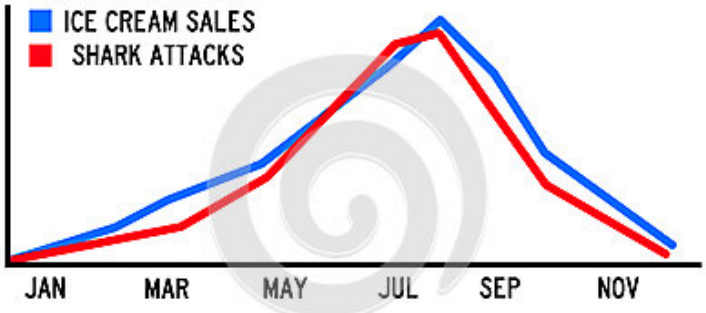
Learning gene causal relationship is important:

- Turning on a gene can be achieved directly or through other genes, which have causal relationship with it.

Clearing up Confusion Between Correlation and Causation

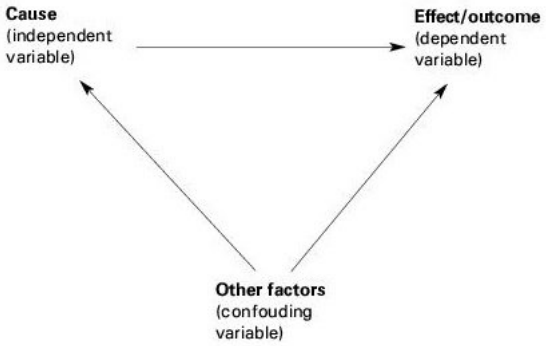


CORRELATION IS NOT CAUSATION!

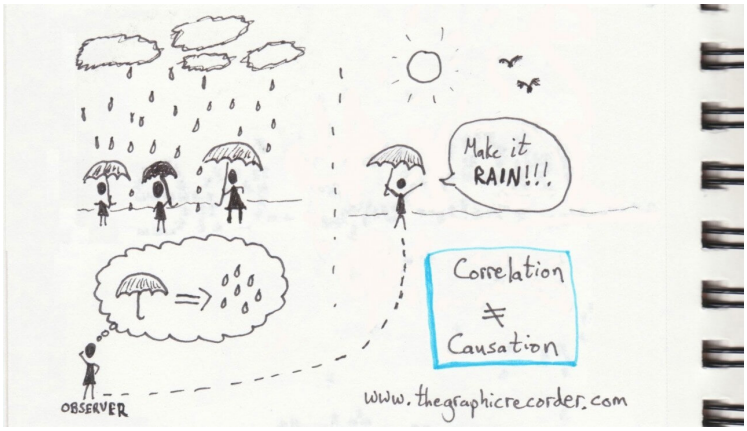


Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea).

Prior knowledge greatly facilitates the meaningful interpretation of data.



Upstream Regulator is the Weather, Downstream Effect is Increased Umbrella Use



Causality vs. Correlation

Example: ***rain*** and ***falling_barometer***

- Observed that they are either both true or both false, so they are related. Then write

rain = falling_barometer

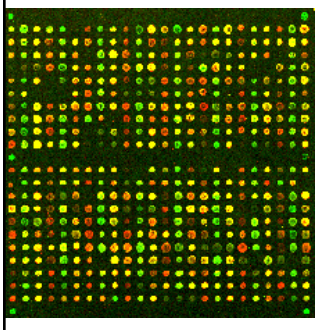
  - Neither *rain* causes *falling_barometer* nor vice-versa.
  - Thus if one wanted *rain* to be true, one could not achieve it by somehow forcing *falling_barometer* to be true. This would have been possible if *falling_barometer* caused *rain*.

- We say that the relationship between *rain* and *falling_barometer* is **correlation**, but **not cause**.

Learning Causal Relationship with Steady State Data

- How to infer causal relationship?
  - In wet-labs, **knocking down** the possible subsets of a gene;
  - Use **time series** gene expression data;
- Problem?
  - Human tissues gene expression data is only available in the **steady state observation**;
- (IC) algorithm by Pearl et al to infer causal information but not in biological domain;

RNAseq Data Treated the Same as Microarray Data



Samples

Genes

Gene Description	1	2	3	4	5	6	7
AFFX-BioB-5_at (endogenous control)	-214	-139	-76	-135	-106	-138	-72
AFFX-BioB-M_at (endogenous control)	-153	-73	-49	-114	-125	-85	-144
AFFX-BioB-3_at (endogenous control)	-58	-1	-307	265	-76	215	238
AFFX-BioC-5_at (endogenous control)	88	283	309	12	168	71	55
AFFX-BioC-3_at (endogenous control)	-295	-264	-376	-419	-230	-272	-399
AFFX-BioDn-5_at (endogenous control)	-558	-400	-650	-585	-284	-558	-551
AFFX-BioDn-3_at (endogenous control)	199	-330	33	158	4	67	131
AFFX-CreX-5_at (endogenous control)	-176	-168	-367	-253	-122	-186	-179
AFFX-CreX-3_at (endogenous control)	252	101	206	49	70	87	126
AFFX-BioB-5_st (endogenous control)	206	74	-215	31	252	193	-20
AFFX-BioB-M_st (endogenous control)	-41	19	19	363	155	325	-115
AFFX-BioB-3_st (endogenous control)	-831	-743	-1135	-934	-471	-631	-1003
AFFX-BioC-5_st (endogenous control)	-653	-239	-962	-577	-490	-625	-761
AFFX-BioC-3_st (endogenous control)	-462	-83	-232	-214	-184	-177	-541
AFFX-BioDn-5_st (endogenous control)	75	182	208	142	32	-94	109
AFFX-BioDn-3_st (endogenous control)	381	164	432	271	213	222	435
AFFX-CreX-5_st (endogenous control)	-118	-141	84	-107	1	-1	-129
AFFX-CreX-3_st (endogenous control)	-565	-423	-501	-101	-260	-140	-399
hum_alu_at (miscellaneous control)	15091	11038	16692	15763	18128	34207	30801
AFFX-DapX-5_at (endogenous control)	7	37	183	45	-28	65	43
AFFX-DapX-M_at (endogenous control)	311	134	378	268	118	154	80
AFFX-DapX-3_at (endogenous control)	-231	-161	-221	-27	-153	-49	-87

Gene up-regulate, down-regulate;

How We Study Gene Causal Network?

We present an algorithm for **learning causal relationship with knowledge of topological ordering** information;

- Studying conditional dependencies and independencies among variables;
- Learning mutual information among genes;
- Incorporating topological information;

How we Study Gene Causal Network?

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 30 no. 4 2014, pages 523–530
doi:10.1093/bioinformatics/btt703

Systems biology

Advance Access publication December 13, 2013

Causal analysis approaches in Ingenuity Pathway Analysis

Andreas Krämer[1,*], Jeff Green[1], Jack Pollard, Jr[2] and Stuart Tugendreich[1]

[1]Ingenuity Systems, 1700 Seaport Boulevard, Redwood City, CA and [2]Translational and Experimental Medicine—Bioinformatics, Sanofi-Aventis, 270 Albany Street, Cambridge, MA, USA

Associate Editor: Jonathan D. Wren

ABSTRACT

Motivation: Prior biological knowledge greatly facilitates the meaningful interpretation of gene-expression data. Causal networks constructed from individual relationships curated from the literature are particularly suited for this task, since they create mechanistic hypotheses that explain the expression changes observed in datasets.

Results: We present and discuss a suite of algorithms and tools for inferring and scoring regulator networks upstream of gene-expression data based on a large-scale causal network derived from the Ingenuity Knowledge Base. We extend the method to predict downstream effects on biological functions and diseases and demonstrate the validity of our approach by applying it to example datasets.

Availability: The causal analytics tools 'Upstream Regulator Analysis', 'Mechanistic Networks', 'Causal Network Analysis' and 'Downstream Effects Analysis' are implemented and available within Ingenuity

effects on cellular and organismal biology. It is critical to infer the identity of upstream regulatory molecules and associated mechanisms to provide biological insight to the observed expression changes. We also aim to predict whether such regulators are activated or inhibited based on the distinct up- and down-regulation pattern of the expressed genes, and determine which causal relationships previously reported in the literature are likely relevant for the biological mechanisms underlying the data. Upstream regulators are not limited to transcription factors; they can be any gene or small molecule that has been observed experimentally to affect gene expression in some direct or indirect way. A similar approach, relying on the same methodology is also used to predict downstream functional effects and phenotypes. Apart from generating likely mechanistic hypotheses, causal inference can also be used to find potential

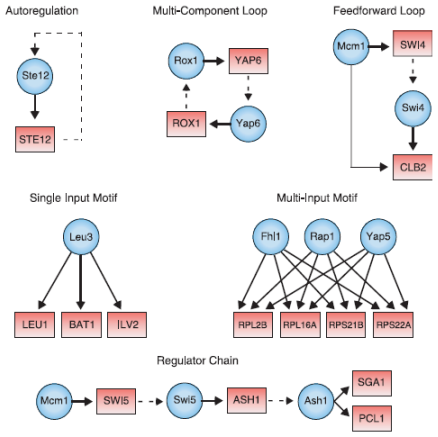
Learning from Multiple Data Sources

We have gene expression data and topological ordering information;

Incorporating some other data sources as prior knowledge for the learning;

- Transcription factor binding location data;
- …

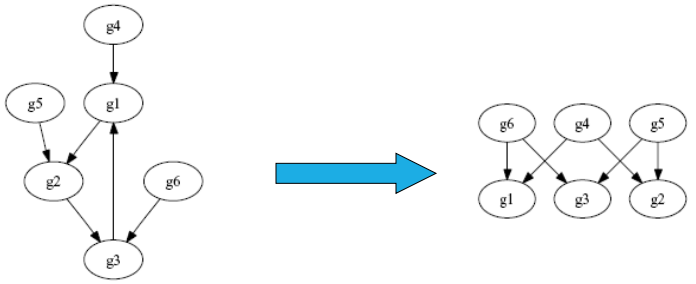
Learning Causality in Motifs



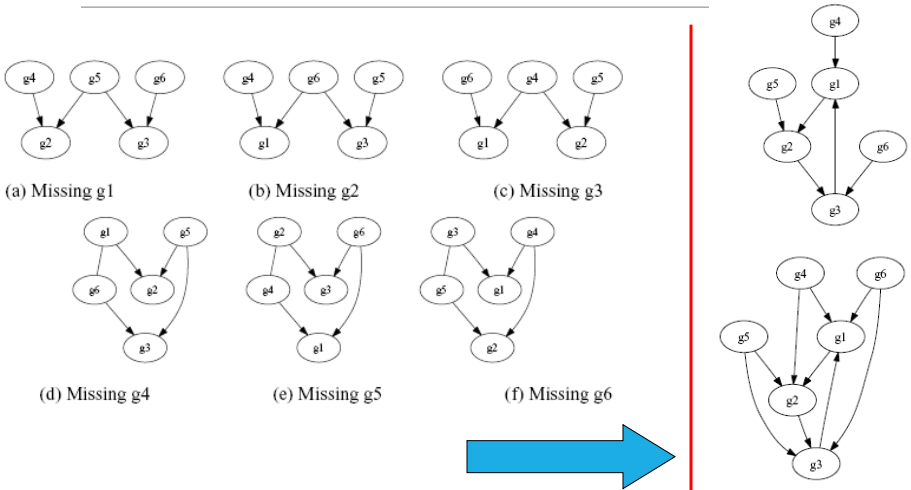
•Network motifs are the simplest units of network architecture.

•They be used to assemble a transcriptional regulatory network.

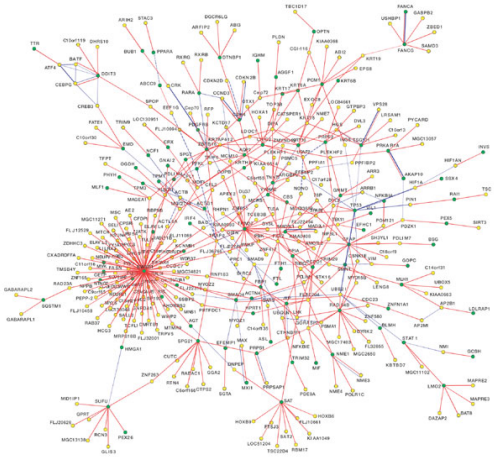
Learning GRN with Feedback Loops



Learning GRN with Feedback Loops (Con'd)



Protein-Protein Interactions



From:
Towards a proteome-scale map of the human protein–protein interaction network
Rual, Vidal et al. Nature **437**, *1173-1178 (2005)*

Why Study Protein-Protein Interactions

Most proteins **perform functions** by interacting with other proteins;

Broader view of how they **work cooperatively** in a cell;

Studies indicate that many **diseases are related** to subtle molecular events such as protein interactions;

Beneficial for the process of **drug design**.

Reference Databases

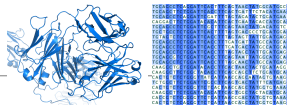
Interactions
- MIPS
- DIP
- YPD
- Intact (EBI)
- BIND/ Blueprint
- GRID
- MINT

Prediction server
- Predictome (Boston U)
- Plex (UTexas)
- STRING (EMBL)

Protein complexes
- MIPS
- YPD

How to Study PPI?

- High-throughput data
  - Two-hybrid systems
  - Mass Spectrometry
  - Microarrays
- **Genomic data**
  - Phylogenetic profile
  - Rosetta Stone method
  - Gene neighboring
  - Gene clustering
- Other Data Sources



NGS : DOWNSTREAM PROCESSING

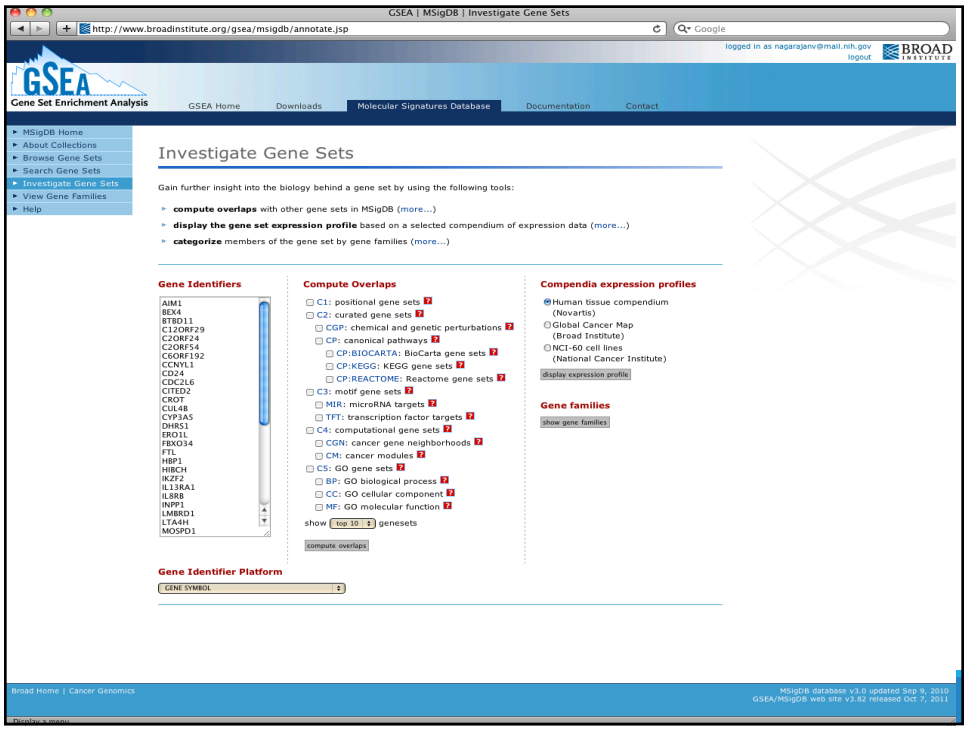
33

The Plan…

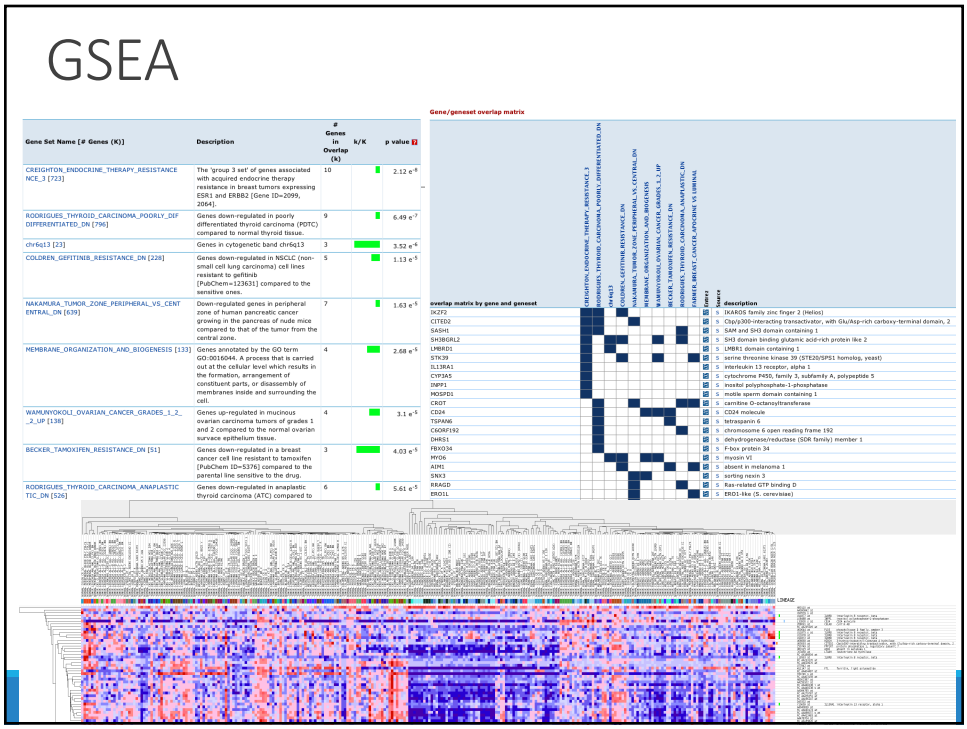
Downstream analysis
- Gene sets
  - Transcriptome studies
  - Cistrome/Epigenome studies
    - Function enrichment
      - GSEA, DAVID, IPA
      - STRING, Cytoscape

Downstream processing
- Genome coordinate sets
  - NGS alignment files (reads/tags)
  - ChIP-Seq peaks
    - BEDTools
    - Galaxy
    - UCSC Table browser

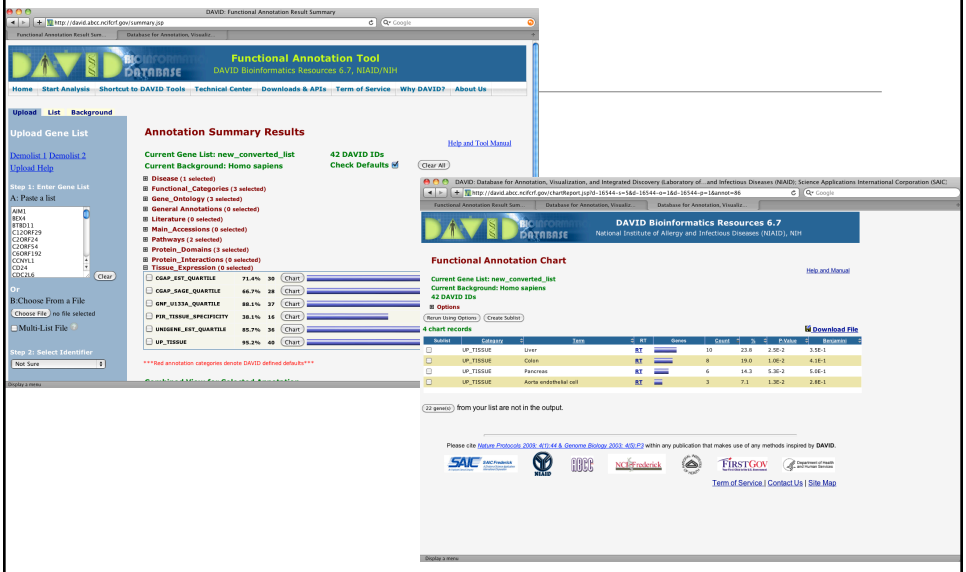
•34


GSEA | MSigDB | Investigate Gene Sets
http://www.broadinstitute.org/gsea/msigdb/annotate.jsp
GSEA
Gene Set Enrichment Analysis
GSEA Home
Downloads
Molecular Signatures Database
Documentation
Contact
MSigDB Home
About Collections
Browse Gene Sets
Search Gene Sets
Investigate Gene Sets
View Gene Families
Help
Investigate Gene Sets
Gain further insight into the biology behind a gene set by using the following tools:
compute overlaps with other gene sets in MSigDB (more...)
display the gene set expression profile based on a selected compendium of expression data (more...)
categorize members of the gene set by gene families (more...)
Gene Identifiers
Compute Overlaps
Compendia expression profiles
Gene families
Gene Identifier Platform
GENE SYMBOL
compute overlaps
display expression profile
show gene families

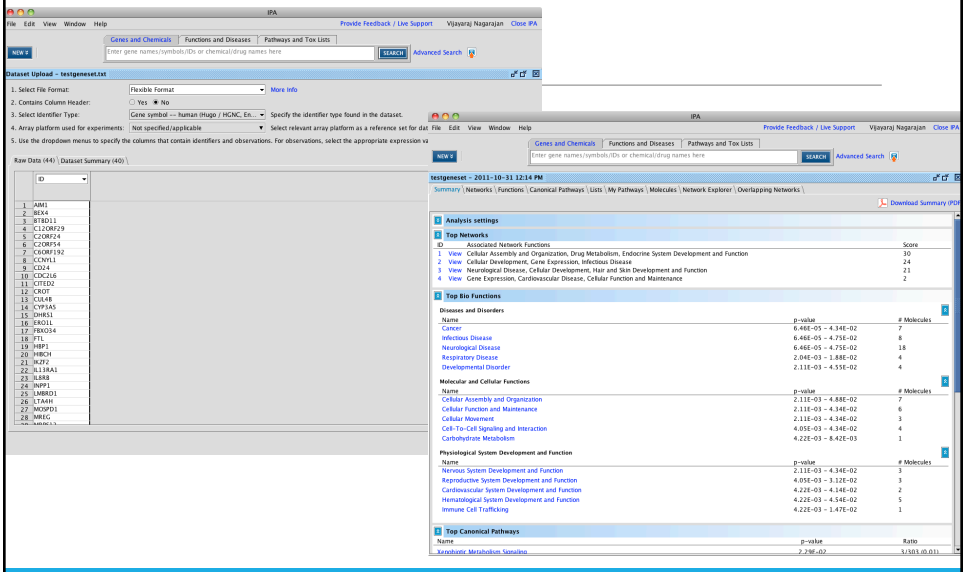
GSEA


Gene/geneset overlap matrix
Gene Set Name [# Genes (K)]
Description
Genes in Overlap (k)
k/K
p value
overlap matrix by gene and geneset
description
LINEAGE

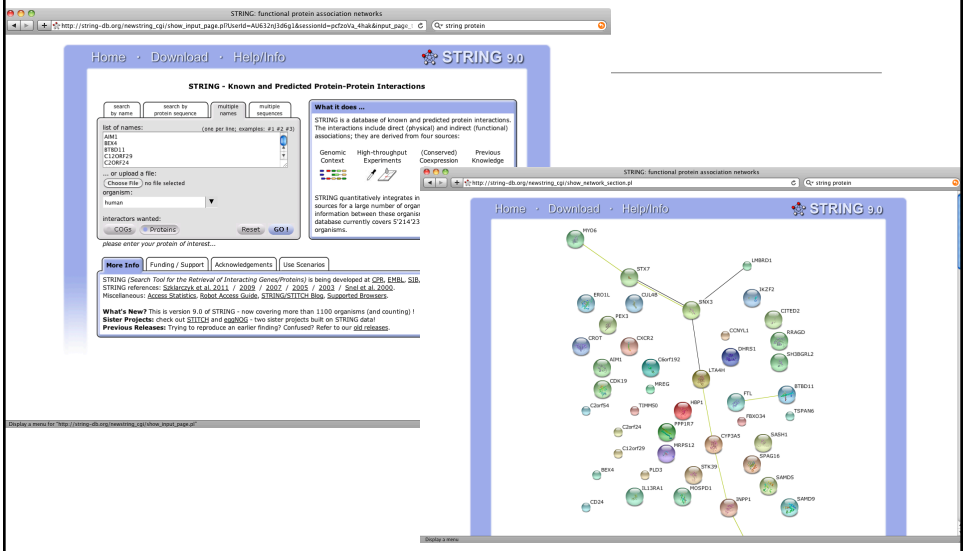
DAVID


Functional Annotation Tool
DAVID Bioinformatics Resources 6.7, NIAID/NIH
Annotation Summary Results
Current Gene List: new_converted_list
Current Background: Homo sapiens
42 DAVID IDs
Check Defaults
Functional Annotation Chart
DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

IPA

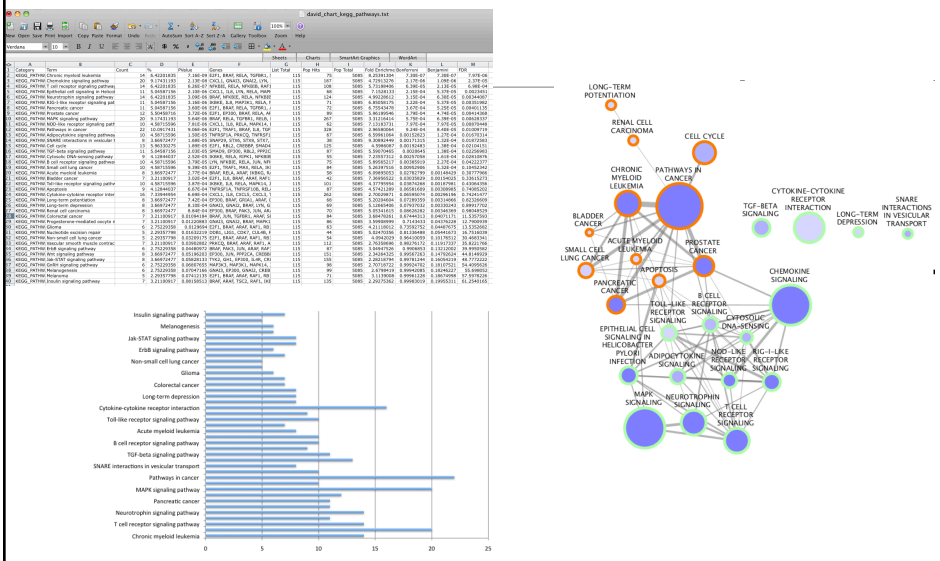

IPA
Dataset Upload - testgeneset.txt
testgeneset - 2011-10-31 12:14 PM
Analysis settings
Top Networks
Top Bio Functions
Top Canonical Pathways

STRING



39

Cytoscape



40

BED-Tools

Command line

Open source

A variety of input formats (BED, BEDPE, SAM/BAM, GFF, VCF)

Things we could do…

- Intersect, Union
- Merge
- Coverage
- Subtract
- Convert
- Closest
- Shuffle
- Group, Sort