A Case Study -- Chu et al.

- An interesting early microarray paper
- My goals
  - Show arrays used in a "real" experiment
  - Show where computation is important
  - Start looking at analysis techniques

1

## The Transcriptional Program of Sporulation in Budding Yeast

S. Chu, * J. DeRisi, * M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, I. Herskowitz

What is Sporulation?

- Under adverse conditions, one yeast cell transforms itself into "spores" -- tetrad of cells with tough cell wall, goes "dormant"
- Yeast is ordinarily diploid; spores are haploid. I.e., genetically, sporulation is analogous to formation of egg/sperm in most sexual organisms -- 2 rounds of meiotic (not mitotic) cell division.
  - And many of the genes/proteins involved in this are recognizably similar to human genes/proteins
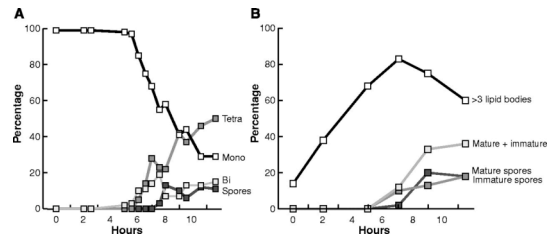
3
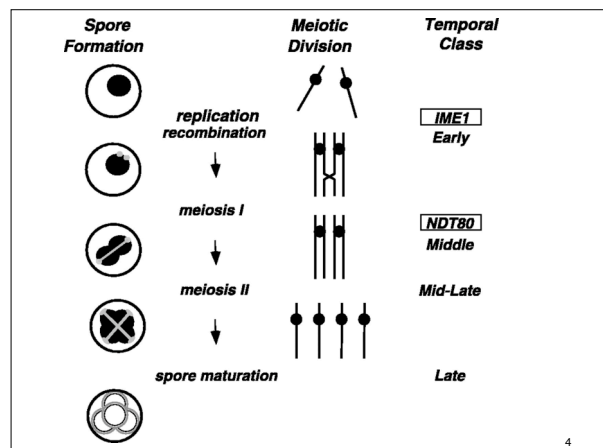
4

The Chu et al. Experiment

- Measure mRNA expression levels of all 6200 yeast genes in 7 time points (0-11 hours) in a (loosely synchronized) sporulating yeast culture
- Compare level at time t to level at time 0 on 2-color cDNA array
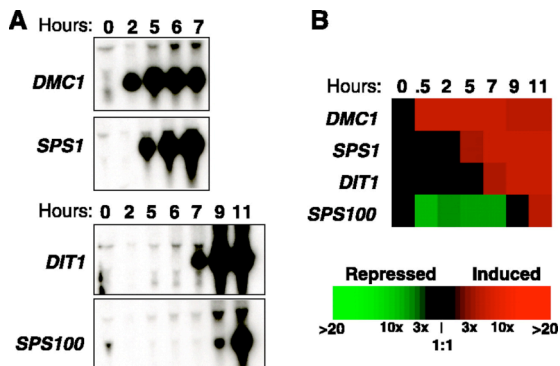- Plus some more standard tests as controls
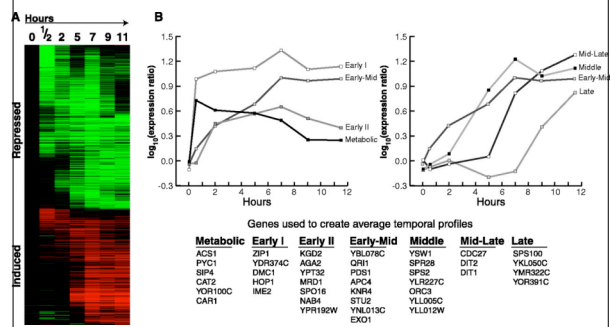
5

Measures of Sporulation

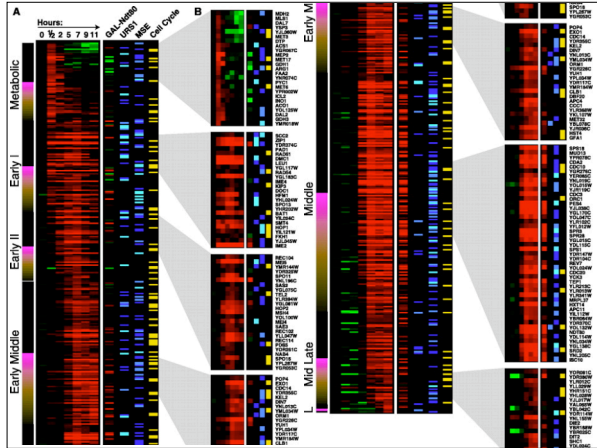NB: < 20% spores, so data are *mixtures* of cell stages

6

Standard Test (Northern) vs Array

Prototype Expression Profiles

8

"Sporulation" Summary, I

- What they did:
  - measured mRNA expression levels of all 6200 yeast genes in 7 time points in a (loosely synchronized) sporulating yeast culture
  - plus some more standard tests as controls
- What they learned:
  - 3-10x increase in number of genes implicated in various subprocesses
  - several subsequently verified by direct knockouts
  - further evidence for significance of some known transcription factors and/or binding motifs
  - several potential new ones
  - evidence for existence of others

10

"Sporulation" Summary, II

- Where computation fits in
  - automated sample handling
  - image analysis
  - data storage, retrieval, integration
  - visualization
  - clustering
  - sequence analysis
    - similarity search
    - motif discovery
  - structure prediction

More on these topics later in the course

11

More on Computation

- Similarity Search -- given a loosely defined sequence "motif", e.g. a transcription factor binding site, scan genome for "matches"
  - "Which genes have an MSE element?"
  - E.g., weight matrix models, Markov models
- Motif discovery -- given a collection of sequences presumed to contain a common pattern, e.g. a transcription factor binding site, find it & characterize it
  - "What motifs are common to Early Middle genes?"
  - E.g., MEME, Gibbs Sampler, Footprinter, …

12

More on Computation

- Finding groups of sequences that plausibly contain common sequence motifs
  - E.g., clustering (co-varying because co-regulated?)

13

Chu's "Supervised" Clustering

- Hand picked ~ 40 prototype genes
  - With significant variation in data set
  - With known function
- Hand-segregated into 7 groups ("Early", …)
- Assign all others to "nearest" group
  - Based on Pearson correlation to per-group averages of prototypes
- For visualization, order within groups by correlation to neighboring groups

14

Critique

+   -

15

2 warnings about arrays & clusters

- Warning 1: expression data often do not separate into nice, compact, well-separated clusters
  - Cf Raychaudhuri et al. (next 2 slides)
- Warning 2: it's hard to visualize high-dimensional data & inadequate visualization may obscure as well as enlighten
  - Cf last 2 slides.

16