

Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*

Lee et al. Science. 2002 Oct

Transcription (again)

- DNA → RNA → Protein (central dogma)
- TranSCRIPTION- $\text{ACTG}_{\text{DNA}} \rightarrow \text{ACUG}_{\text{RNA}}$
 - Proceeds from 3' to 5' on DNA
- TRANSLATION- RNA → Protein
 - Proceeds from 5' to 3' on RNA; creates protein N-terminus to C-terminus

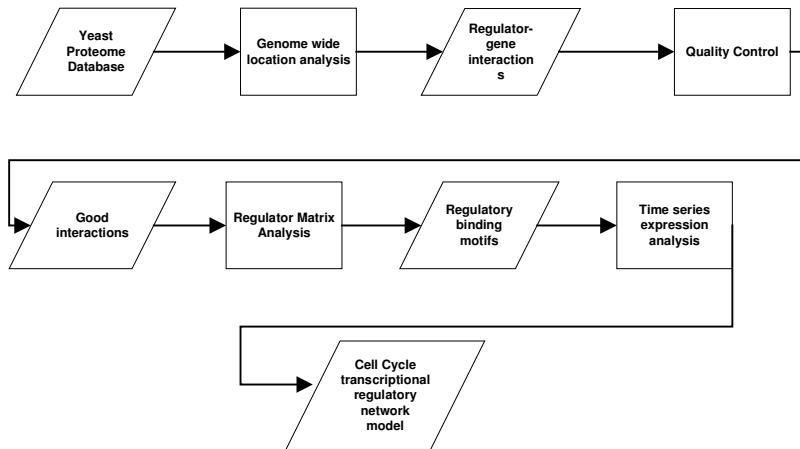
~Terminology

- **Transcription factor AKA regulator**-is a protein that binds DNA at a specific promoter region or site, where it regulates transcription
- **Promoter**-(core) a DNA sequence that enables a gene to be transcribed by facilitating the binding of RNA polymerase
- **Promoter**-(proximal, distal) a DNA sequence that is bound by a regulator
- **Antibody**-(immunoglobulin) a protein synthesized by an immune system cell (B lymphocyte) that acts as the causative agent for an immune system response by binding to foreign molecules
- **Epitope**-the part of a foreign molecule that is recognized by an antibody
- **Epitope tag**-a tag placed on a molecule of interest that has affinity for an antibody to be used

Purpose

- Find the regulator-gene interactions in Yeast
- Use the interactions to create network motifs (an abstracted set of interactions)
- Use motifs to construct a transcriptional regulatory network

Experimental Design



Epitope Tagging of Regulators

1. Amplify a segment of DNA that contains Myc epitope, TRP selectable marker, and sequences designed to recombine with the 3' end of the regulator

5' → ----- 3' Translation of regulator RNA

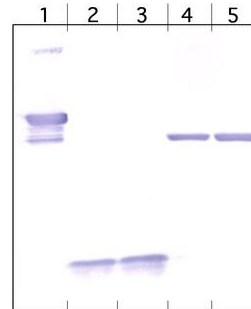
5' -Myc epitope-TRP marker- 3' PCR product

2. Transform the yeast strain with the PCR product (homologous recombination)
3. Select for clones of the yeast strain that have the PCR product by growing on TRP-plate
4. Confirm correct insertion by PCR and non-botched expression of the regulator by Western blotting using anti-Myc antibody

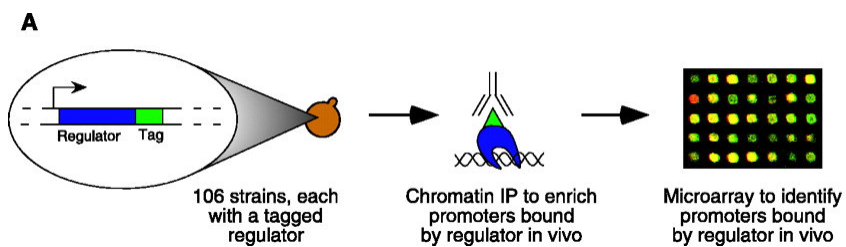
PCR and Western Blot

- Polymerase Chain Reaction
 - Taq polymerase and nucleotides
 - Cycle annealing and elongation

- Western blotting
 1. Gel electrophoresis
 2. Probe with antibody

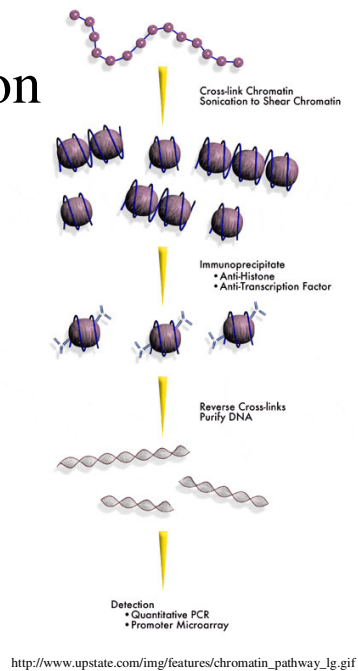


Genome-wide location analysis



Chromatin Immunoprecipitation (ChIP)

1. Cross-link proteins to DNA using formaldehyde (effects amino groups NH_2)
2. Smash chromatin to pieces
3. Grab the ones with the epitope
4. Undo Cross-linking
5. Amplify/dye IP-enriched sequences
6. Amplify/dye DNA that's not IP-enriched
7. Analyze binding of IP-enriched sequences to microarray with all intergenic sequences



Microarray Data Analysis

- Two channels IP and genomic DNA
- Normalization
 - Median intensity of control blanks subtracted
 - Genomic Normalization Factor = Median intensity of IP channel / Median intensity of genomic channel
 - Log-ratio of IP/Genomic for each intergenic region across all hybridization experiments

Microarray Data Analysis

- All log ratios for specific region were normalized by subtracting average log ratio for that intergenic region (bias in IP procedure)
- Adjusted intensity values calculated from ratios
- Chip error model used to calculate confidence values for each intergenic region (Rosetta Inpharmatics)

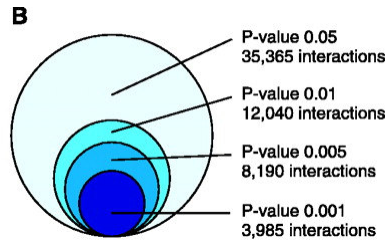
Chip error model

- According to this error model, the significance of a measured ratio at a spot is defined by a statistic X , which takes the form

$$X = \frac{a_2 - a_1}{(\sigma_1^2 + \sigma_2^2 + f^2(a_1^2 + a_2^2))^{1/2}}$$

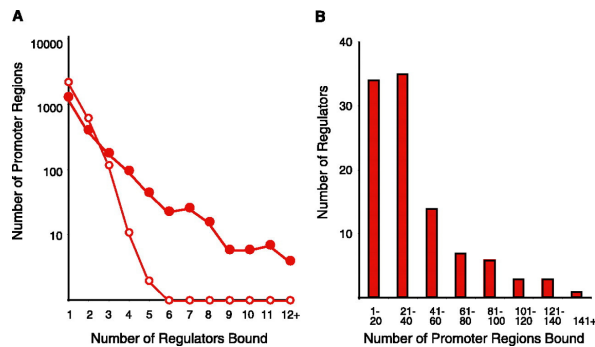
- where $a_{1,2}$ are the intensities measured in the two channels for each spot, $\sigma_{1,2}$ are the uncertainties due to background subtraction, and f is a fractional multiplicative error such as would come from hybridization non-uniformities, fluctuations in the dye incorporation efficiency, scanner gain fluctuations, etc. The distribution of X across all spots on a chip is approximately normal. The parameters $\sigma_{1,2}$ and f were chosen based on control hybridizations such that X had unit variance. The significance of a change of magnitude x is then calculated as $p = 1 - \text{Erf}(X)$

Resulting interactions



- The number of promoters bound by a regulator (0-181) for p-value=.001
- Average 38 promoters per regulator

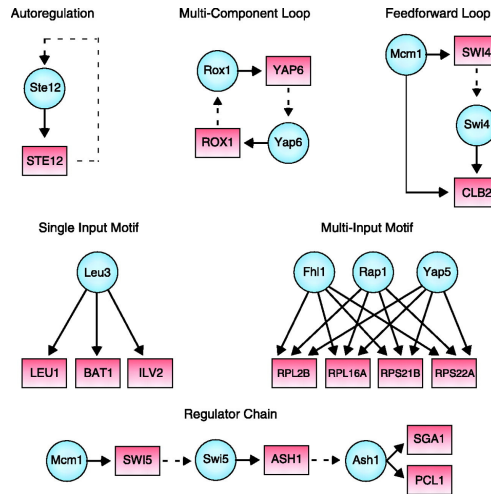
Distribution of Interactions



Genome-wide distribution of transcriptional regulators. **(A)** Plot of the number of regulators bound per promoter region. The distribution for the actual location data (red circles) is shown alongside the distribution expected from the same set of P values randomly assigned among regulators and intergenic regions (white circles). At a P value threshold of 0.001, significantly more intergenic regions bind four or more regulators than expected by chance. **(B)** Distribution of the number of promoter regions bound per regulator

Network Motifs

- Network Motif- "simplest units of commonly used transcriptional regulatory network architecture provide regulatory capacities such as positive and negative feedback loops"



Network Motifs

- Used a program called MEME to find motifs
- 39/106 regulators showed involvement in feedforward regulation
- What are the values for other motifs? They provide the number of motifs but not the number of regulators involved in each motif
- Motifs found
 - 188 chains ranging in size from 3 to 10 regulators
 - 81 multi-input motifs in our data, potentially regulating a total of 343 genes
 - 90 single input modules (we cannot assign these SIMS with high confidence using the location data obtained thus far because it is unlikely that we have assayed binding of all regulators under all conditions necessary to reveal true single input modules)
 - 3 multi-component loops
 - 49 feedforward loops
 - 10 autoregulation motifs

Creation of regulatory motifs

- Two data matrices
- Overall matrix D consists of binary entries D_{ij} , where a 1 indicates binding of regulator j to intergenic region i with a p-value of less than or equal to 0.001
- Regulator matrix R is a subset of D , containing only the rows corresponding to the intergenic region assigned to each regulator

Matrix evaluation

- **Autoregulatory motif:** Find each non-zero entry on the diagonal of R
- **Feedforward loop:** For each master regulator (column of R), find non-zero entries, which correspond to regulators bound. For each master regulator / secondary regulator pair, find all rows in D bound by both regulators
- **Multi-component loop:** For each regulator (column of R), find the regulators to which it binds. For each of these, find the regulators it binds. If any of these are the original regulator, you have a multi-component loop of two. For all others, find regulators to which they bind. If any of these are the original, you have a multi-component loop of three. Repeat to find larger loops
- **Single input module:** Find the intergenic regions bound by only one regulator. That is, take the subset of rows of D such that the sum of each row is 1. Then for each regulator (column), find non-zero entries. Each set (greater than three intergenic regions) is a SIM.
- **Multi-input module:** Find the intergenic regions bound by more than one regulator. That is, take the subset of rows of D such that the sum of each row is greater than 1. Then, for each row, find any other row bound by the same regulators. The collection of rows bound by the same regulators correspond to a MIM. Once a row is assigned to a MIM, remove it from further analysis
- **Regulator chain:** For each regulator (column of R), use a recursive algorithm to find chains of all lengths. That is, for each regulator whose promoter is bound by the regulator before it in the chain, find the regulator promoters to which it binds. Repeat until the chain ends. There are three possible ways to end a chain: a regulator that does not bind to the promoter of any other regulator, a regulator that binds to its own promoter, or one that binds to the promoter of another regulator earlier in the chain.

Assembling a network structure

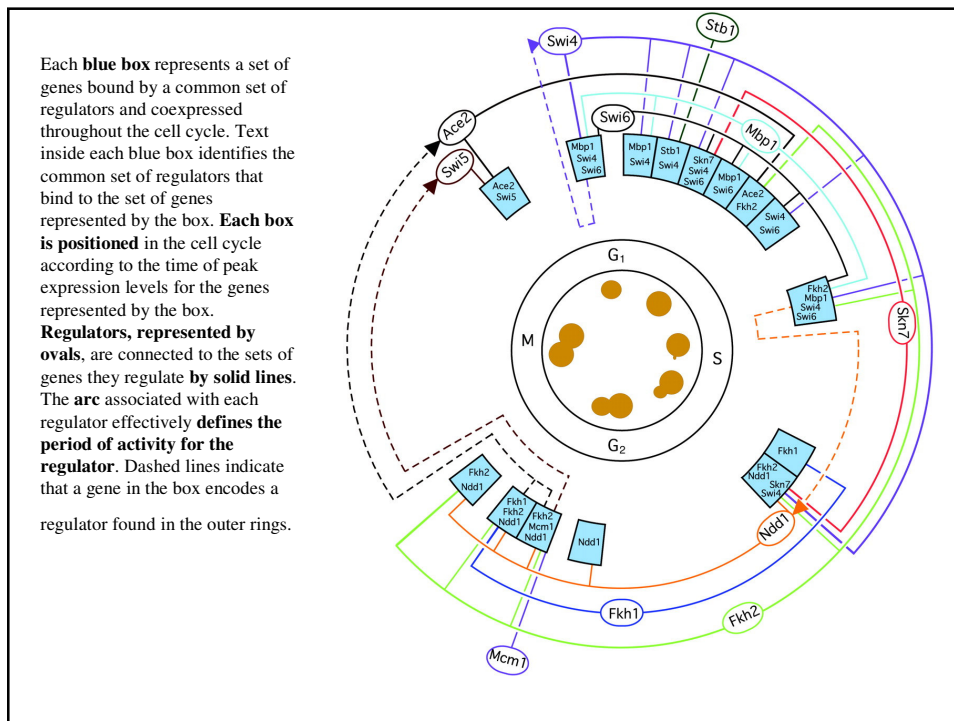
- Find a gene that is coordinately bound and expressed
- Define a set of genes G that are bound by a set of regulators S with $P\text{-value} \leq .001$
- Find a large subset that are similarly expressed and drop the rest
- Scan the remainder of the genome for genes with similar gene expression profiles

Assembling a network structure

- The resulting sets of regulators are multi-input motifs refined for common expression
- Use time series data to align the sets around the cell cycle on the basis of peak expression

Results

- Correctly assigns all regulation to stages of cell cycle
- Two regulators that are ill-defined could be assigned within the network on the basis of binding data
- Regulators show involvement in multiple cellular processes



Letters to nature
**Transcriptional regulatory code of a
eukaryotic genome**

Harbison et al.

NATURE | VOL 431 | 2 SEPTEMBER 2004

- Applies the same methodology
- Virtually the same paper
- Characterize regulator behavior under different conditions
- They find differences in binding under different environmental conditions

Results

