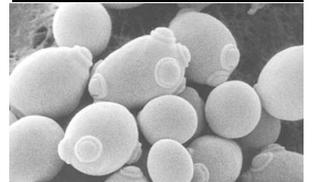
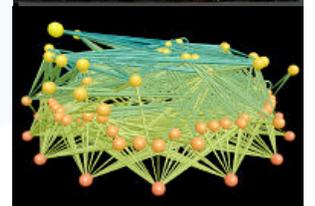
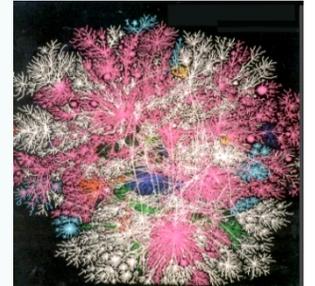
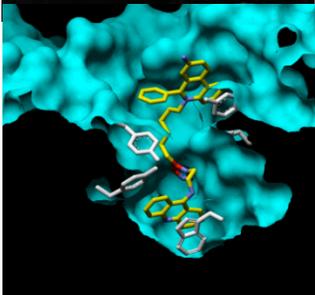
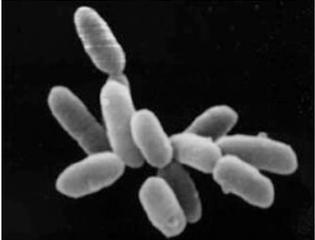
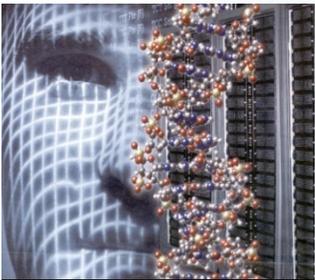


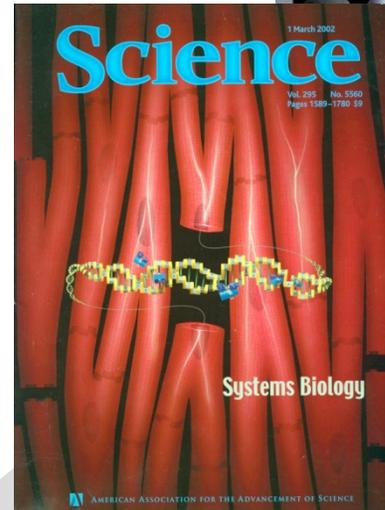
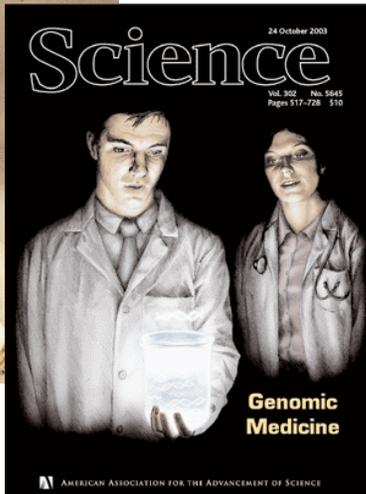
# Complexity and Computing at the Extremes: Next Generation Computational Biology?

David Galas

*Chicago, August 17, 2009*



# A Grand Convergence



Genetics, genomics

*Complex  
Biological Systems*

Systems biology



Modern Computing

# The big questions

- What are the major computational problems in biology facing us?
- Which ones require extreme computing – that is, that we can't solve otherwise?
- Do we have the means to attack these problems?

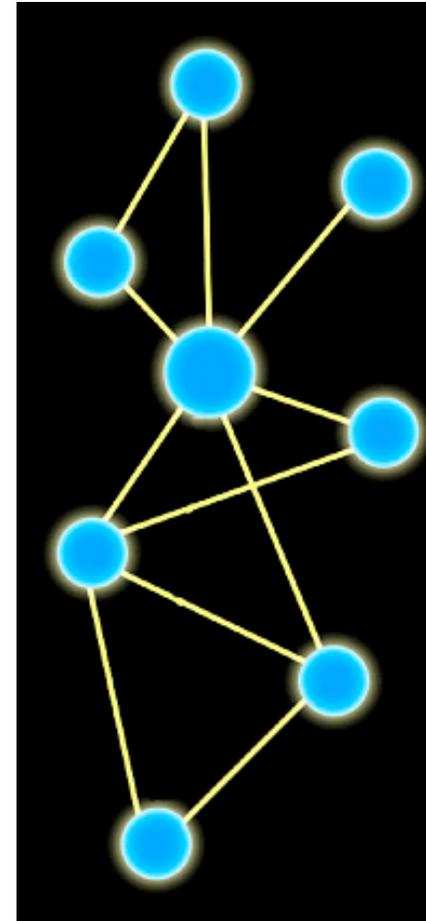
# My main points

- *Global systems approaches to biology are essential*
- ***Interconnectedness of components is extreme***
- *Predictive quantitative models are important, but the first important challenges are not in the models, but rather...*
- *In the synthesis, integration and analysis of data:*
  - *comparative genomics & proteomics, metabolomics*
  - *network comparisons -- evolutionary questions*
  - *genetics & gene interactions*
  - *Putting it all together and iterating models (a very hard problem)*
- *Computing in the design of new experiments*

# Biological Systems

## *Dynamic Networks*

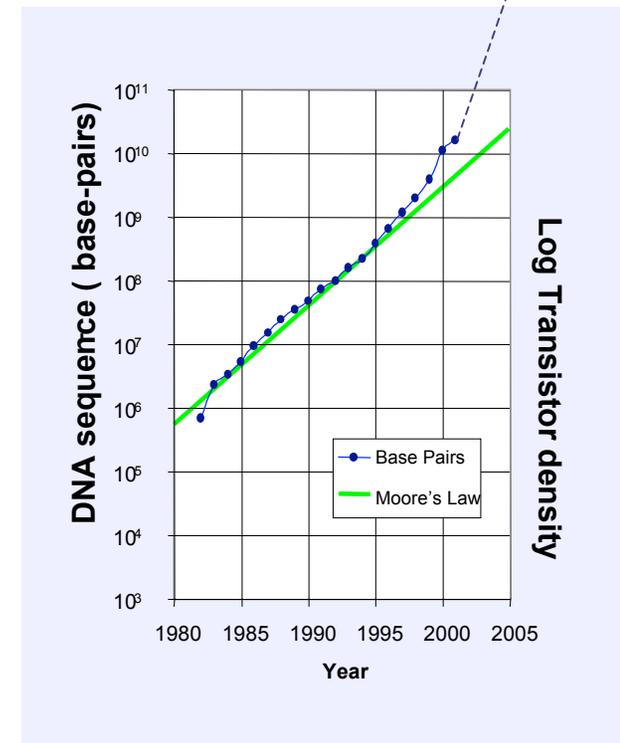
- **Elements** (*genes, proteins, miRNAs, metabolites, complexes, cells, organs.....*) – “nodes”
- **Interactions** between the elements – “edges”—*dynamic (directed, undirected)*
- **Structure of network is also dynamic!**
- Elements and their interactions are affected by the **Context of other systems within--cells and organisms**
- Interactions between/among elements give rise to the system’s **Emergent properties**
- **Unique features**
  - *Global character is essential*
  - *Integration of different data types*
  - *Millions to billions of data measurements per experiment*



# Data gathering technologies

*exponential increases in capacity*

- Sequencing and molecule-counting transcriptomics:
  - Next generation: 454, Solexa, ABI
  - Next-next generation: Helicos & PB ...
    - Single molecule fragment sequencing,
    - High degrees of multiplexing
- Proteomics
  - High-end MS
  - Protein chips (PCA's on nano sized arrays)
  - Nano-wires, SPR etc..
- Prediction: human genome for ~\$5K next year



# Data Types

- mRNA levels (in time and space)
  - lambdas and ratios
- Protein levels (in time and space)
  - Protein probabilities (protein prophet)
  - ASAP ratios
  - Peptide counts
- Protein associations
  - functional relationships/associations (phylogenetic pattern)
  - genome organization (operons chromosomal proximity)
- Protein-protein interactions (MS/MS)
  - Protein probabilities (protein prophet)
  - ASAP ratios
  - Peptide counts
- Cellular Structure
  - Images
  - Cellular fractionation
- Protein-DNA interactions (ChIP-chip, seq)
  - lambdas and ratios
  - genome localization
- Chromatin structure in nucleus
- Gene function annotations
  - Pfam domains
  - COGs
  - Protein Structure
  - TIGRFam
- Metabolic pathway dynamics
- Phenotypes
  - Growth/survival, apoptosis
  - Cell morphology
  - Structure based: imaging (organelle localization etc.)
- Regulatory networks
  - Cis-regulatory motifs
  - Co-regulated gene profiles
  - Protein interactions/associations
  - Regulatory relationships
  - Cell-cell signalling networks

# Major Computational Challenges

- Define the key problems well
- Approaches to dimension reduction in analysis of biological data
- Understand the uses of models vs data analysis
- Key issues:
  - Amount and types of data
  - Complexity & interrelatedness
    - Cell type specificity
    - Cell-cell interaction complexity
    - Dynamical complexity
  - Genetics !



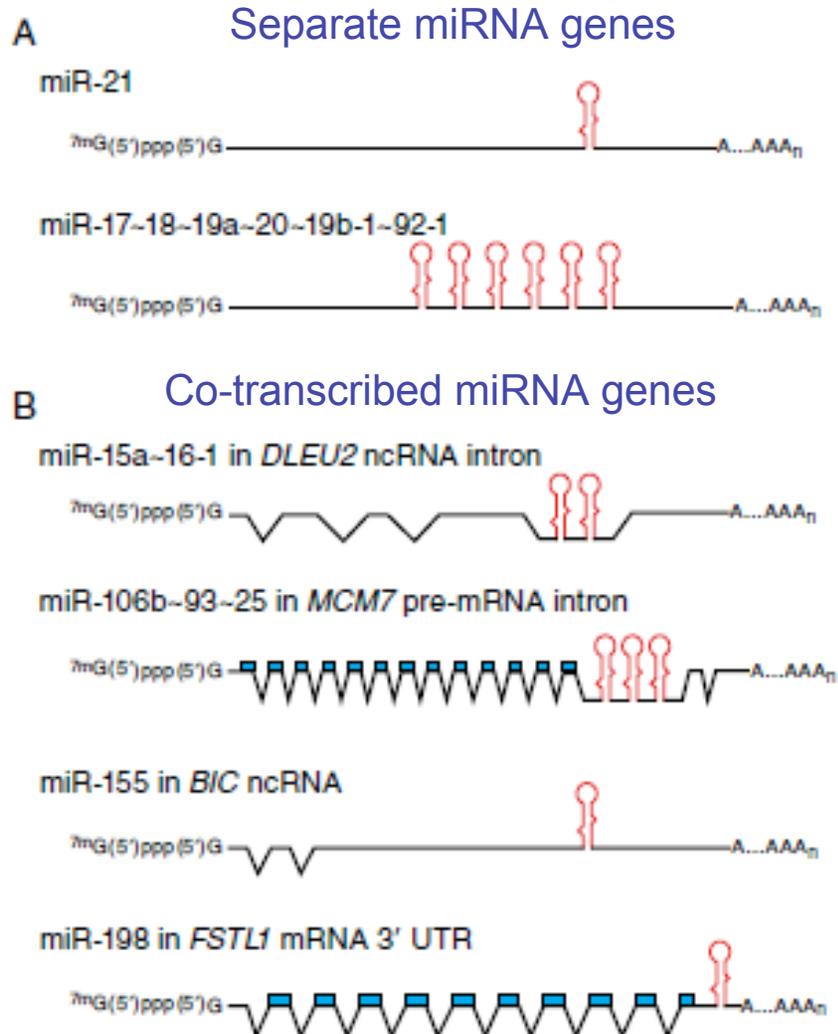
**Need tools to match the problems!**

# Three insidious problems

- Dynamics of networks includes structure – feedback on network structure
- Genetic interactions are dominant. Genetic variation modifies the systems in what ways?
- There are large numbers of weak effects.

# A Poorly understood System

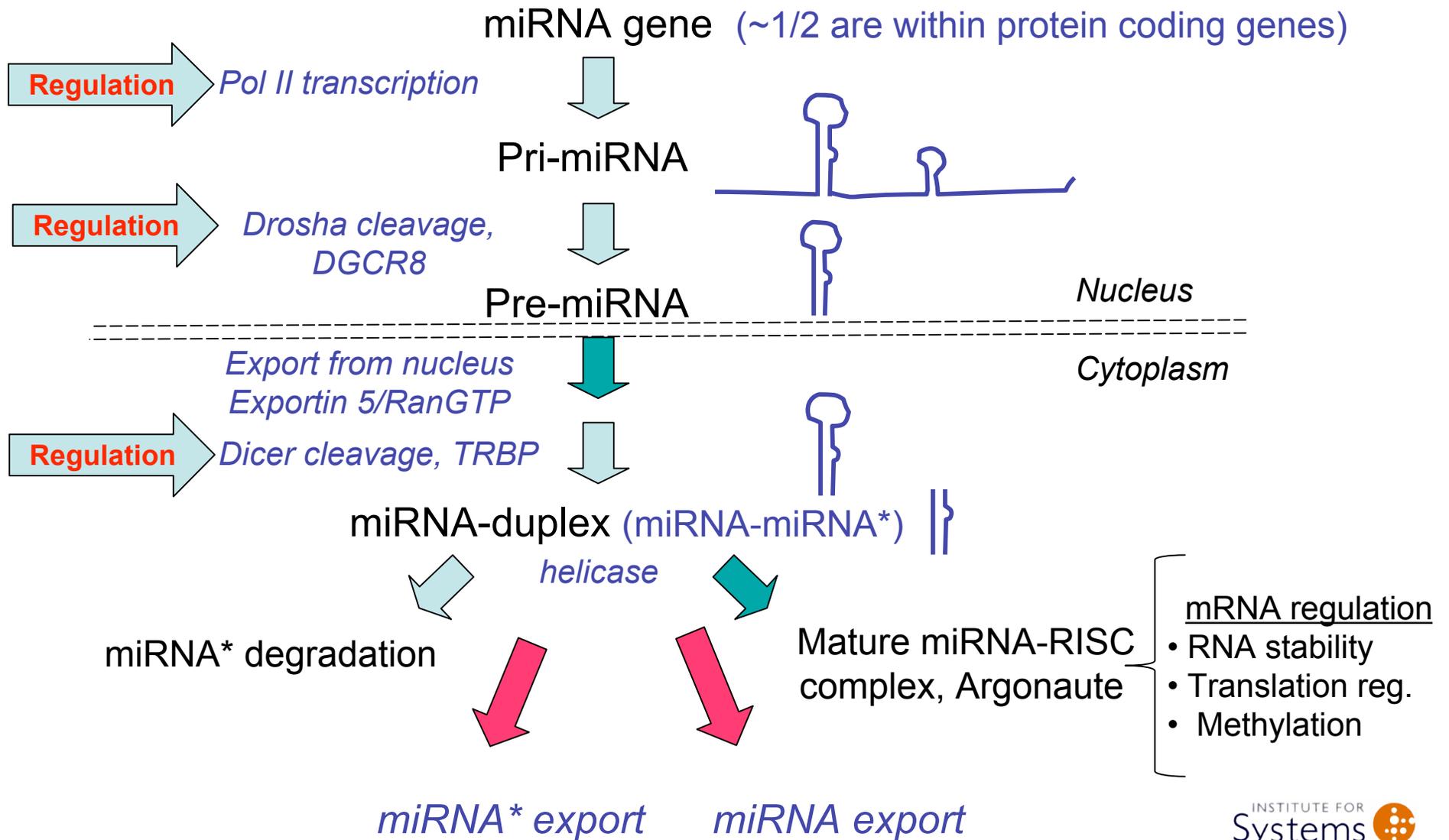
## ~1000 microRNAs in Mammals



### Some Problems

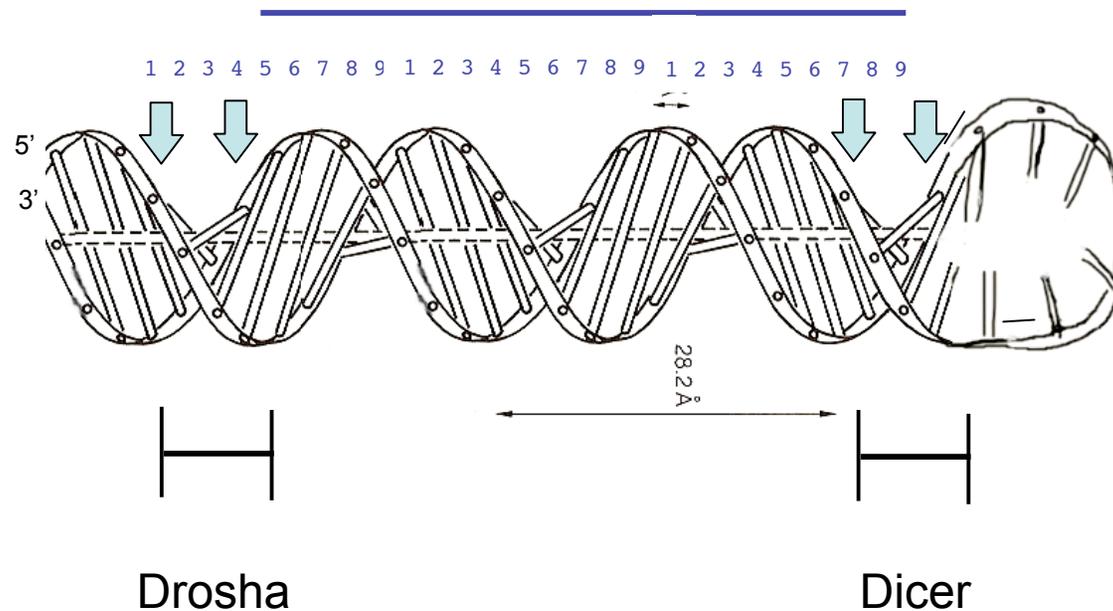
- Regulatory targeting not well understood
- Multiple miRNAs reg. same gene
- Modifications of miRNA can change functions

# microRNA pathways



# Processing of Pre-miRNA and Consequences

*Structure is drawn for perfect base-pairing*



## Modification of regulatory Functions

- Several miRNA from same gene (processing)
- Modification and editing of miRNAs
- Strand switching from miRNA to miRNA\*

# Three insidious problems

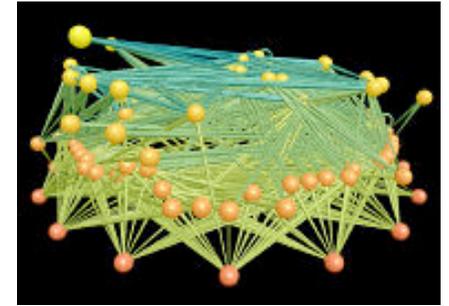
- Dynamics of networks includes structure – feedback on network structure. *Example: miRNAs*

- Genetic interactions are dominant. Genetic variation modifies the systems in what ways?

- There are large numbers of weak effects.

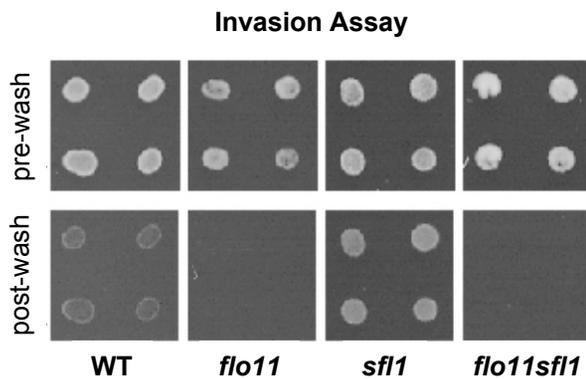
# Genetic Interactions

*little understood, poorly integrated*

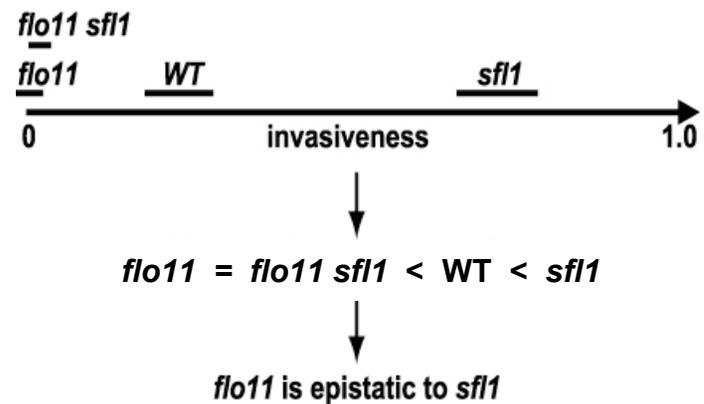


- Pairwise perturbation of genes
- Two genes *combine* to affect phenotype  
(Hereford & Hartwell, 1974)
- Types of interactions can reveal network structure
  - Non-additive effects
  - Synthetic lethals
  - Epistasis
  - Multicopy suppression
- Loss-of-function, gain-of-function, dominant-negative, etc.
- Interaction depends on **phenotype** measured!

# Genetic Interaction Study in Yeast

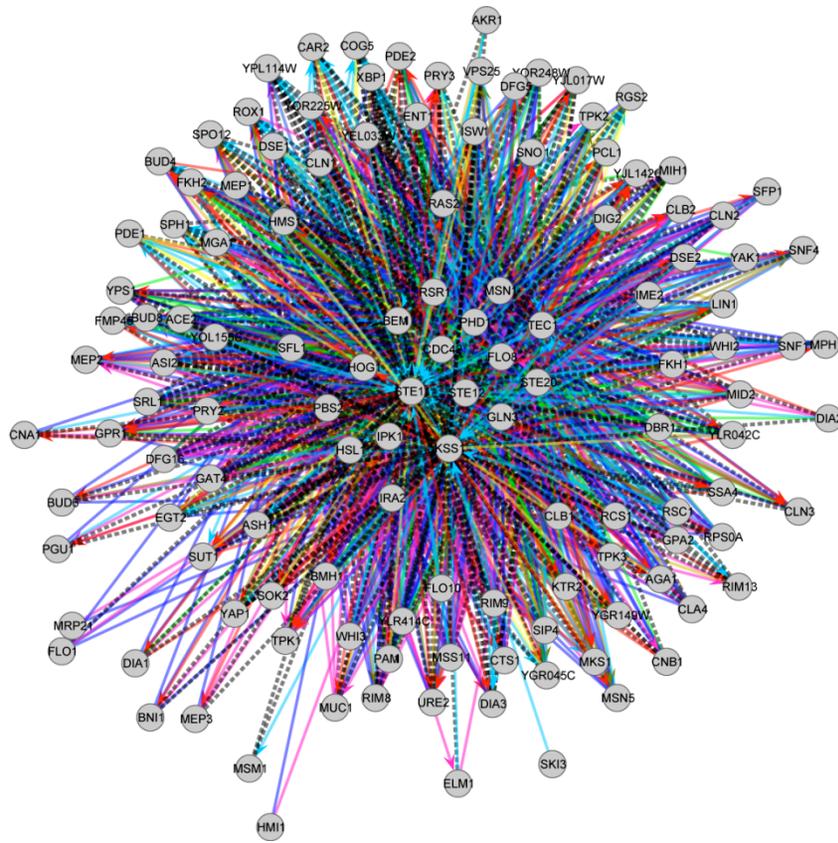


Repeated for ~2000 genetic interactions



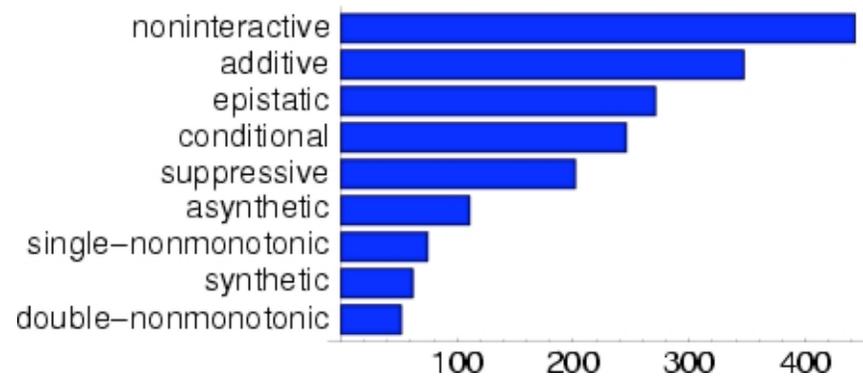
Drees, Thorsson, Carter, et al., *Genome Biol.* 2005

# Genetic Interaction Network Example



Yeast invasion network  
~2000 tested interactions  
among 130 genes

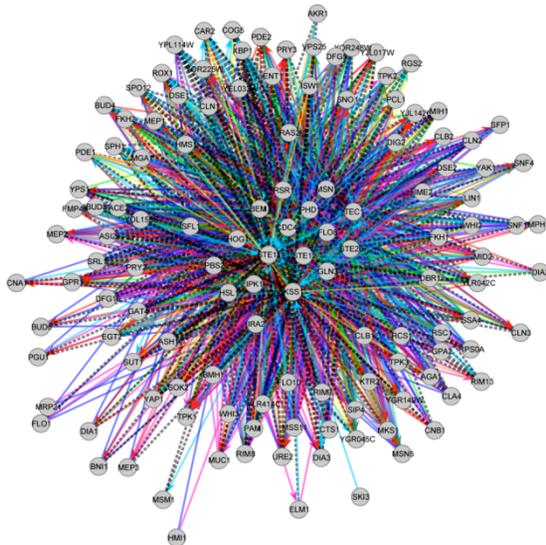
Distribution of Types of Interactions



Galitski & Carter, ISB

# Data Sets

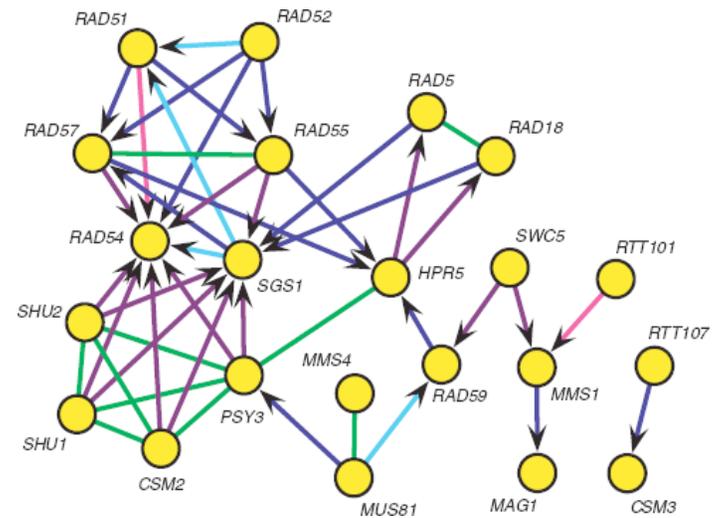
## Invasion data



Drees, et al. *Genome Biology* 2005

130 genes, ~2000 pairs tested

## MMS fitness data

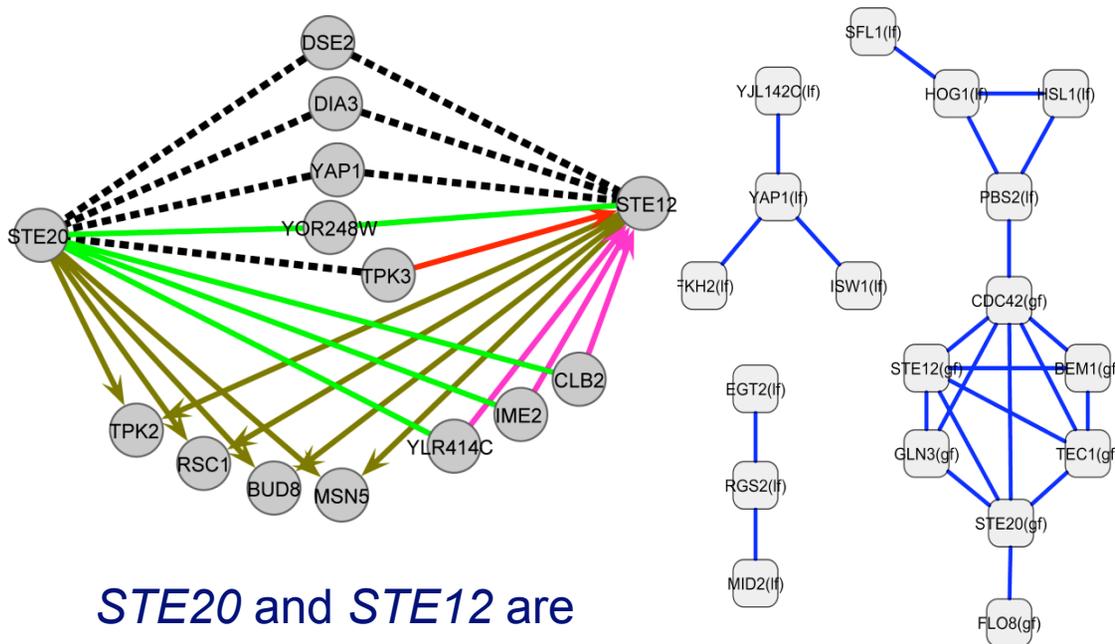


St. Onge, et al. *Nature Genetics* 2007

26 genes, 325 pairs tested

# Mutual Information & Set Complexity

Gene pairs that exhibit systematic interaction patterns



*STE20* and *STE12* are mutually informative ( $p = 10^{-16}$ )

Modular maps

$$\Psi = \sum K_i m_{ij} (1 - m_{ij})$$

$K_i$  is the information of element  $i$ ,  
 $m_{ij}$  is the mutual information between  $i$  and  $j$ ,

$$0 \leq m_{ij} \leq 1 \text{ and } 0 \leq \Psi \leq 1$$

# Understanding Gene interactions

*independent test against “biological information”*

MMS fitness:

116K possible networks

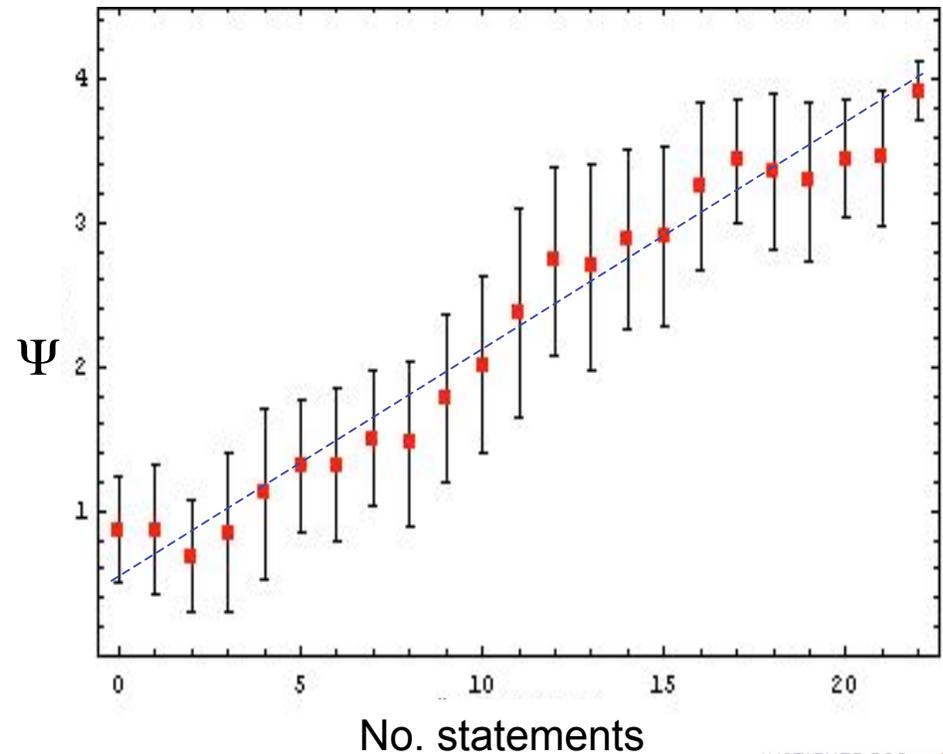
(all classifications of 10 interactions on 26 genes)

## BIOLOGICAL INFORMATION

• A **biological statement** is a result of the interaction classes and annotation information on the genes (from database)

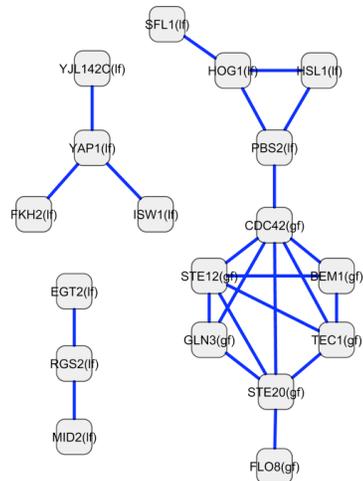
- These statements are real biological information derived from the classifications
- The number of biological statements is

proportional to  $\Psi$ !

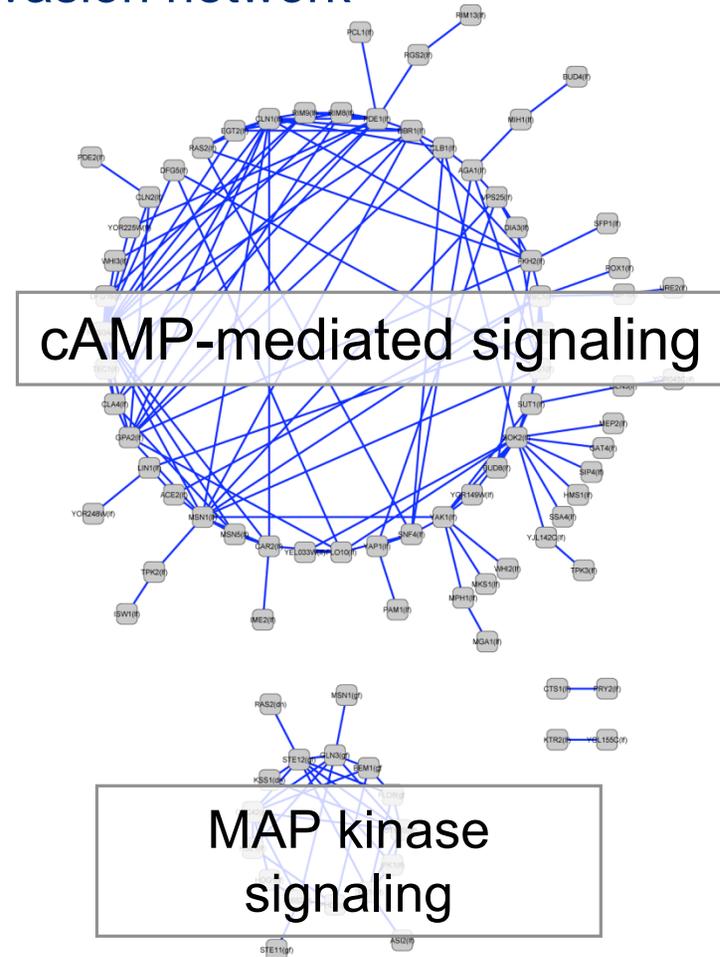


# “Maximally Complex” Networks

Mutual information in the yeast invasion network



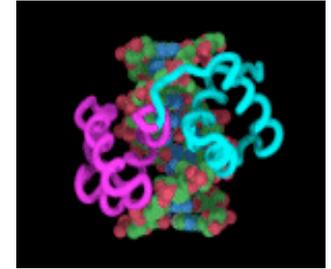
original network ( $\Psi = 0.57$ )



maximally complex network ( $\Psi = 0.79$ )

# Integration:

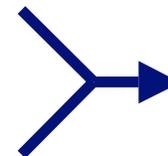
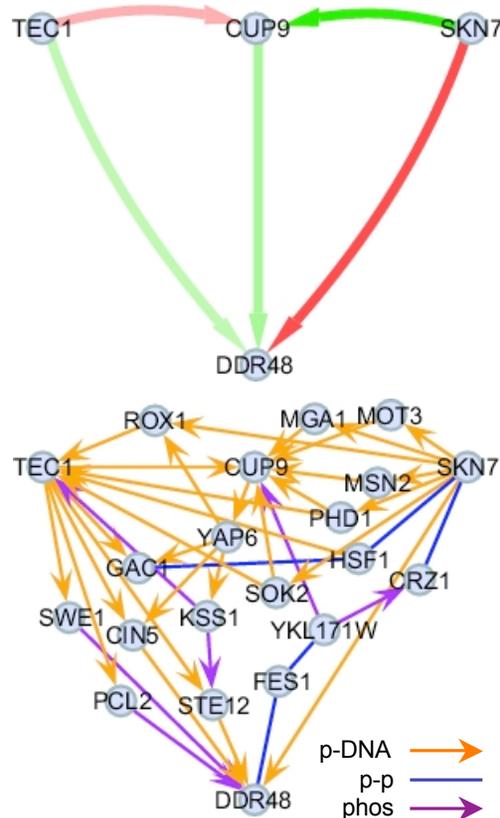
*gene interaction to molecular interactions*



Generate molecular hypotheses for information flow

example: expression of *DDR48*

*functional*  
influences  
Network  
(from time series  
global expression,  
Genetic interactions)

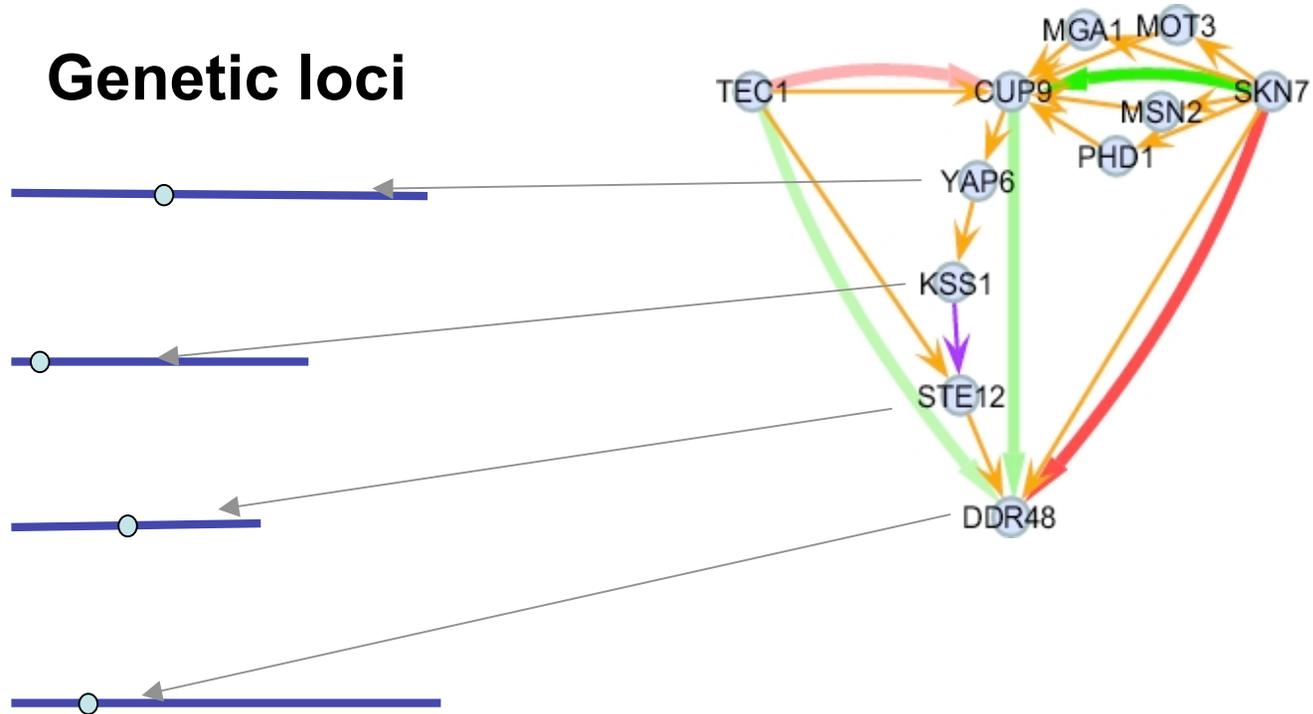


integrated  
network

*physical*  
molecular  
network

Carter, et al., *Genome Res.* 2006  
Carter, et al., *Mol. Syst. Biol.* 2007

# DDR48 expression determined by >10 genes



- 2 variants for each gene in population
- There are 1000 combinations
- One quantitative phenotype

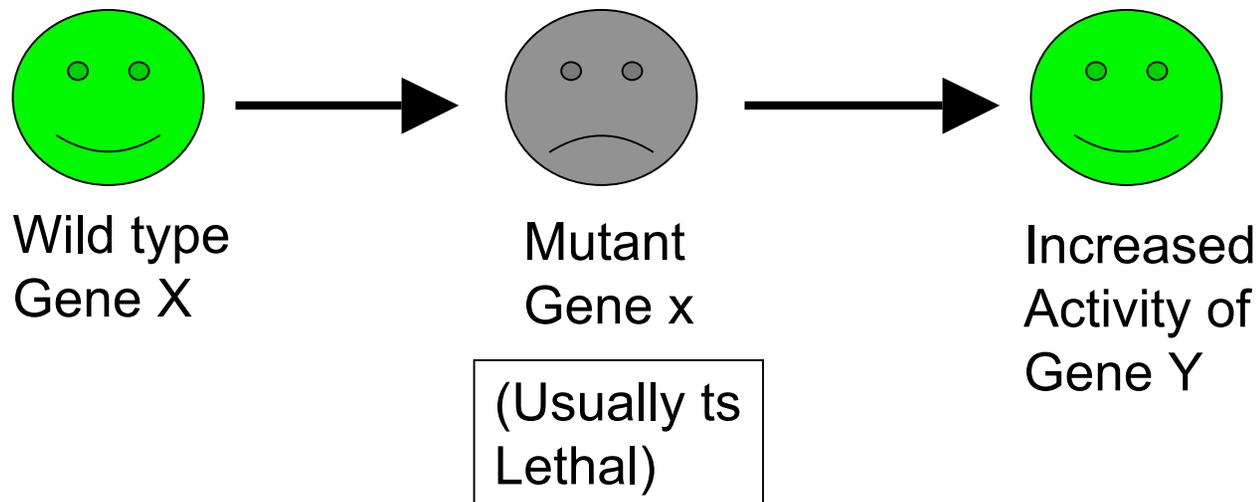
# Three insidious problems

- Dynamics of networks includes structure – feedback on network structure. *Example: miRNAs*
- Genetic interactions are dominant. Genetic variation modifies the systems in what ways? *Examples: Gene interaction networks*
- There are large numbers of weak effects.

# Genome-wide Dosage Suppression Network

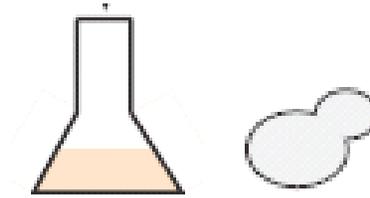
- An unexpectedly rich view of the network of genes regulating cell cycle
- Studies the plasticity of network function
- Reveals normally weak network interactions

## Dosage Suppression



Conditional mutants

Grow at permissive conditions



Suppressor  
identification

Transform with MORF Plasmids

( Whole genome ORF collection under Gal control)



Plate cells and allow them to recover in selective media/permissive temp.



Scrape and pool all transformants

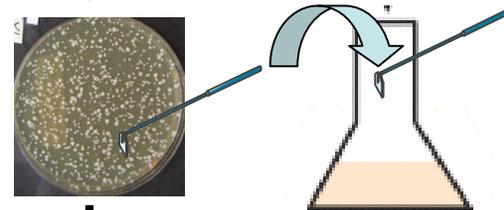


Plate cells in inducing media and shift to restrictive conditions



**Yeast cells grow if mutation is suppressed!**

# 80 *ts* strains

Cell Cycle	
<i>cak1-23</i>	<i>cdc48-9 (Y-F)</i>
<i>cdc123-4</i>	<i>cdc5-1</i>
<i>cdc13-1</i>	<i>cdc6-1</i>
<i>cdc15-2</i>	<i>cdc7-1</i>
<i>cdc16-1</i>	<i>cdc9-1</i>
<i>cdc20-1</i>	<i>cks1-35</i>
<i>cdc24-H</i>	<i>ctf13-30</i>
<i>cdc25-1</i>	<i>kin28-ts</i>
<i>cdc2-7</i>	<i>med11-ts</i>
<i>cdc28-td</i>	<i>med4-6</i>
<i>cdc33-E72G</i>	<i>mms21-1</i>
<i>cdc35-1</i>	<i>pob3-7</i>
<i>cdc36-16</i>	<i>pol5-2</i>
<i>cdc37</i>	<i>pti1-ts7</i>
<i>cdc39-1</i>	<i>rad3-ts14</i>
<i>cdc40-ts</i>	<i>rsp5-1</i>
<i>cdc4-1</i>	<i>smc2-8</i>
<i>cdc45-27</i>	<i>spt6-14</i>
<i>cdc46-1</i>	<i>taf12-9</i>
<i>cdc47</i>	<i>tfb3-ts</i>

RNA-related	
<i>abd1-5</i>	<i>mcm10-1</i>
<i>abf1-102</i>	<i>mcm2-1</i>
<i>afg2-18</i>	<i>mot1-1033</i>
<i>arp4-G161D</i>	<i>nab3-11</i>
<i>arp7-E411K</i>	<i>nog2-1</i>
<i>cft2-1</i>	<i>nop2-3</i>
<i>cus1-3</i>	<i>orc2-1</i>
<i>dbf2-1</i>	<i>orc3-70</i>
<i>dbp5-1</i>	<i>pcf11-ts10</i>
<i>dcp2-7</i>	<i>prt1-1</i>
<i>ded1-199</i>	<i>rap1-1</i>
<i>dim1-2</i>	<i>rat1-1</i>
<i>esal-L254P</i>	<i>rnt1-ts</i>
<i>ess1-H164R</i>	<i>rsc3-1</i>
<i>fcpl-1</i>	<i>spt15-P65S</i>
<i>gcd10-506</i>	<i>sup35-td</i>
<i>god1-502</i>	<i>swd2-1</i>
<i>hsf1-848</i>	<i>tfcl-E447K</i>
<i>hts1-1</i>	<i>tor2-21</i>
<i>hyp2-1</i>	<i>yef3-F650S</i>

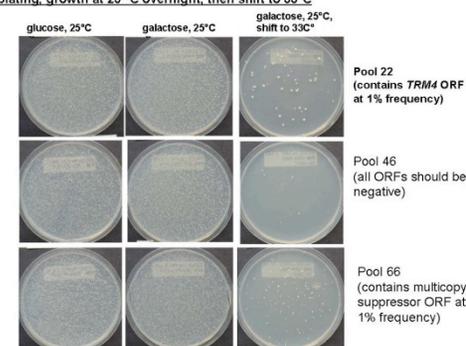
# Multi-copy suppression (MCS)

~70% of genes tested could not be suppressed

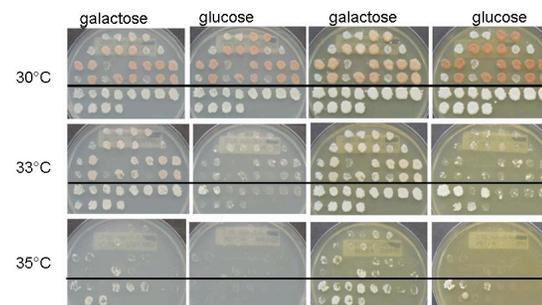
Examples of the 30% that did

<u>Mutant</u>	<u>Number of MCS genes</u>
<i>trm4Δ trm8Δ</i>	4
<i>prp3-1</i>	7
<i>rpb1-1</i>	1
<i>prp8-1</i>	6
<i>nup116Δ</i>	2
<i>prp16-1</i>	1
<i>prp11-1</i>	5
<i>prp21-1</i>	12

Direct selection of suppressors of *trm4 trm8* mutant by transformation, plating, growth at 25 °C overnight, then shift to 33°C



Re-test for galactose dependent suppression for *prp3-1* candidates and *rpb1-1* candidates



Dosage suppressor network (FIBR+SGD) involving *cdc*  
& related mutant nodes screened by FIBR

# Network of shared suppressors of *cell cycle* related mutants

# A SUMO Ligase substitutes for Ubitquitin Ligase

# Some Computational and Mathematical Challenges

1. How to do all-vs-all comparisons of many 1000s of human genomes, proteomes etc...
  - How to infer phenotype from genotype
  - Understand the noise, biological and measurement induced
2. How to integrate multiple high-throughout data types, including images and structure
3. How to visualize & explore large-scale, multi-dimensional biological data
4. How to infer protein, miRNA and gene regulatory networks from genetic, expression, etc.. Highly heterogeneous data
  - How to find the common core and the significant genetic variants
  - How to attribute differential effects to the alleles
5. How to build useful, predictive models across multiple scales of time & space, and connect logically to large data sets

# Dynamics of PrP network in Mouse brain

## 6 to 22 wks



# Three insidious problems

- Dynamics of networks includes structure – feedback on network structure
- Genetic interactions are dominant. Genetic variation modifies the systems in what ways?
- There are large numbers of weak effects.

# Our challenge

*to focus on the right problems*



# Acknowledgments

- Greg Carter, ISB
- Ilya Shmulevitch, ISB
- Nathan Price, ISB
- Kai Wang, ISB
- Ji-Hoon Cho, ISB
- Nikita Sakhanenko
- Matti Nykter, ISB
- Tim Galitski, ISB
- Lee Hood, ISB
- Nat Goodman, ISB
- Eric Phizicky, U of Rochester
- Animesh Ray, KGI

## Support

- NSF *FIBR Program*; NIH, Battelle Memorial Institute,  
Grand Duchy of Luxembourg