02-710
Computational Genomics

Reconstructing signaling and dynamic regulatory networks

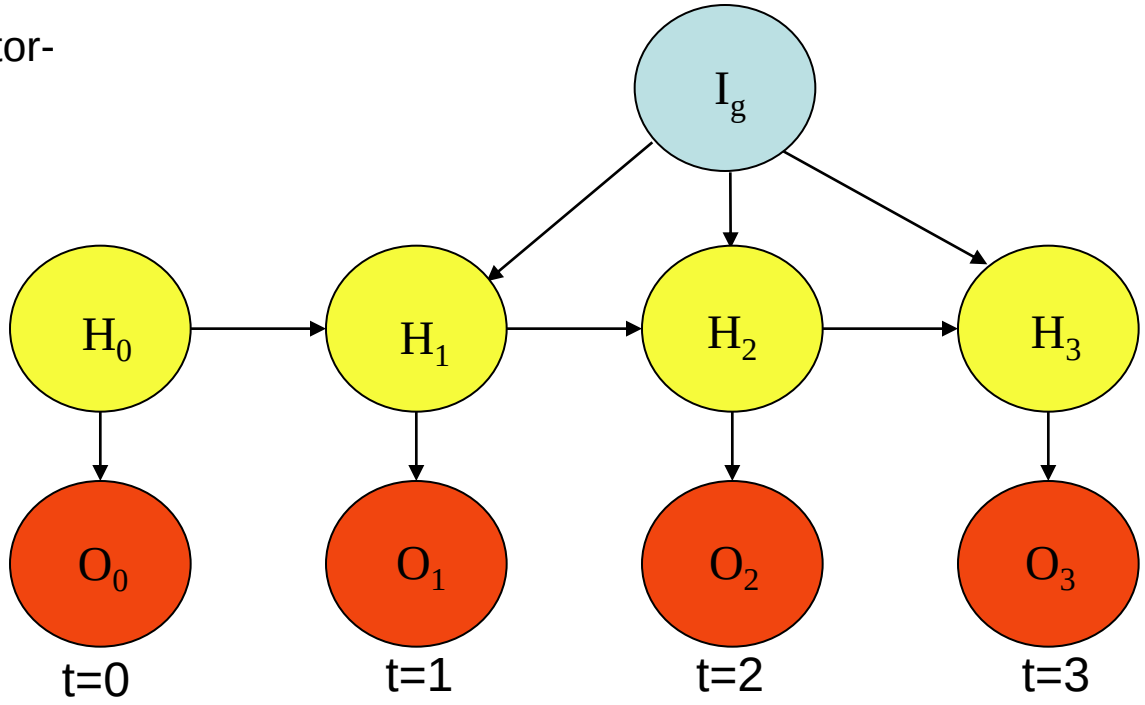
Input – Output Hidden Markov Model

Bengio and Frasconi, *NIPS* 1995

Input (Static transcription factor-gene interactions)

Hidden States Variables
(We constrain transitions between states to form a tree structure)

Output State Variables
(Gaussian distribution for expression values)



Log Likelihood

$$r(G|M) = \sum_{g \in G} \log \sum_{q \in Q} \prod_{t=1}^{n-1} f_{q(t)}(o_g(t)) \prod_{t=1}^{n-1} P(H_t = q(t) | H_{t-1} = q(t-1), I_g)$$

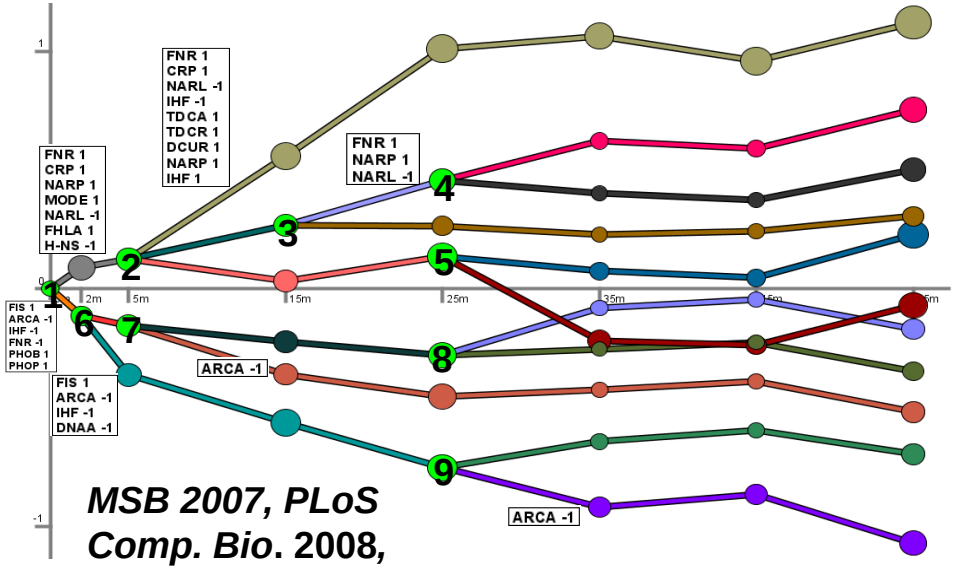
Sum over all genes

Sum over all paths *Q*

Product over all Gaussian emission density values on path

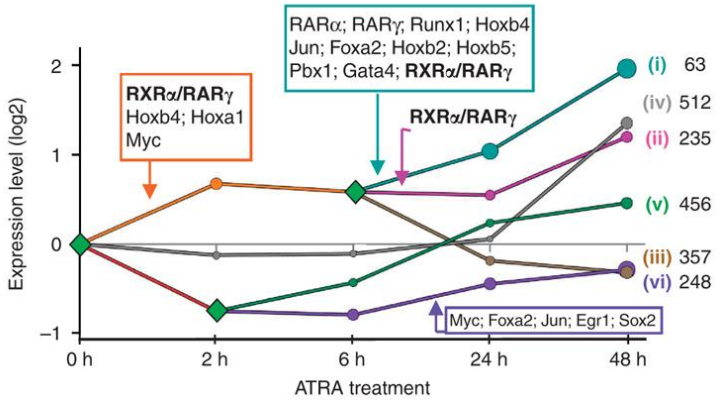
Product over all transition probabilities on path

Stress and hormone response



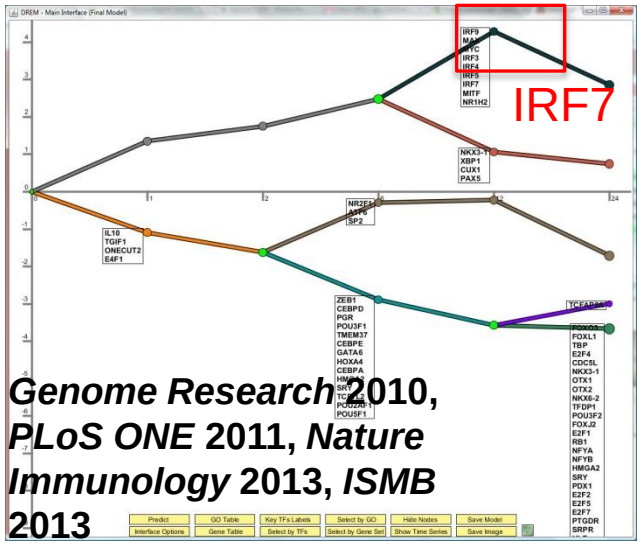
MSB 2007, PLoS Comp. Bio. 2008, *eLife* 2013

Stem cells differentiation



MSB 2011

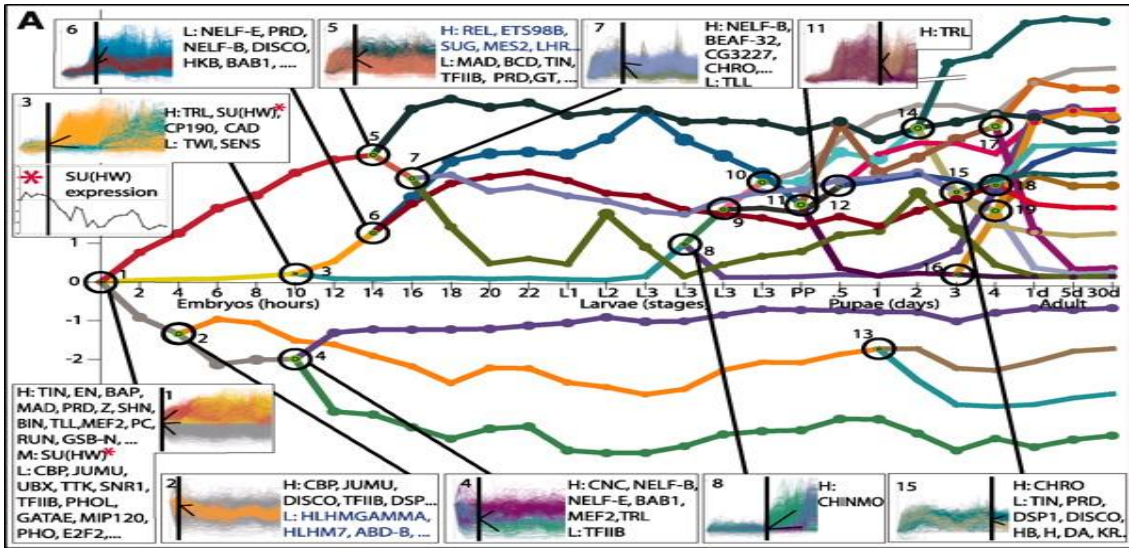
Immune response



Genome Research 2010, *PLoS ONE* 2011, *Nature Immunology* 2013, *ISMB* 2013

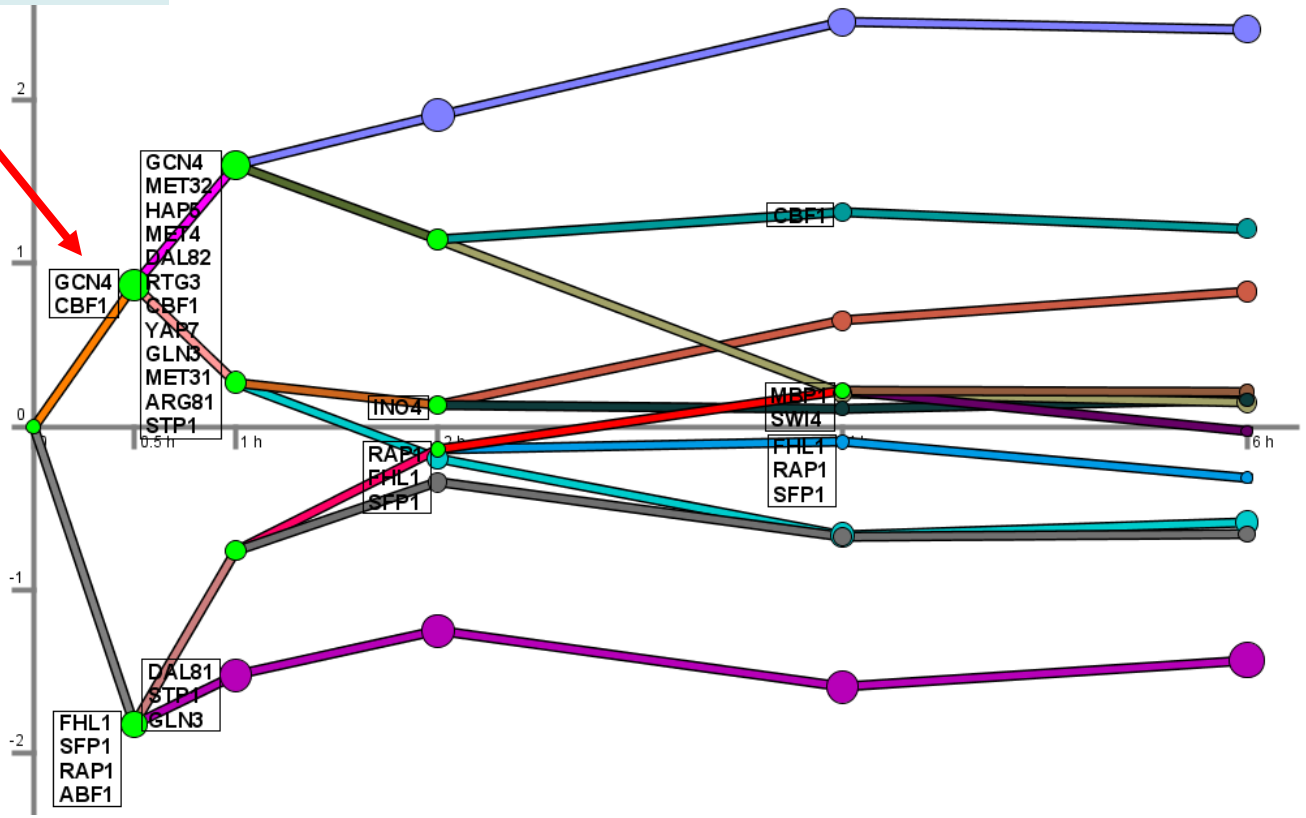
Fly development

Science 2010

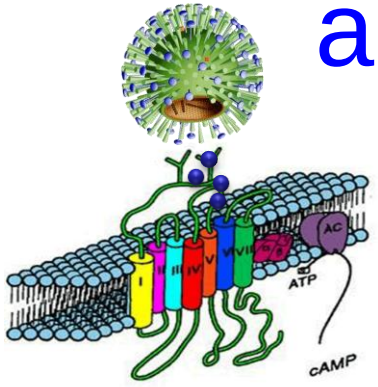


DREM is useful, but several questions remain …

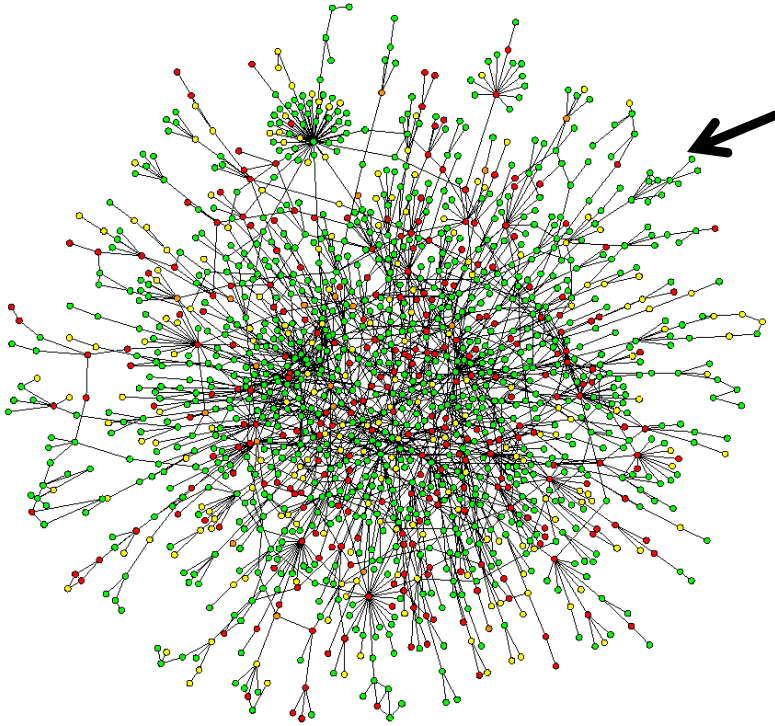
Who controls the master regulators?

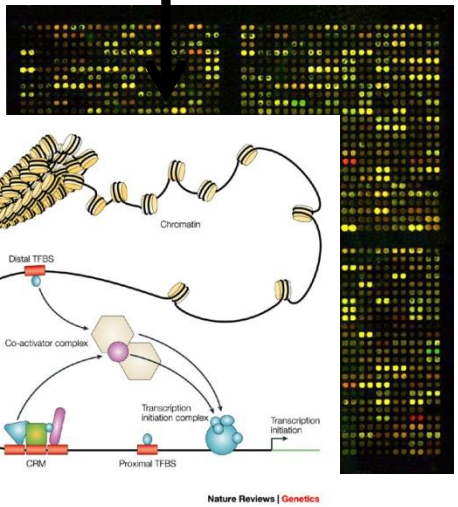


Linking the signaling and regulatory networks



What's going on here?!?



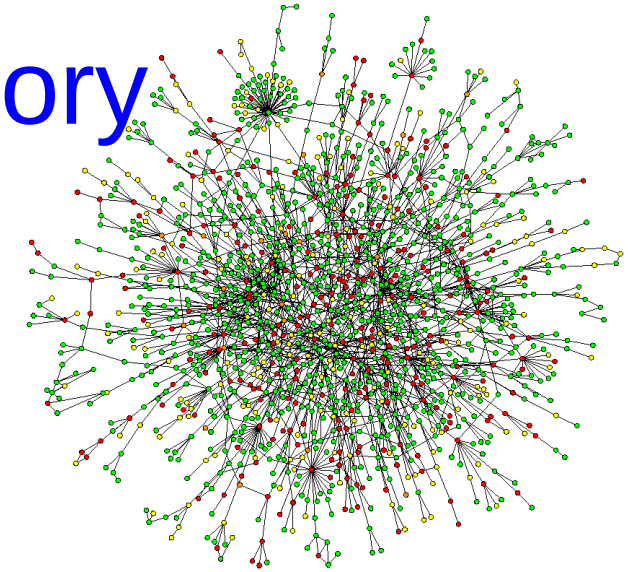


SDREM: Extending DREM to model signaling networks

Inputs:

- Condition specific inputs:
  - Time series expression data following treatment
  - (A few) receptors interacting with invader or activated by condition of interest
- General interaction data (not necessarily from the same condition):
  - Protein-DNA interactions
  - Motif information
  - Protein interaction networks

Iterative method for reconstructing dynamic signaling and regulatory networks



Identify TFs actively regulating gene expression

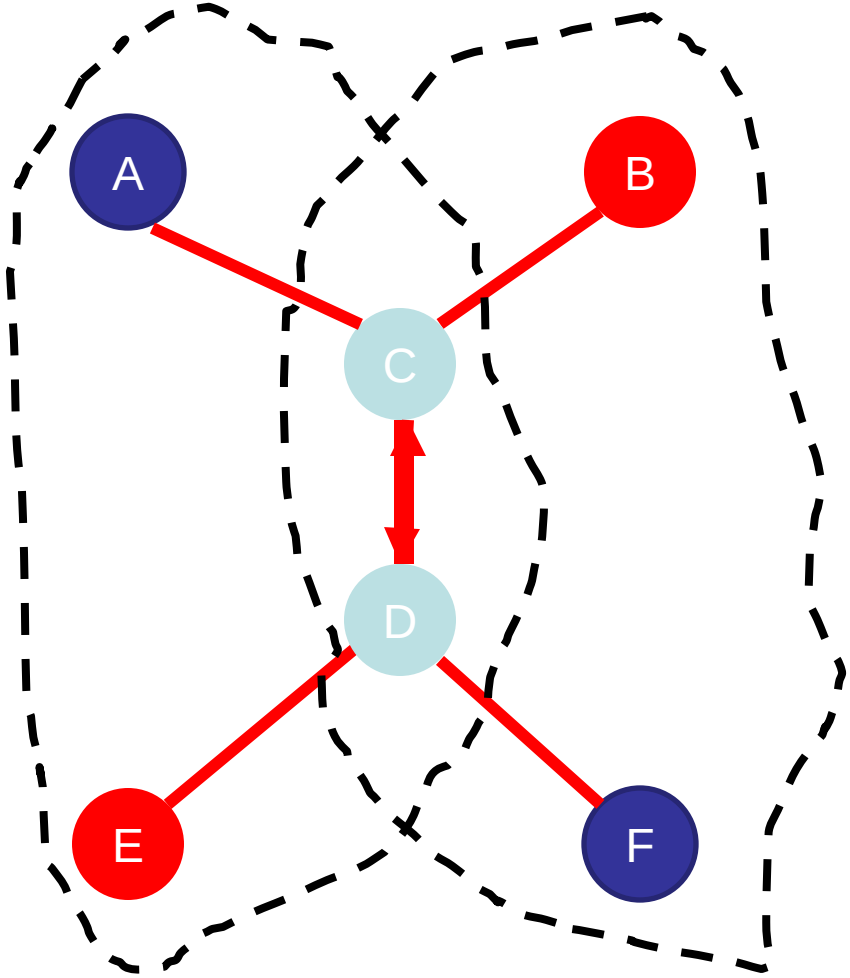
Determine which active TFs are well-connected in the PPI network?

How do we orient the network?

- We are given the receptors / sensing proteins and the TFs
- We need to determine an appropriate path from receptors to TFs
- The orientation should
  - Use short source-target pathways
  - Prefer high-confidence interactions
  - Encourage parallel pathways

Example

Source target pairs:
{A,E}, {F,B}



Maximum Edge Orientation (MEO)

- Mixed graph $G = (V, E)$
- Known set of sources S and targets T (from DREM)
- Maximum path length k
- Consider all simple paths P from $s \in S$ to $t \in T$
- Path weights given

Can be converted into a satisfiability problem and approximated using SAT solvers

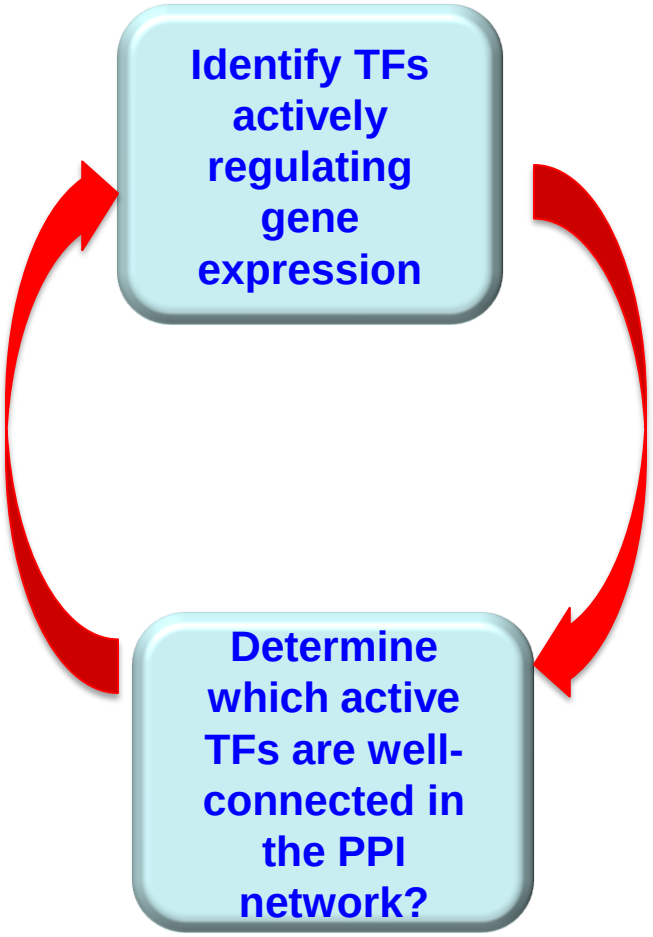
- Objective is to maxi

$$\sum_{p \in P} I_S(p) * w(p)$$

weight
(from DREM)

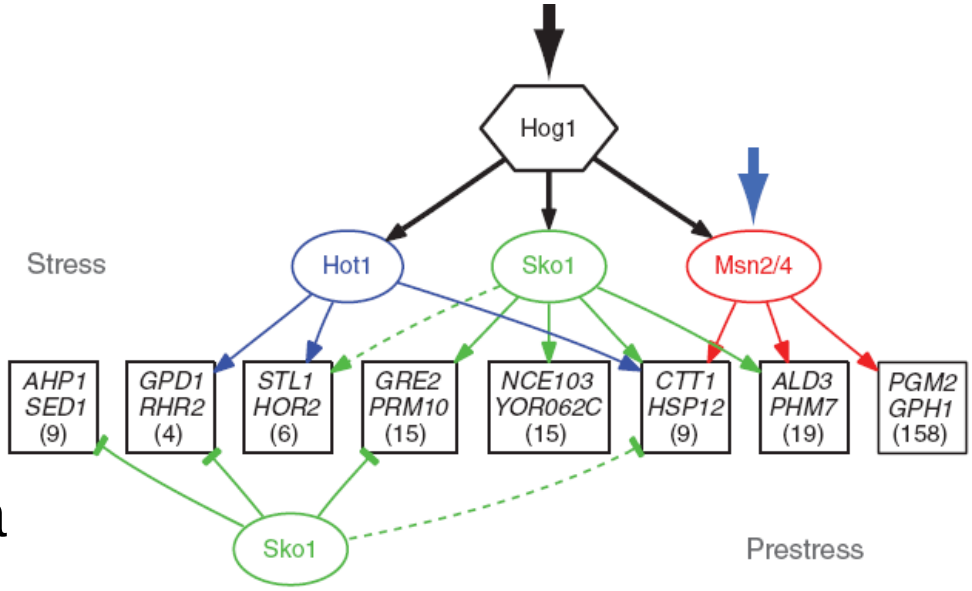
where $I_S(p)$ indicates if p is satisfied

Yeast response to osmotic stress



HOG pathways analysis

- The high osmolarity glycerol (HOG) pathway is activated by increased environmental osmolarity and results in a rise of the cellular glycerol concentration.



Capaldi et al *Nature Genetics* 2008

- **Condition specific input:**
- Time series expression data after treatment with sorbitol
- 5 known proteins that sense this condition

input sensory proteins

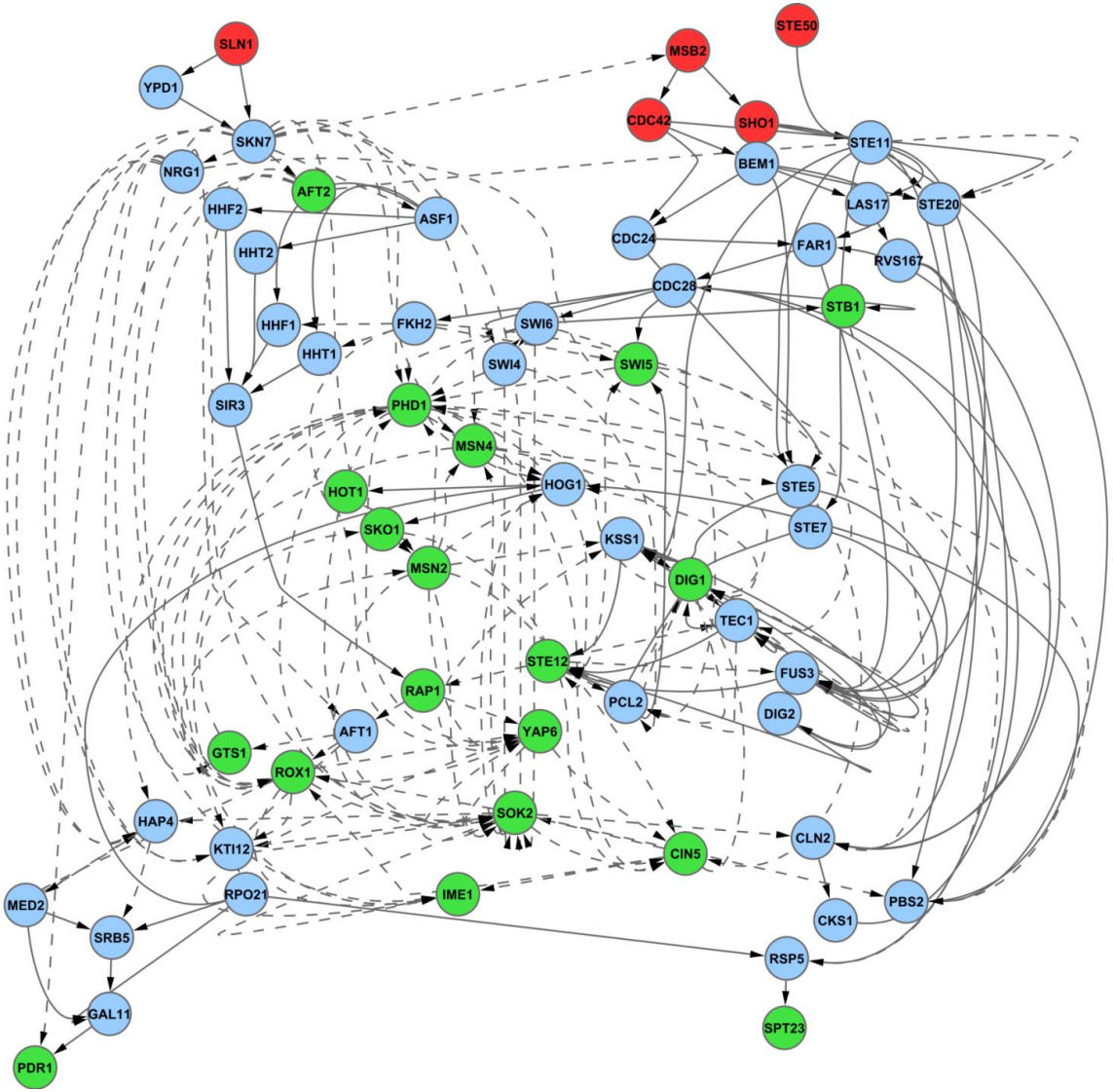
signaling network proteins

transcription factors (from DREM)

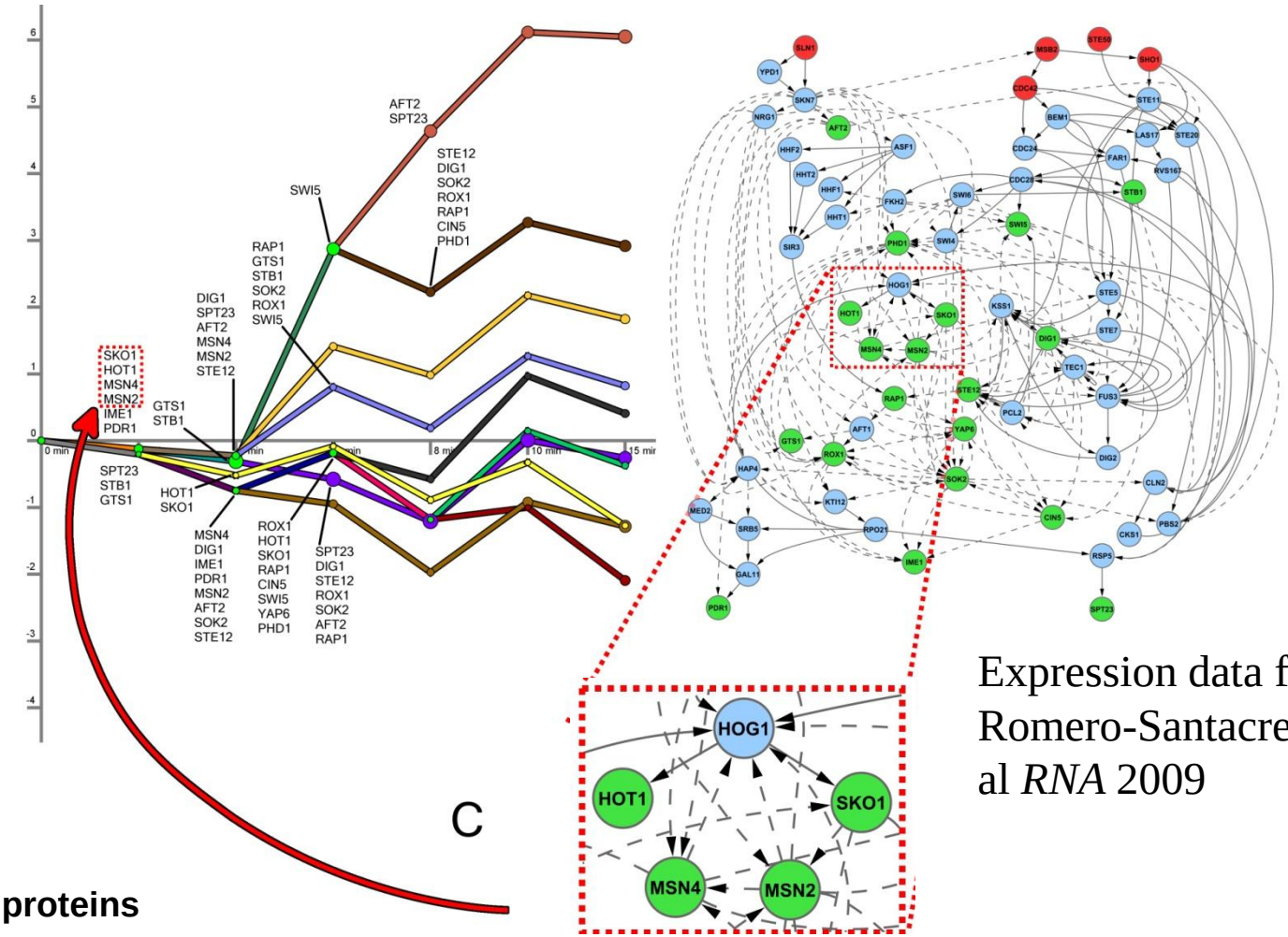
Yeast response to osmotic stress

Identify TFs actively regulating gene expression

Determine which active TFs are well-connected in the PPI network?



Reconstructed HOG pathway: Short time series

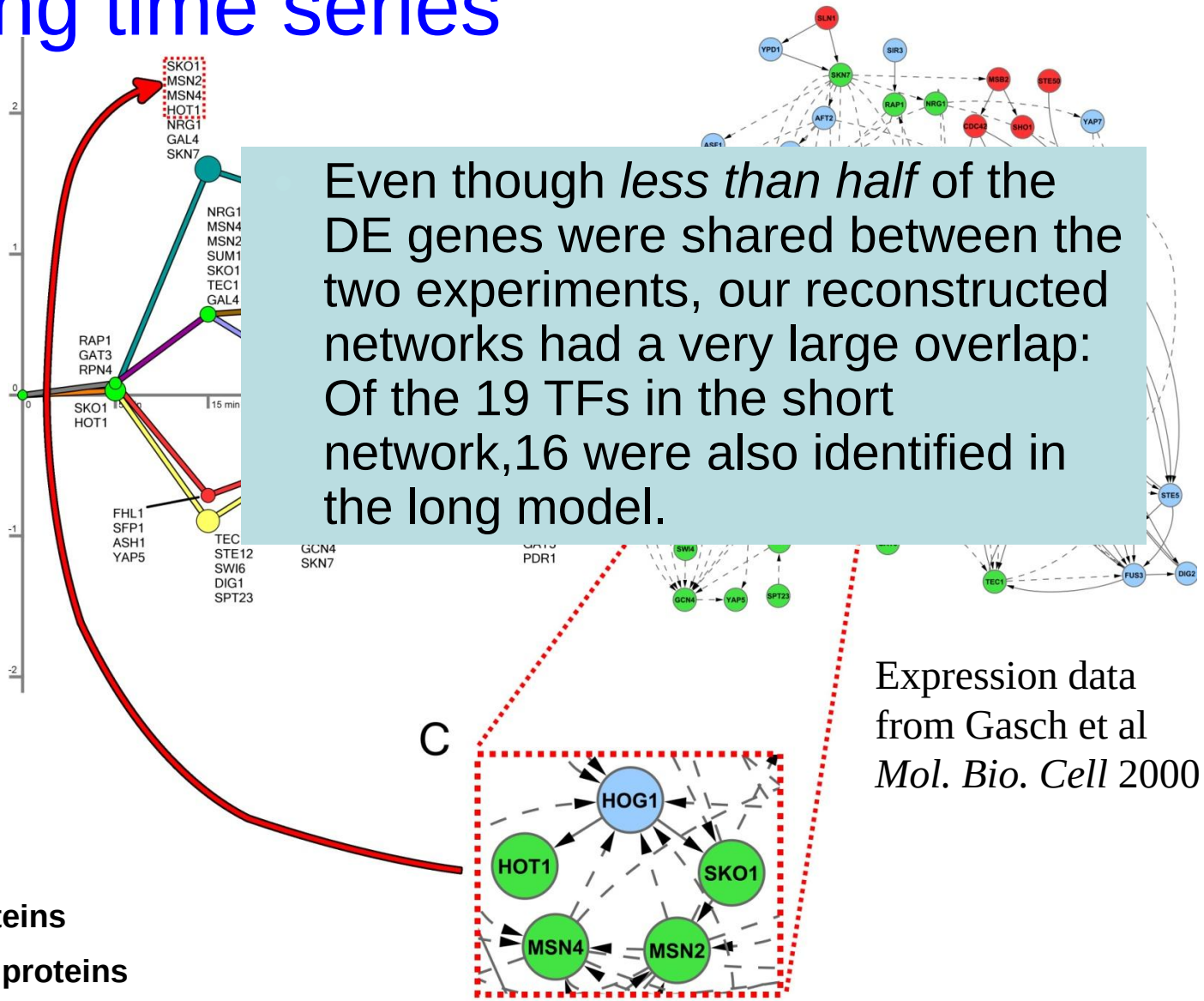


Expression data from Romero-Santacreu et al *RNA* 2009

- **input sensory proteins**
- **signaling network proteins**
- **transcription factors (from DREM)**

Gitter *et al Genome Research* 2013

Reconstructed HOG pathway: Long time series



Even though *less than half* of the DE genes were shared between the two experiments, our reconstructed networks had a very large overlap: Of the 19 TFs in the short network,16 were also identified in the long model.

Expression data from Gasch et al *Mol. Bio. Cell* 2000

- **input sensory proteins**
- **signaling network proteins**
- **transcription factors (from DREM)**

HOG predictions

- We used two different HOG time series expression datasets

Short time series

- 11 of 19 TFs in gold standard (59%)
- 27 of 39 internal proteins in gold standard (69%)

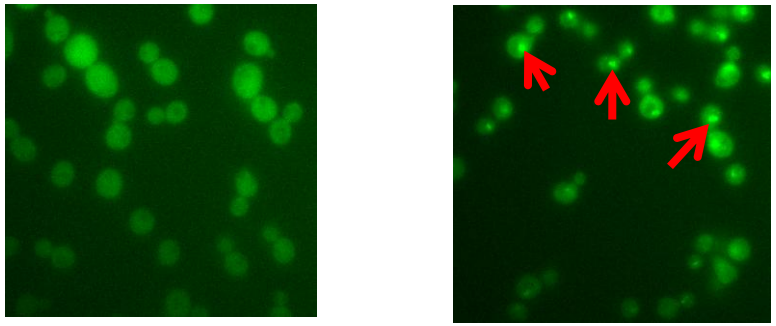
Long time series

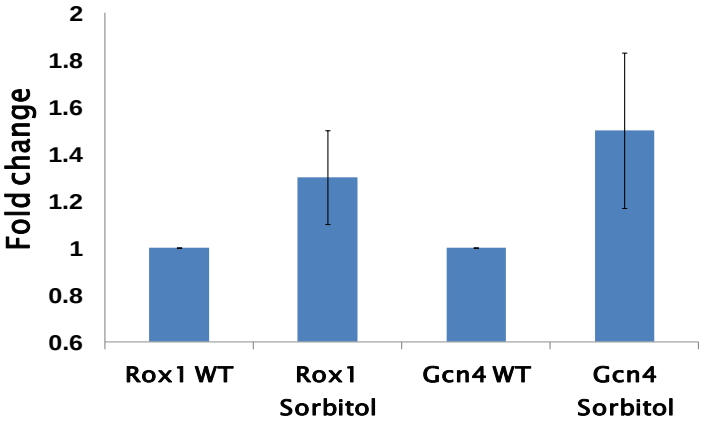
- 13 of 28 TFs in gold standard (46%)
- 16 of 23 internal proteins in gold standard (70%)

Validating computational predictions

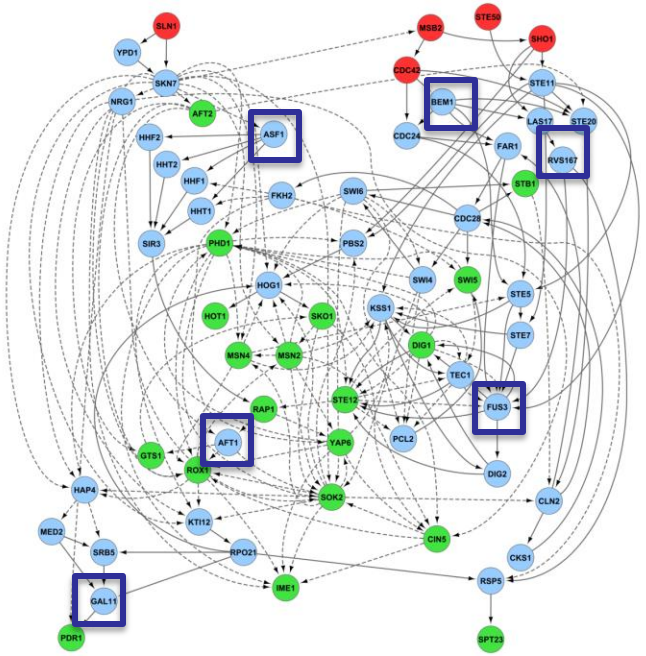
- Predictions for active TFs

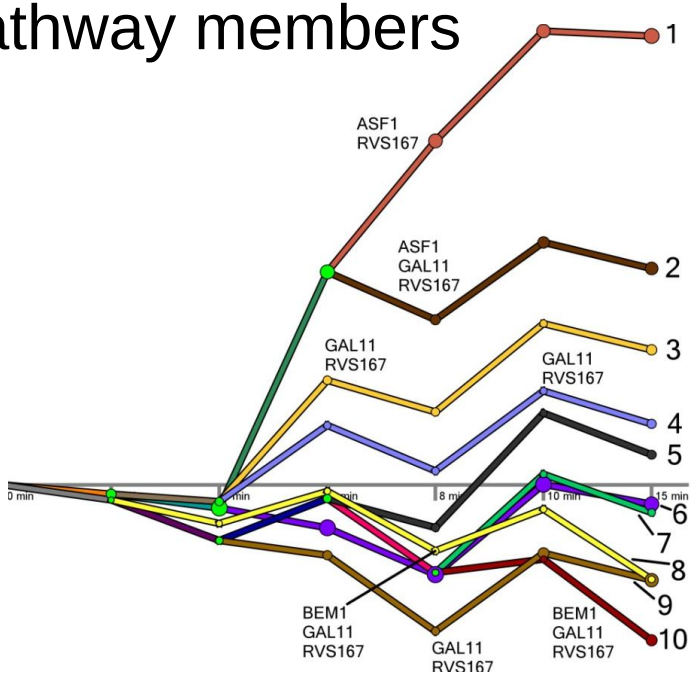
Before treatment **3 min after treatment**





- Knockouts validate internal pathway members



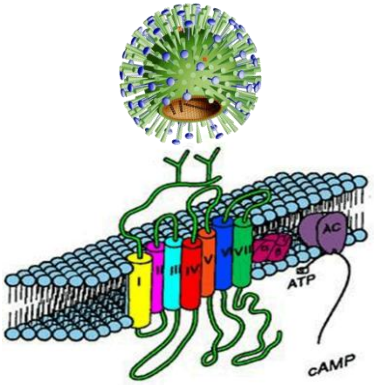


Gitter *et al Genome Research* 2013

Applications of SDREM to immune response

Applications to viral infection

- Treatments directly targeting virus can fail
  - Viral mutation
- Instead interested in host response



- Reconstruct infection model
  - Start with host-virus PPI
  - Find pathways in human PPI network
  - Connect the TFs driving observed transcription

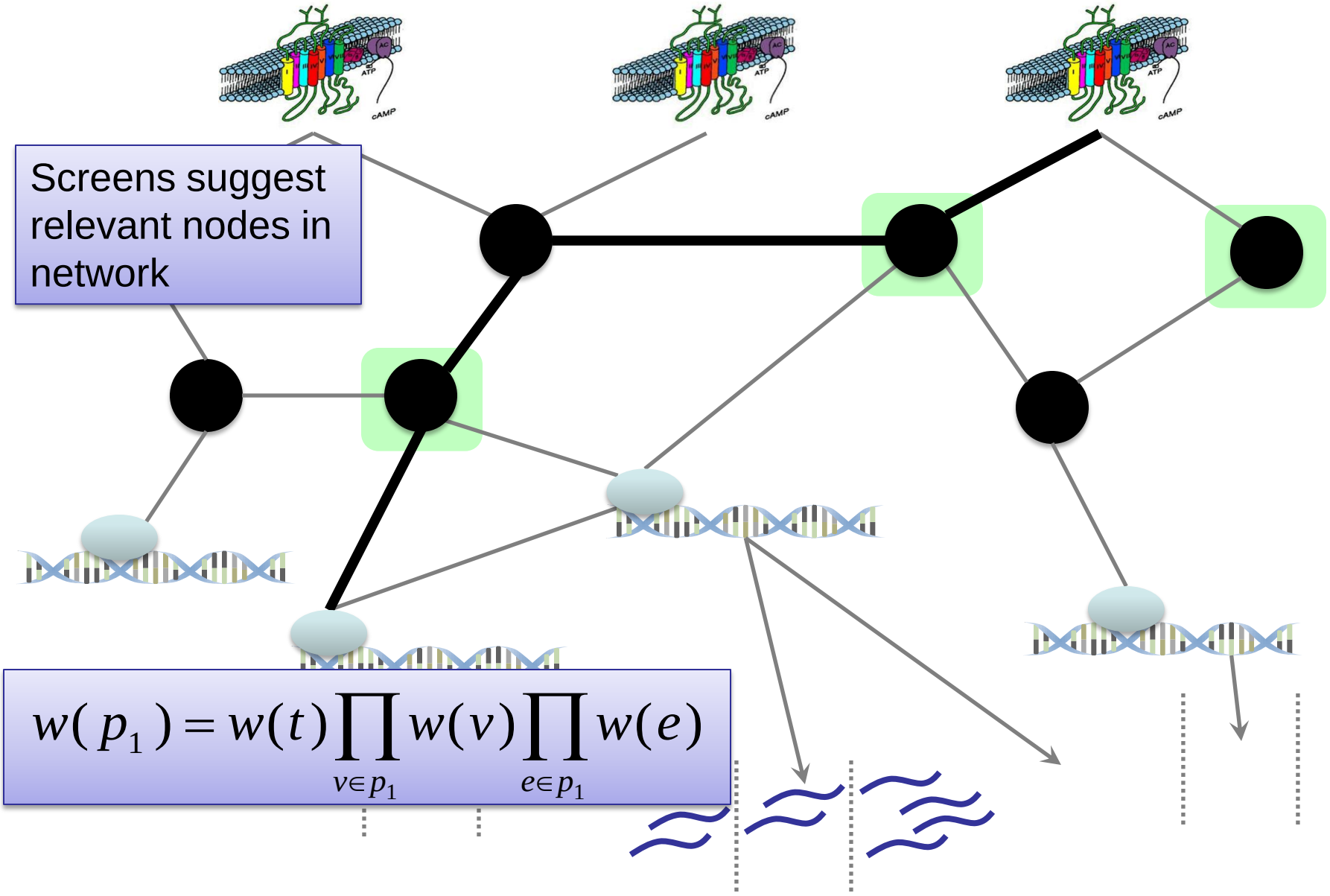
Disease-specific pathways

- Human networks contain millions of paths
  - Want pathways specific to the disease
- RNAi screening provides functional relevance

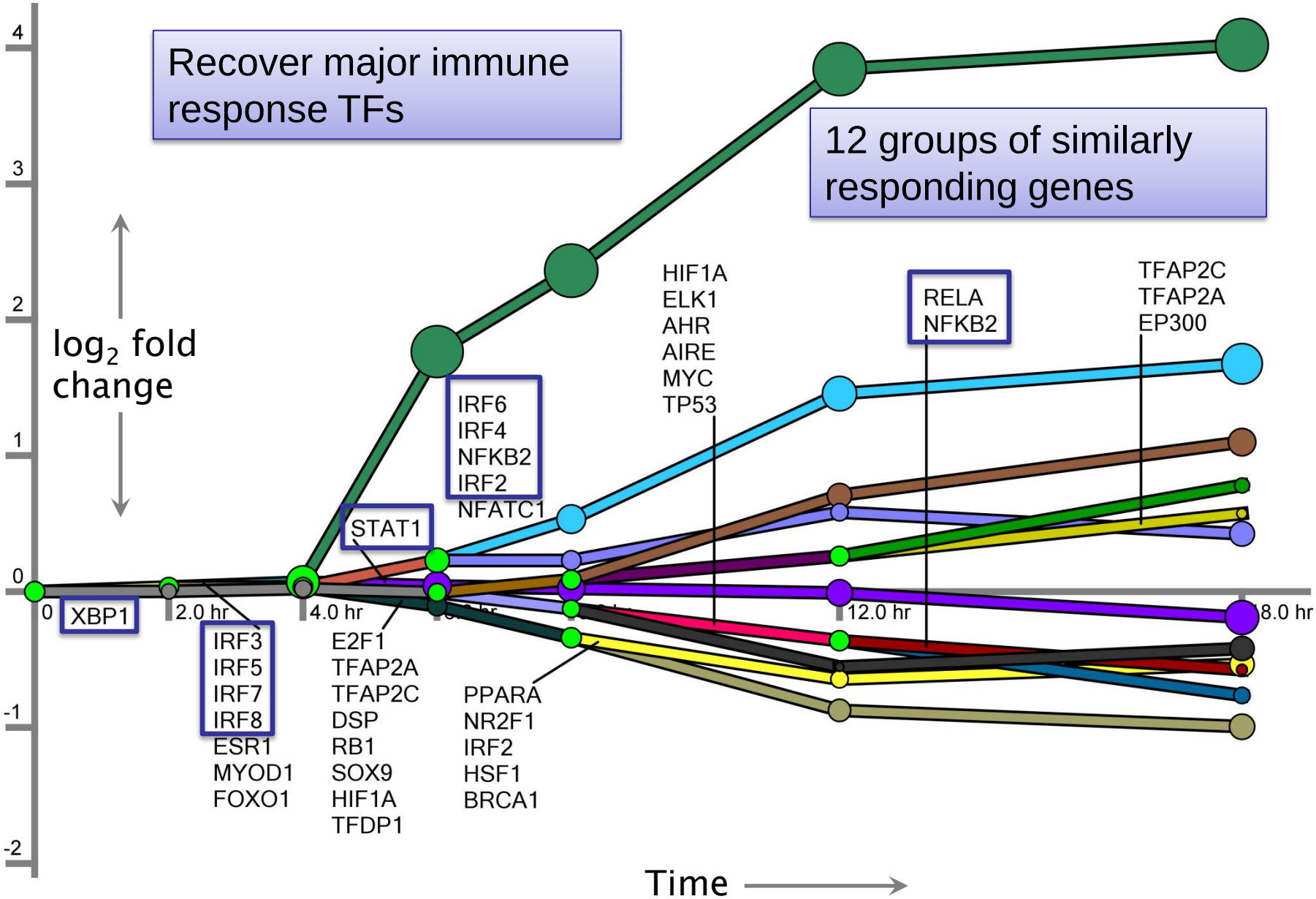
- Not reproducible across labs

Virus	Independent screens	Genes common to 3+ screens
H1N1 influenza	5	0.7%
HIV	3	0.4%

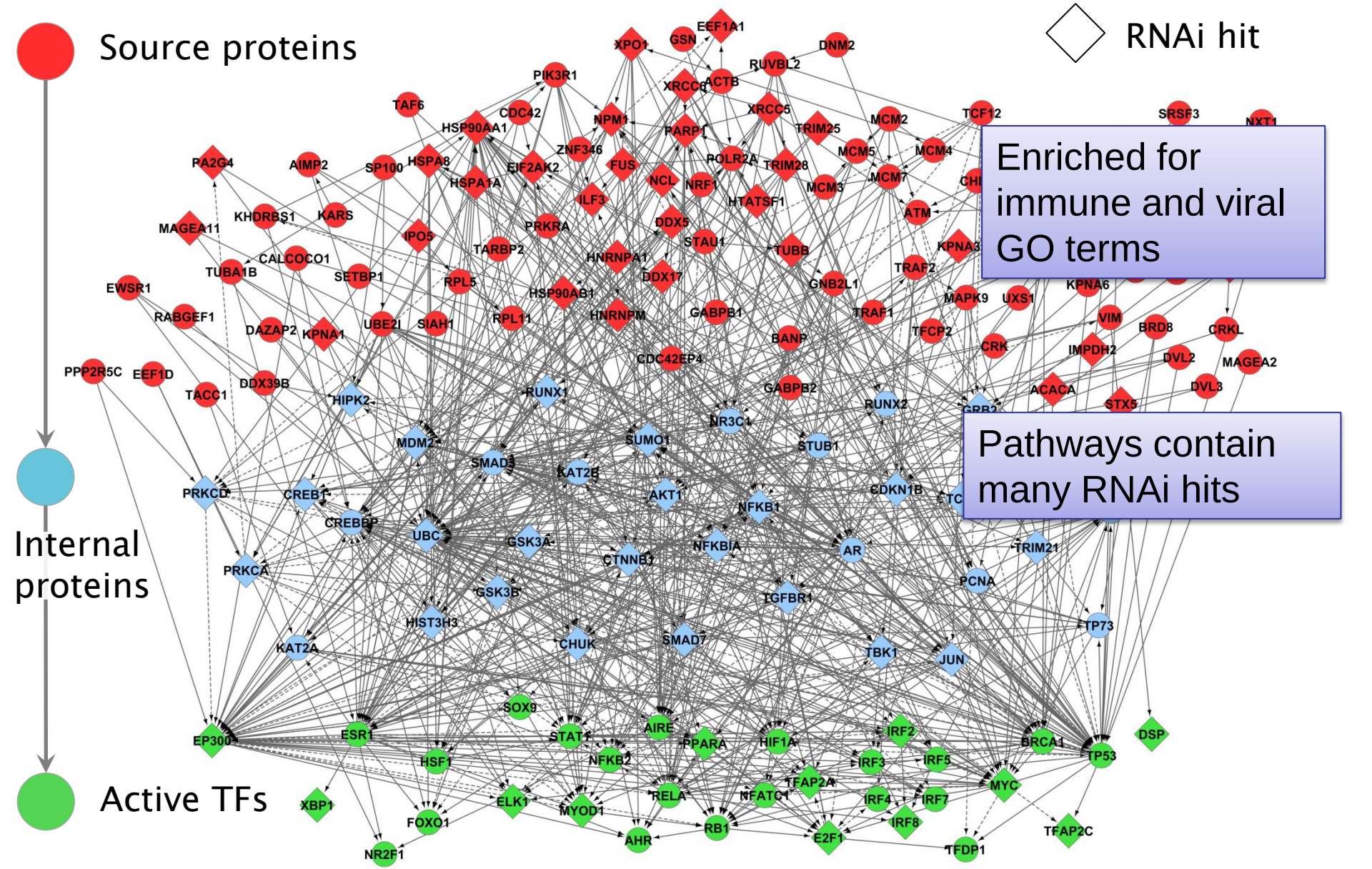
Incorporating RNAi screens



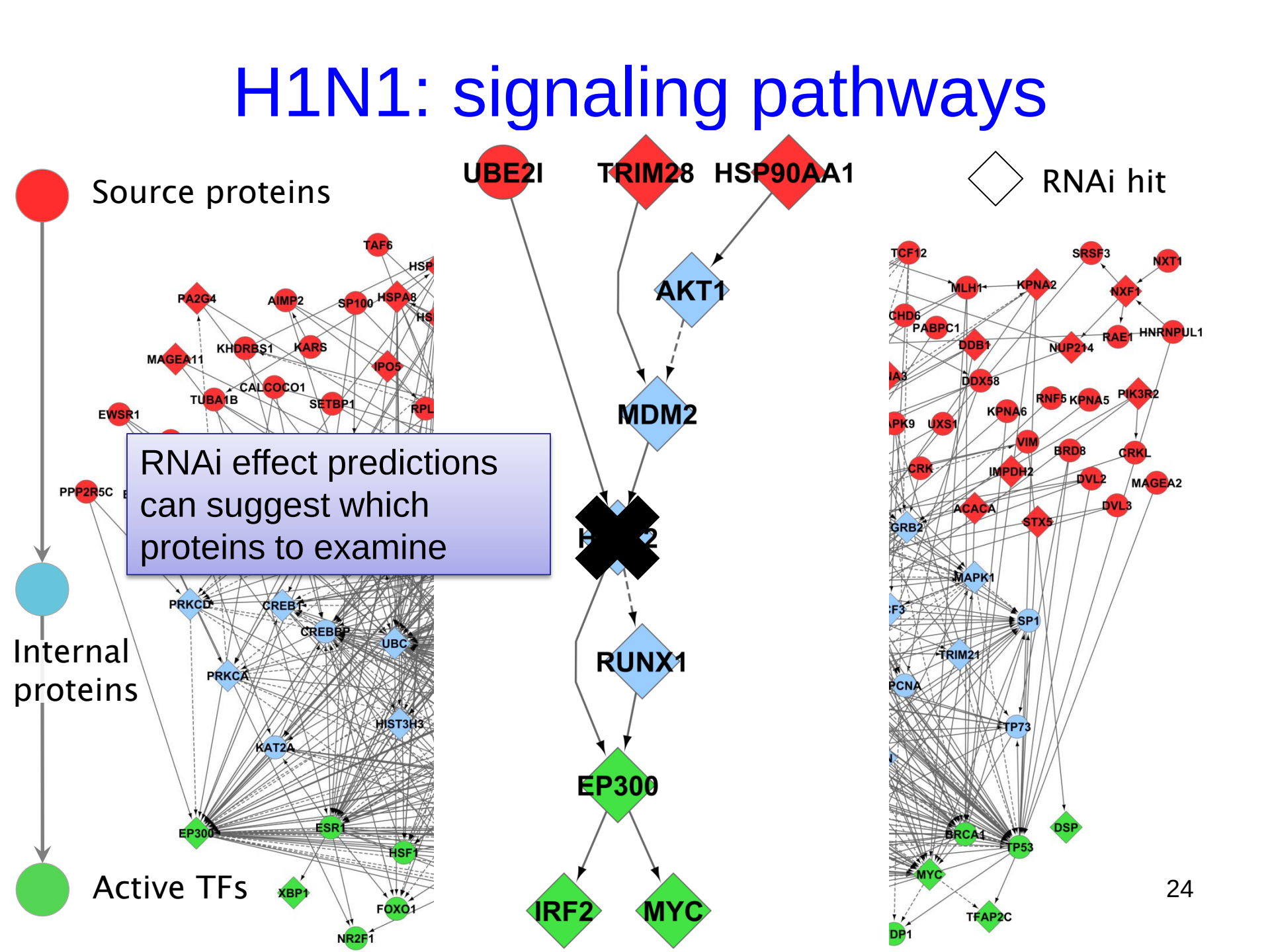
H1N1: temporal TF activity



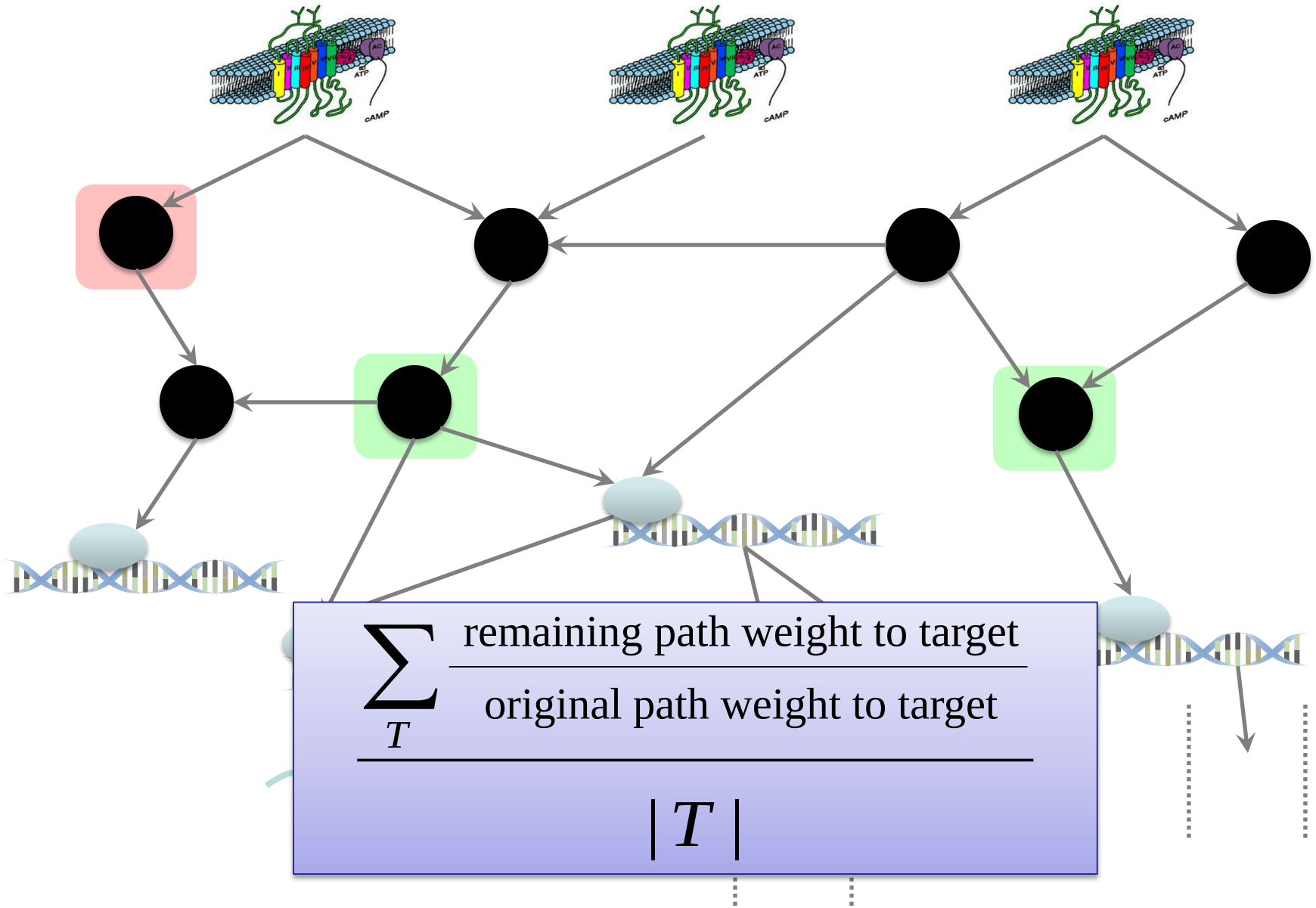
H1N1: signaling pathways



H1N1: signaling pathways



Predicting RNAi screen hits



$$\frac{\sum_{T} \frac{\text{remaining path weight to target}}{\text{original path weight to target}}}{|T|}$$

Predicting RNAi screen hits

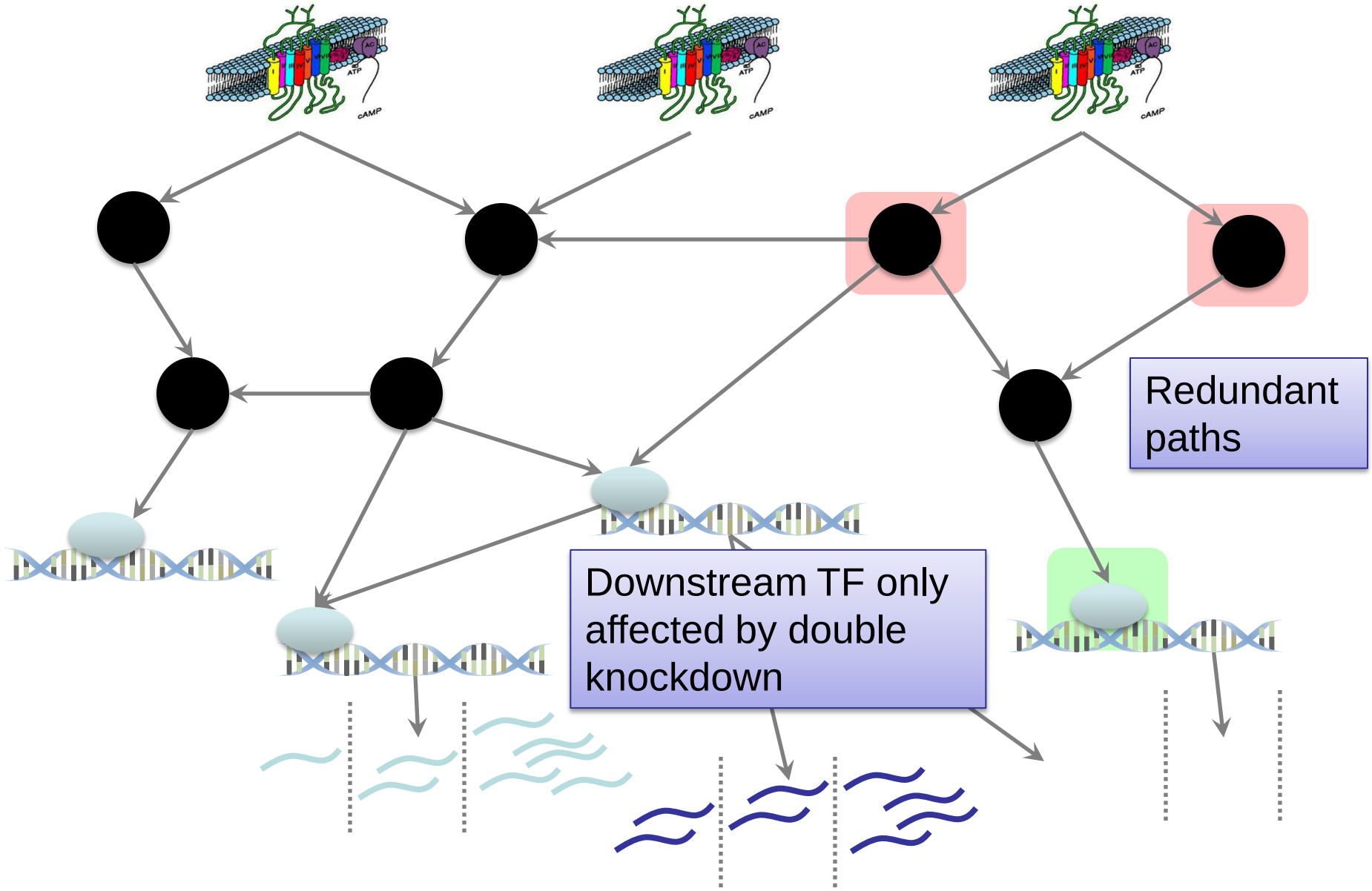
- Hold out all RNAi data and rerun SDREM
- Predict H1N1 knockdown effects

	Top 10	Top 20	Top 50	Top 100
Correct predictions	6	8	18	42
Significance	1.97 E-5	3.44 E-5	3.24E-9	9.42 E-23

- Can also be used to make predictions about double KO which are infeasible to test genome-wide
- SDREM produces ranked pairs to test
  - Disease-specific predictions

Gitter *et al ISMB (Bioinformatics)* 2013

Predicting genetic interactions



Predicting genetic interactions

- Genetic interaction definition from pioneering yeast studies

$$g = ob_{AB} - ex_{AB} = ob_{AB} - ob_A ob_B$$

- *In silico* observed phenotype is calculated using the directed pathways

$$ob_A = \frac{\sum_T \frac{\text{remaining path weight to target}}{\text{original path weight to target}}}{|T|}$$

- Calculate same *in silico* phenotype for double knockdowns