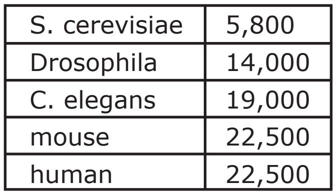
Lecture 22
Eukaryotic Genes and Genomes III

In the last three lectures we have thought a lot about analyzing a regulatory system in *S. cerevisiae*, namely **Gal** regulation that involved a hand full of genes. These studies monitored the increased transcription of **Gal** genes in the presence of galactose (and the absence of glucose); we saw that this regulation is achieved by particular proteins, or multiprotein complexes that bind to specific sequences in the promoter region upstream from their target genes.

What if I told you that it is now possible to do the following in *S. cerevisiae*:

- Monitor mRNA expression level for every gene in *S. cerevisiae*, in one single experiment.
- Monitor all the binding sites in the *S. cerevisiae* genome for each transcription factor in a single experiment.
- Determine all possible pair-wise interactions for every *S. cerevisiae* protein.

Obviously I wouldn't mention these possibilities if they weren't already happening. What I want to do today is to introduce you to the idea of carrying out genetic analyses on a global, genome-wide scale, and hopefully give you some examples that are relevant to what we have already learned along the way. So, this will be a technology oriented lecture, but with some application to what we have already learned about gene regulation in eukaryotes. It should also be mentioned that what will be described for *S. cerevisiae*, is theoretically possible for any organism whose genome has been completely sequenced and the location of all the genes in that genome have been established. What we will learn today is already being, or will be, applied to higher eukaryotes and mammals.



S. cerevisiae	5,800
Drosophila	14,000
C. elegans	19,000
mouse	22,500
human	22,500

Figure by MIT OCW.

Monitor mRNA expression level for every gene in *S. cerevisiae*, in one single experiment: Global transcriptional profiling.

Before we consider how it is possible to measure the levels of thousands of mRNA species, we will have to step back to consider how the levels of one or two mRNA species can be measured by Northern Blot analysis….and I know you must have learned this in 7.01 if not in high school. Northern blot analysis is based upon the fact that DNA and RNA molecules that possess complementary base sequences will hybridize together to form a double stranded molecule. If the complementarity is perfect the duplex molecule is stable, if it is imperfect (with base pair mismatches) it is relatively less stable. This provides the specificity needed to identify perfectly matched DNA:RNA duplexes (on Northern Blots) and DNA:DNA duplexes (on Southern Blots). This specificity is needed to be sure we are measuring the level of one particular transcript and that this is not contaminated with signal from closely related transcripts. RNA is isolated from cells, size fractionated on a gel; the thousands of mRNAs species form a smear on the gel which is punctuated by the strong ribosomal RNA bands (28S and 18S) that do not interfere with the analysis.

Image removed due to copyright reasons.
Please see
http://www.accessexcellence.org/RC/VL/GG/nucleic.html

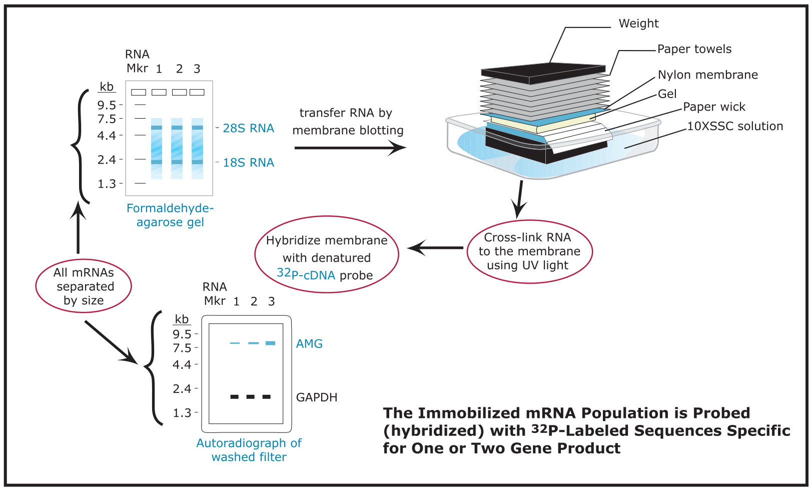

RNA
Mkr 1 2 3
kb
9.5
7.5
4.4
2.4
1.3
28S RNA
18S RNA
Formaldehyde-agarose gel
transfer RNA by membrane blotting
Weight
Paper towels
Nylon membrane
Gel
Paper wick
10XSSC solution
Cross-link RNA to the membrane using UV light
Hybridize membrane with denatured 32P-cDNA probe
All mRNAs separated by size
RNA
Mkr 1 2 3
kb
9.5
7.5
4.4
2.4
1.3
AMG
GAPDH
Autoradiograph of washed filter
The Immobilized mRNA Population is Probed (hybridized) with 32P-Labeled Sequences Specific for One or Two Gene Product

Figure by MIT OCW.

Northern Blots

Immobilized mRNA population hybridized with labeled DNA probe representing one or two genes

DNA Microarrays

Immobilized DNA probes representing **all** possible genes hybridized with labeled mRNA population

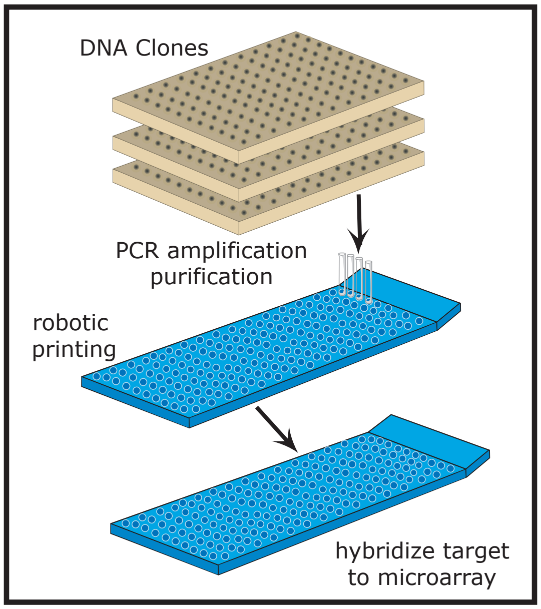
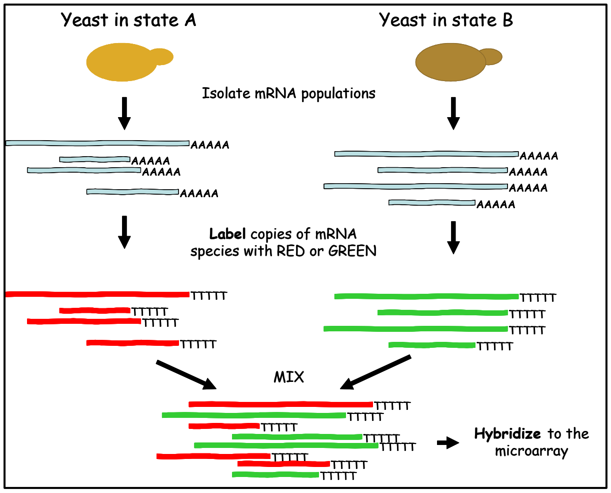
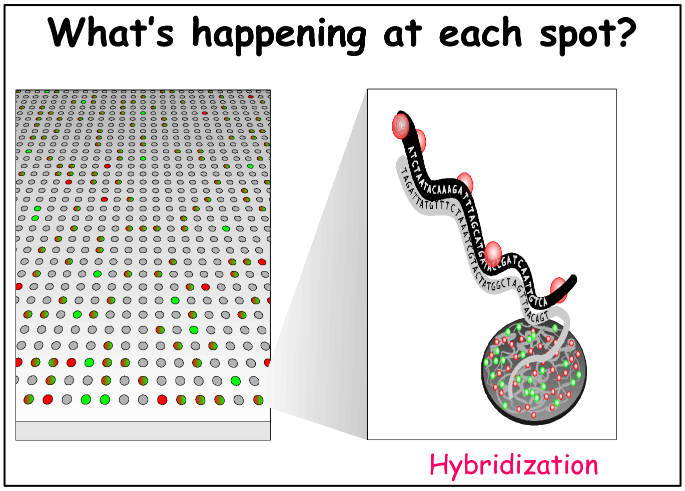


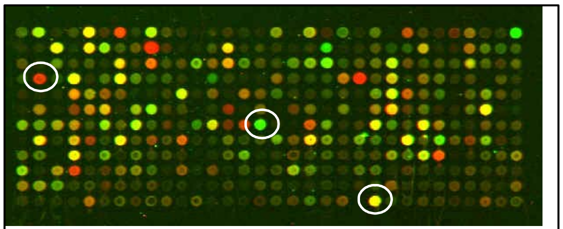
Figure by MIT OCW.

The breakthrough in developing microarrays for analyzing mRNA levels was to reverse the logic – instead of immobilizing the mRNAs for hybridization with one or two labeled complementary DNA (cDNA) probes, all possible cDNA probes are immobilized on a solid surface (usually glass slides). The spotting of probes is achieved robotically; the DNA probes are designed to specifically hybridize to only one nucleic acid sequence that represents a single mRNA species. The thousands of DNA probes are dispensed from 96-well, or 384-well plates to an addressable site on the solid surface. The mRNA population from each cell type purified and then copied such that the copy is fluorescently labeled. This fluorescent population is hybridized to the immobilized probes, and the intensity of the fluorescence at each probe spot is proportional to the number of copies of that specific mRNA species in the original mRNA population.

So let's look at how this would actually work in a real experiment. **mRNA** is isolated from yeast cells in **state A** (e.g., minus galactose) and from yeast cells in **state B** (e.g., plus galactose), and copies of each population is made such that one fluoresces red and the other fluoresces green. After mixing, these fluorescent molecules are hybridized to the slides containing ~5,800 DNA probes, each one specific for detecting hybridization of many copies of an individual mRNA species.


Yeast in state A
Yeast in state B
Isolate mRNA populations
AAAAA
Label copies of mRNA species with RED or GREEN
TTTTT
MIX
Hybridize to the microarray





🔴 mRNA present much higher in State A than State B

🟢 mRNA present much higher in State B than State A

🟡 mRNA present at equal levels in States A and B

The location and identity of each probe on the microarray slide is known, and each probe is specific for a single mRNA. The color and intensity of the fluorescence is measured by scanning the slide with lasers, and the relative abundance of each mRNA in the cells of **State A** vs **State B** can be calculated from the emitted fluorescence. i.e., the relative level of 5,800 mRNAs can be compared between two populations of yeast cells.

Presenting data for thousands of mRNA transcripts is clearly a challenge. You could present endless tables of data, but our brains are much more adept at recognizing shapes, patterns and colors. Colored representations of up and down regulation of transcripts levels is the preferred way to present data.

Northern Blot vs. Microarray

Each colored vertical line in the horizontal lane displays the relative expression level of a single mRNA

Images removed due to copyright reasons. Please see Lodish, Harvey, et. al. *Molecular Cell Biology*. 5th ed. New York : W.H. Freeman and Company, 2004.

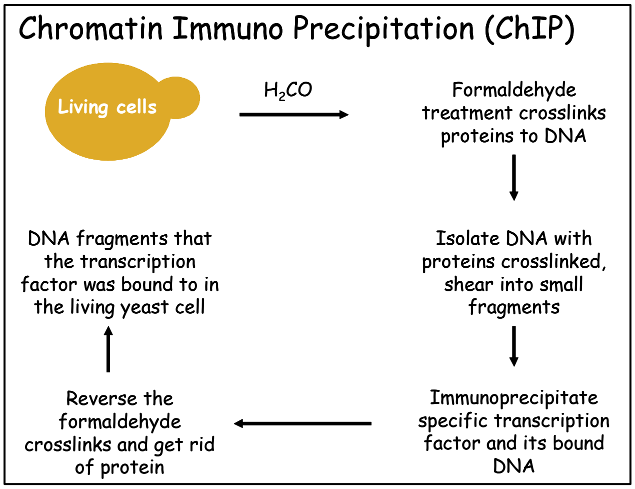
For our purposes here, let's look at what genes are up-regulated when a glucose grown culture of *S. cerevisiae* is shifted into galactose; what genes are up-regulated under these conditions? Obviously transcripts for **Gal1, Gal7** and **Gal10** genes will be up-regulated, as we have discussed in the last couple of lectures. In addition **Gal2** (galactose permease) and **Gal80** (the negative regulator of the **Gal4** transcriptional activator) are also induced; this was previously known, although we didn't discuss it directly in the previous lectures. But upon looking globally, it has become clear that some other genes are also up-regulated. (This figure shows just a small snapshot of the response.) These additional genes are **Fur4, Gcy1, Mth1,** and **Pcl10,** and their co-regulation along with the **Gal** genes was previously unrealized**.** We will be coming back to this later in the lecture.

What transcripts have increased levels when shifted from glucose to galactose?

Images removed due to copyright reasons.
Please see Ren, Bing., et.al. "Genome-wide Location and Function of DNA Binding Proteins."
Science 290, no. 5500 (Dec. 22, 2000): 2306-9.

Monitor all the binding sites in the *S. cerevisiae* genome for each transcription factor in a single experiment.

In the last lecture we talked about deletion analysis of cis-acting regulatory sequences identifying the location of **UAS** and **URS** sequences upstream of the **Gal1** gene. That the **Gal4** transcriptional activator protein binds to the DNA sequence present at the **URS$_{\mathbf{GAL1}}$** can be shown to happen in the test tube, but showing that it is actually bound in a living cell is another matter. A method was recently developed for doing just that, and this method has been further developed to determine transcription regulator binding across the whole genome.



Chromatin Immuno Precipitation (ChIP)

Living cells
H_2CO
Formaldehyde treatment crosslinks proteins to DNA
Isolate DNA with proteins crosslinked, shear into small fragments
Immunoprecipitate specific transcription factor and its bound DNA
Reverse the formaldehyde crosslinks and get rid of protein
DNA fragments that the transcription factor was bound to in the living yeast cell

Images removed due to copyright reasons. Please see
Figure 2 in Weinmann, Amy S. Novel ChIP-based Strategies to Uncover Transcription Factor Target Genes in the Immune System.
Nature Reviews Immunology 4 (2004): 381-386.

This method takes advantage of the fact that formaldehyde crosslinks proteins to DNA in a way that can later be reversed.

For galactose grown yeast cells chromatin immunoprecipitation (ChIP) with an antibody that pulls down the **Gal4** protein revealed some surprises. In addition to confirming that **Gal4** binds to the promoters regions upstream of the expected **Gal** genes, the **Gal4** protein also binds to the promoter regions of 4 other genes, namely **Fur4, Pcl10, Mth1** (shown in the adjacent figure) and **Gcy1** (not shown). Note that these genes were shown to be induced by galactose in the previous section. Just how the up-regulation of **Fur4, Pcl10** and **Mth1** might contribute to optimizing the metabolism of galactose is shown in this figure, but the role **Gcy1** plays is unclear. Clearly, taking a global look at what genes are up-regulated in the presence of galactose, and taking a global look at what promoters are bound by the **Gal4** regulator, has clearly enriched our view of how *S. cerevisiae* adapts to the presence of this sugar.

A more complete view of galactose induced gene expression in *S. cerevisiae*

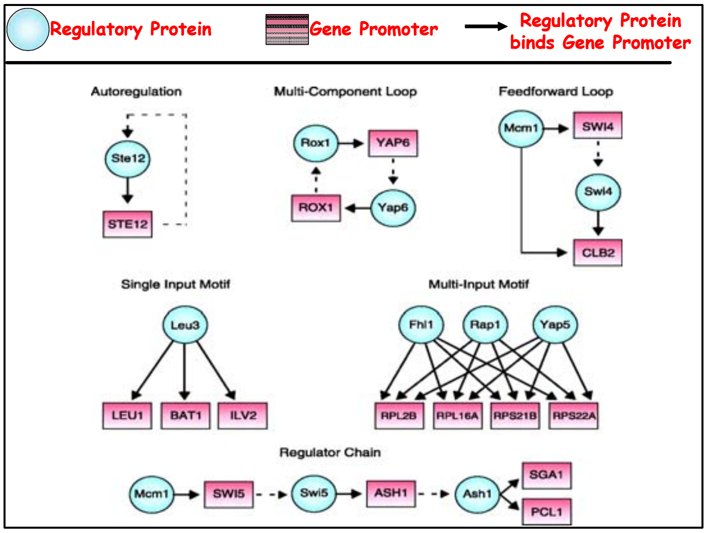
Images removed due to copyright reasons.
Please see Ren, Bing, et. al. "Genome-wide Location and Function of DNA Binding Proteins." *Science* 290, no. 5500 (Dec. 22, 2000): 2306-9.

The ChIP approach, followed by hybridization to DNA microarrays, was originally limited to monitoring binding of transcriptional regulators for which there were good precipitating antibodies. However, this limitation was recently eliminated by fusing an epitope TAG to each regulator gene. This epitope TAG is recognized by a strong antibody, and so a single antibody can "pull down" (immunoprecipitate) >100 different regulatory proteins, each of which is expressed in its own yeast strain.

Images removed due to copyright reasons.
Please see Ren, Bing, et. al. "Genome-wide Location and Function of DNA Binding Proteins." *Science* 290, no. 5500 (Dec. 22, 2000): 2306-9.

Arrayed probe sequences represent the upstream cis-acting regions of all 5,800 genes

This has enabled a massive study to identify all of the target genes for each of 106 transcriptional regulators in *S. cerevisiae* growing in a defined medium. A compilation of all the data has revealed a number of fundamentally different regulatory motifs; these are shown in the



adjacent figure. For the most part the **Gal4** regulatory network (not shown) represents a simple Single Input Motif.

This approach has already been extended to human cells and it will not be long until detailed regulatory mechanisms are defined for humans, in the way it is now happening in yeast. It is now possible to go on to monitor which genes the transcriptional regulators bind to under different environmental conditions, and from there to build more dynamic models for how these genetic regulatory mechanisms operate and ultimately how they co-operate with each other.

Determine all possible pair-wise interactions for every *S. cerevisiae* protein.

The third global scale analysis we will consider is the systematic determination of protein-protein interactions in *S. cerevisiae*. This essentially involves a systematic test of all pair-wise combinations between all 5,800 yeast proteins. Individual matings to test >33 million combinations isn't feasible, so mating pools of 100 strains in all combinations has become the preferred approach. Only the diploid strains where the **Gal4 DB-**fusion and the **Gal4 AD-**fusion proteins interact will be able to grow on galactose medium without uracil and histidine, as well as turning blue when grown on galactose and X-gal. The plasmids present in such diploids are then sequenced to determine which proteins are fused to the **Gal4 AD** and **DB** domains.

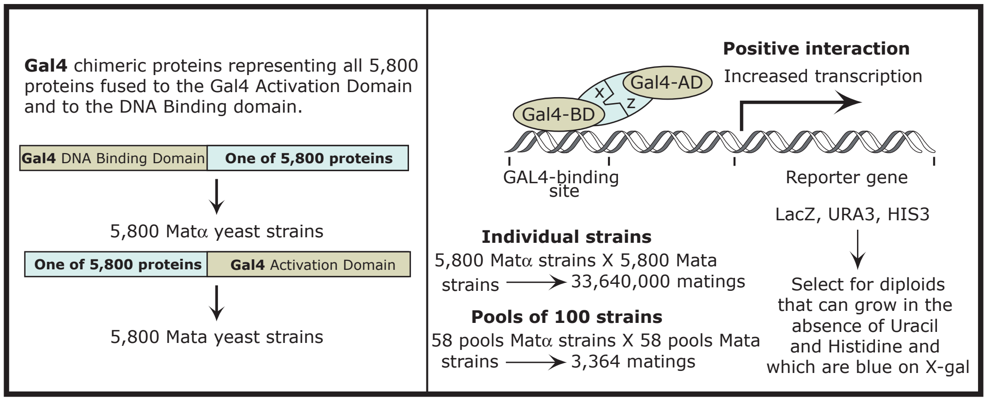
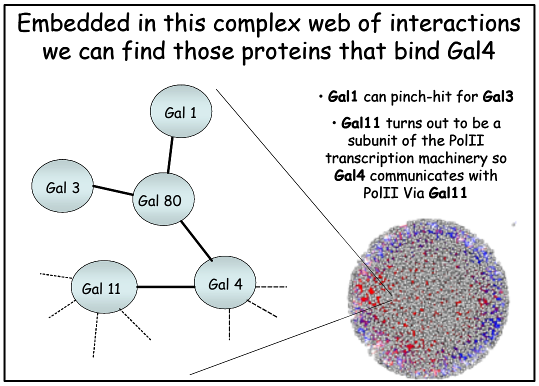

Gal4 chimeric proteins representing all 5,800 proteins fused to the Gal4 Activation Domain and to the DNA Binding domain.
Gal4 DNA Binding Domain
One of 5,800 proteins
5,800 Matα yeast strains
One of 5,800 proteins
Gal4 Activation Domain
5,800 Mata yeast strains
Positive interaction
Increased transcription
Gal4-AD
X
Z
Gal4-BD
GAL4-binding site
Reporter gene
LacZ, URA3, HIS3
Individual strains
5,800 Matα strains X 5,800 Mata strains → 33,640,000 matings
Pools of 100 strains
58 pools Matα strains X 58 pools Mata strains → 3,364 matings
Select for diploids that can grow in the absence of Uracil and Histidine and which are blue on X-gal

Figure by MIT OCW.

This systematic approach to cataloguing all possible protein-protein interactions for yeast proteins yielded many more interactions that originally thought. Admittedly the yeast two hybrid is quite noisy, giving many false positive interactions, but even so alternative methods (that we do not have time to consider in detail) have confirmed many of these interactions. When all of the known protein-protein interaction data is assembled, we see the surprising fact that > 5,000 proteins can be connected together by > 14,000 protein


Embedded in this complex web of interactions we can find those proteins that bind Gal4
Gal 1
Gal 3
Gal 80
Gal 11
Gal 4
• Gal1 can pinch-hit for Gal3
• Gal11 turns out to be a subunit of the PolII transcription machinery so Gal4 communicates with PolII Via Gal11

interactions in a continuous web. Indeed, the interaction data for **Gal4** embedded within this web makes sense and adds some new information. Such "Interactomes" are being developed for all the usual organisms, and the *C. elegans* interactome is particularly well developed. One of the major revelations has been that proteins from pathways that were previously thought to be totally unconnected, turn out to have interacting proteins.