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

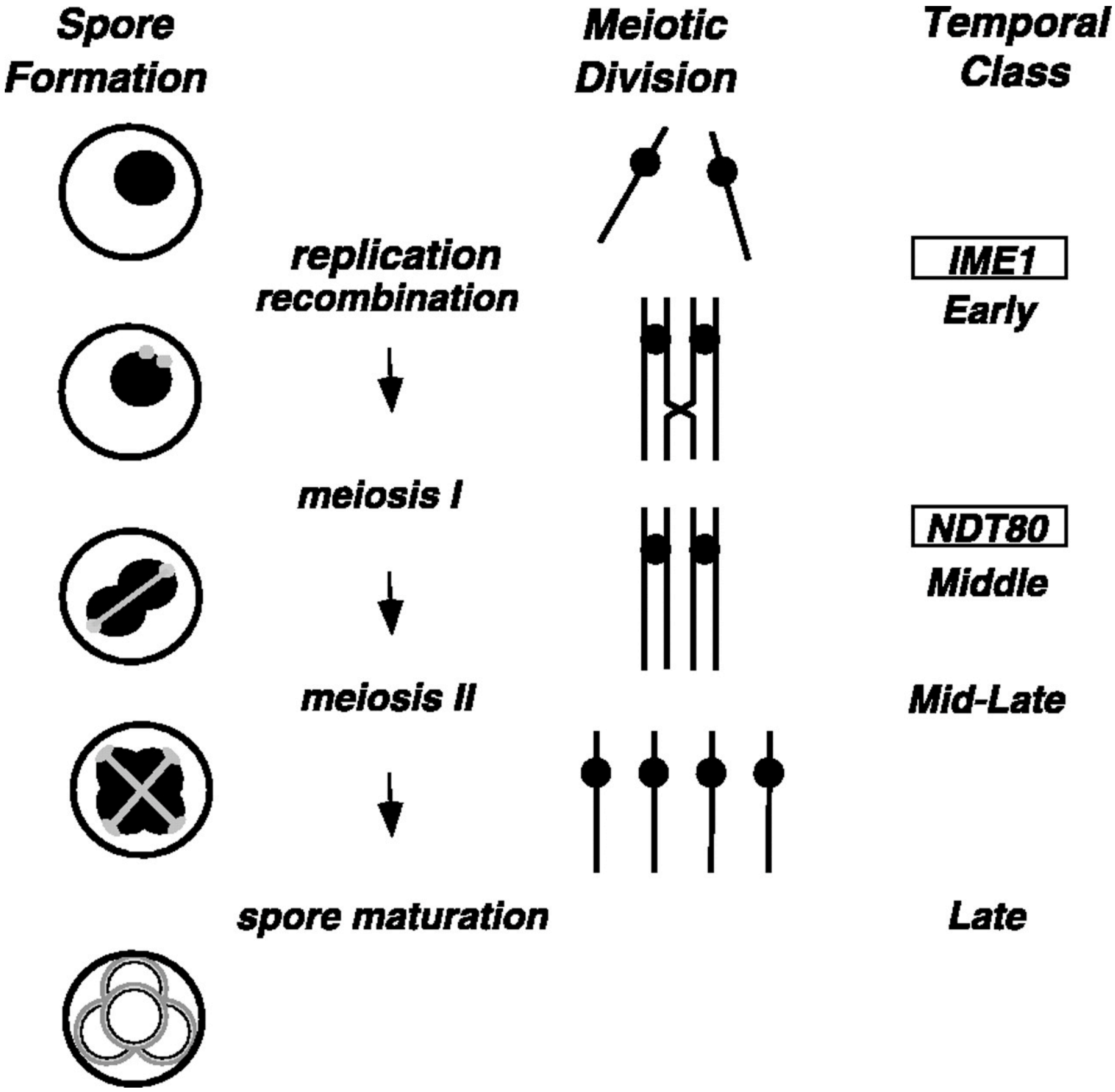
The Transcriptional Program of Sporulation in Budding Yeast

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

Science, 282 (Oct 1998) 699-705

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



Spore Formation

Meiotic Division

Temporal Class

replication
recombination

IME1

Early

meiosis I

NDT80

Middle

meiosis II

Mid-Late

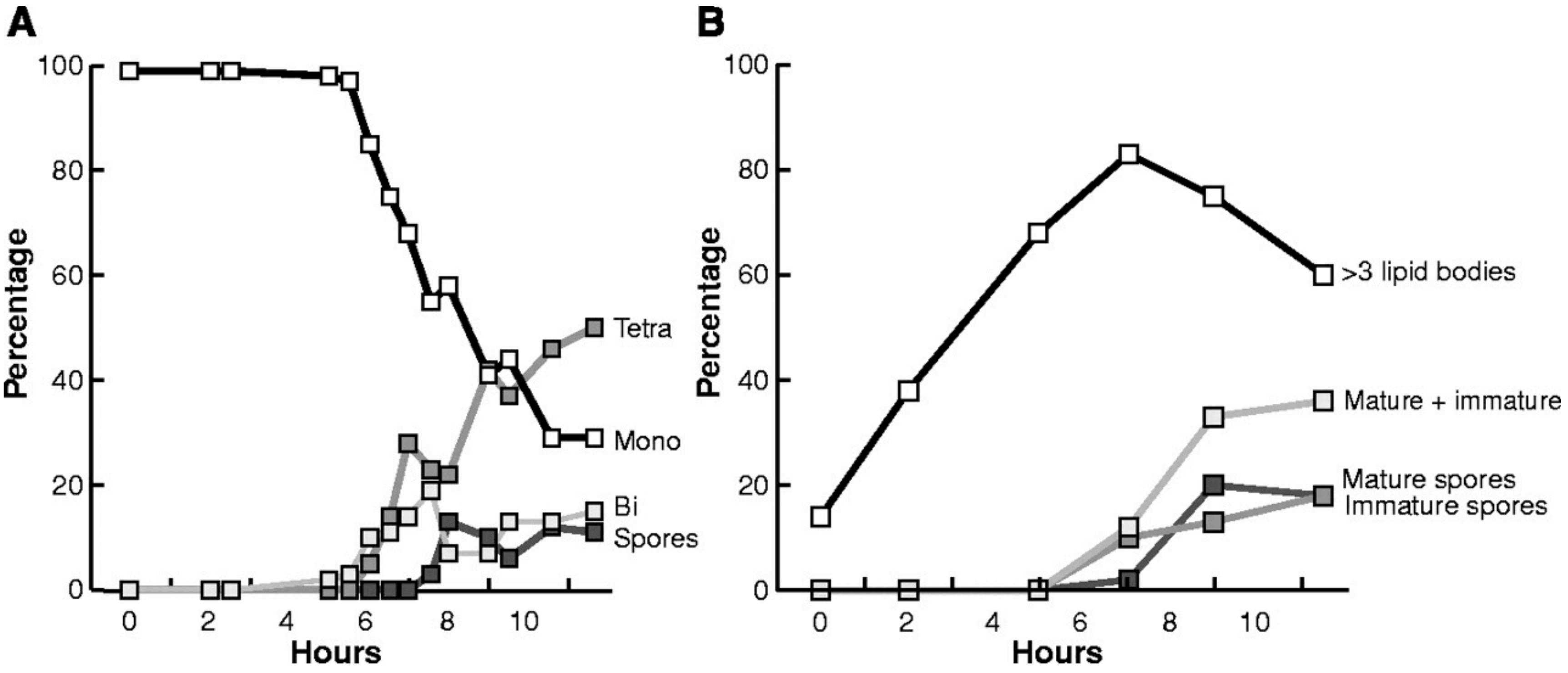
spore maturation

Late

The Chu et al. Experiment

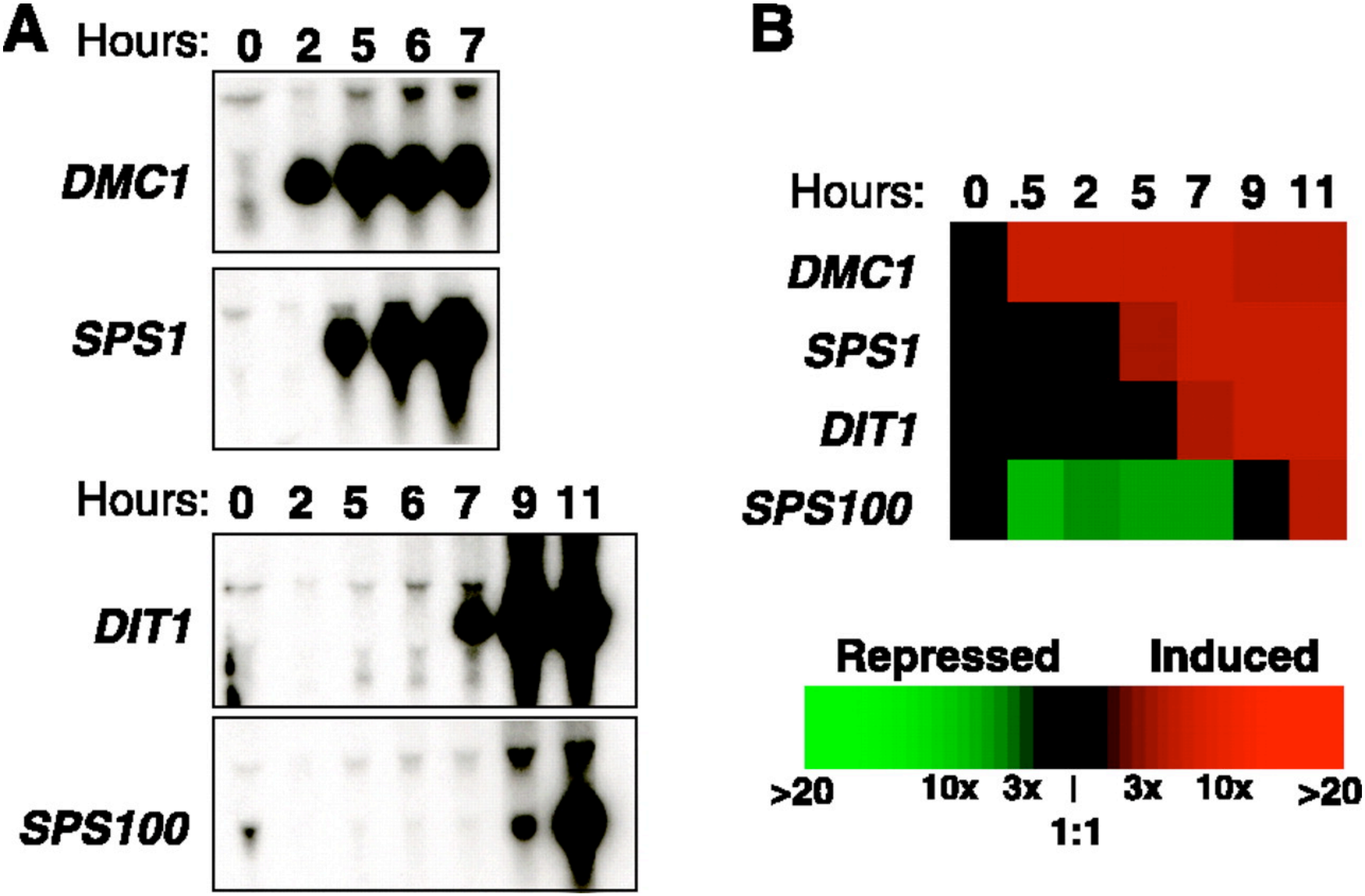
- Measure mRNA expression levels of all 6200 yeast genes in 7 time points (0-11 hours) in a (loosely synchronized) sporulating yeast culture
- Compare level at time t to level at time 0 on 2-color cDNA array
- Plus some more standard tests as controls

Measures of Sporulation

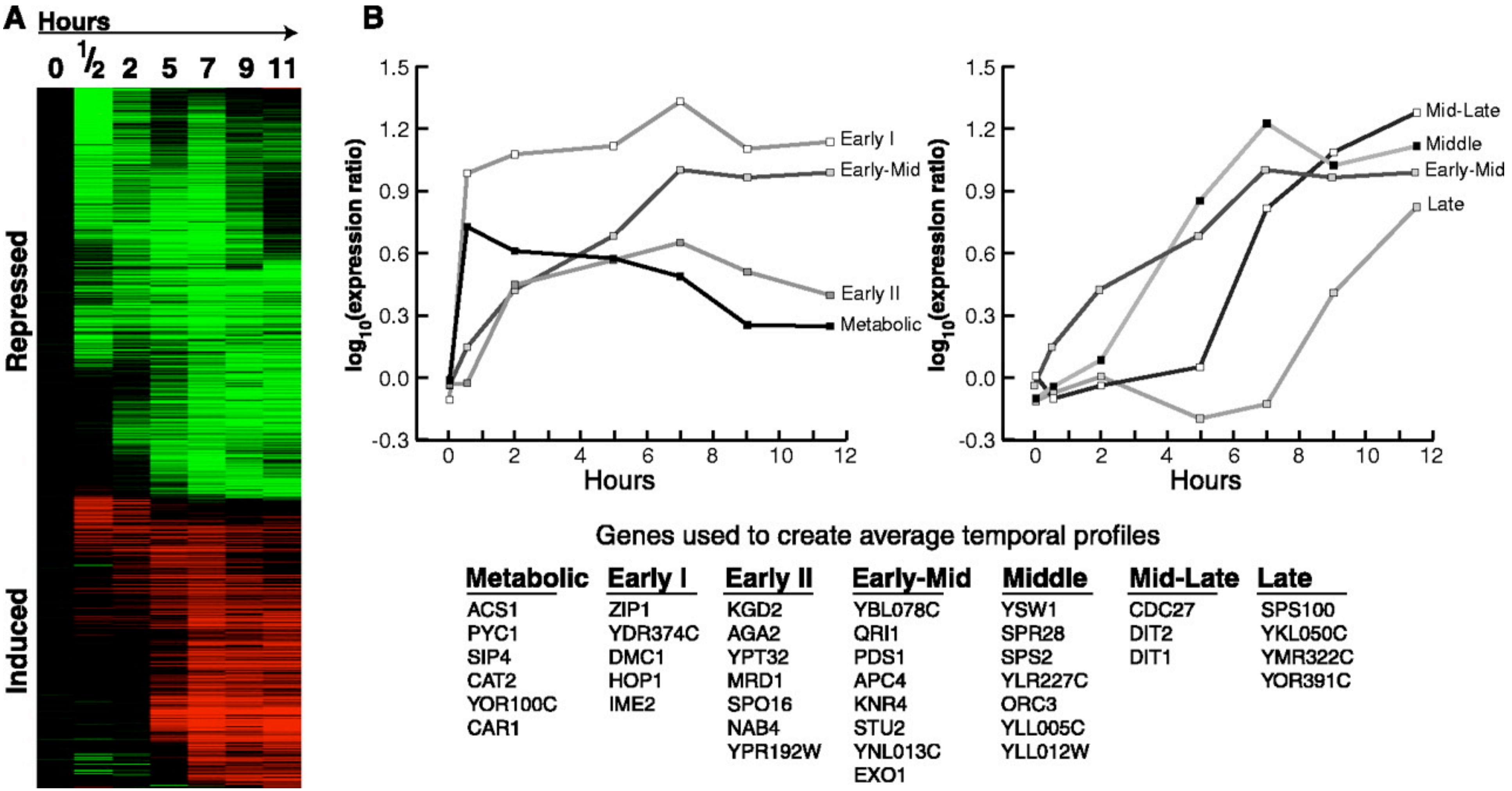


NB: < 20% spores, so data are *mixtures* of cell stages

Standard Test (Northern) vs Array

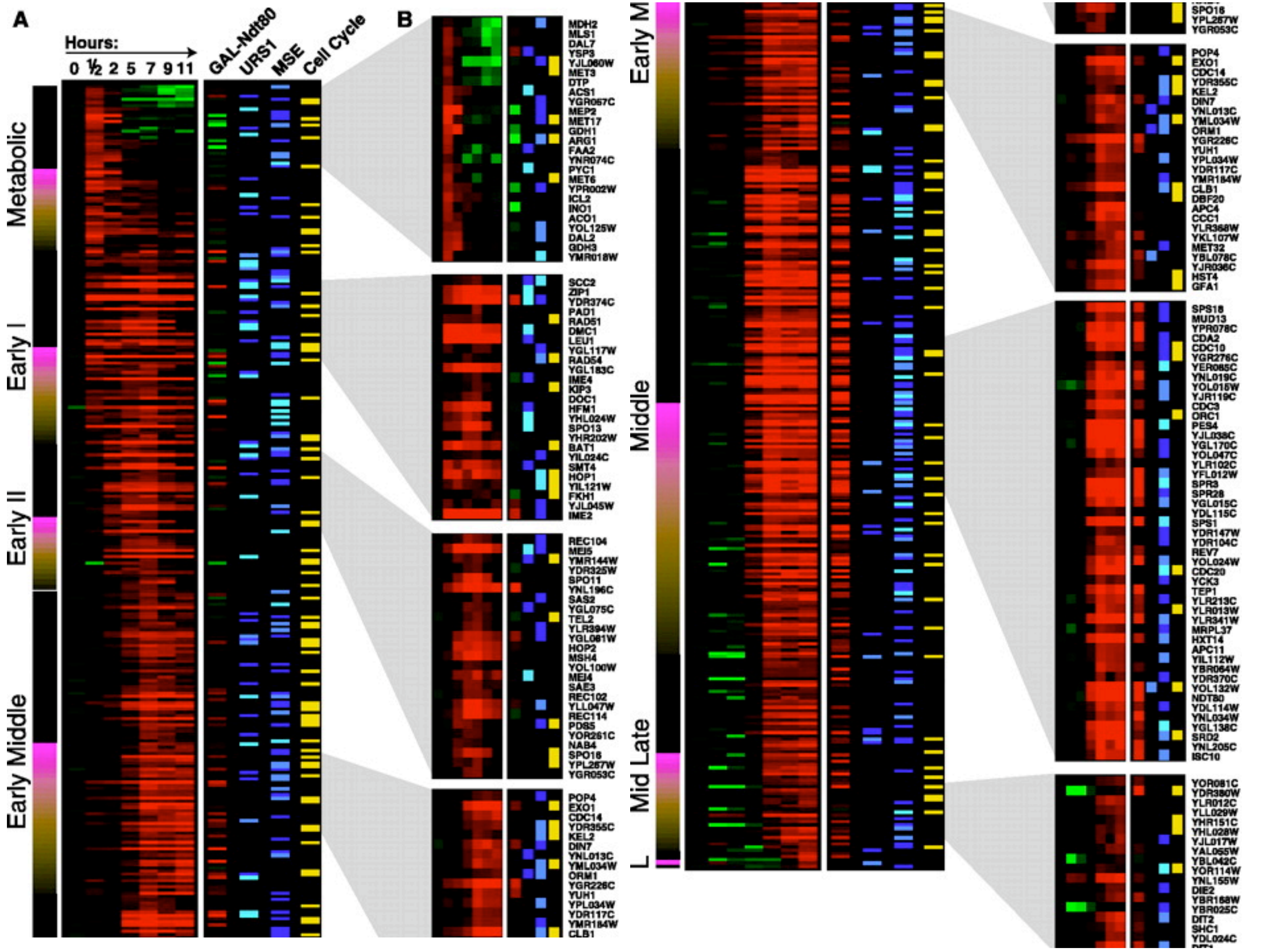


Prototype Expression Profiles



Genes used to create average temporal profiles

Metabolic	Early I	Early II	Early-Mid	Middle	Mid-Late	Late
ACS1	ZIP1	KGD2	YBL078C	YSW1	CDC27	SPS100
PYC1	YDR374C	AGA2	QRI1	SPR28	DIT2	YKL050C
SIP4	DMC1	YPT32	PDS1	SPS2	DIT1	YMR322C
CAT2	HOP1	MRD1	APC4	YLR227C		YOR391C
YOR100C	IME2	SPO16	KNR4	ORC3		
CAR1		NAB4	STU2	YLL005C		
		YPR192W	YNL013C	YLL012W		
			EXO1			


A
Hours:
0 ½ 2 5 7 9 11
GAL-Ndt80
URS1
MSE
Cell Cycle
Metabolic
Early I
Early II
Early Middle
B
MDH2
MLS1
DAL7
YSP3
YJL060W
MET3
DTP
ACS1
YGR057C
MEP2
MET17
GDH1
ARG1
FAA2
YNR074C
PYC1
MET6
YPR002W
ICL2
INO1
ACO1
YOL125W
DAL2
GDH3
YMR018W
SCC2
ZIP1
YDR374C
PAD1
RAD51
DMC1
LEU1
YGL117W
RAD54
YGL183C
IME4
KIP3
DOC1
HFM1
YHL024W
SPO13
YHR202W
BAT1
YIL024C
SMT4
HOP1
YIL121W
FKH1
YJL045W
IME2
REC104
MEI5
YMR144W
YDR325W
SPO11
YNL196C
SAS2
YGL075C
TEL2
YLR394W
YGL081W
HOP2
MSH4
YOL100W
MEI4
SAE3
REC102
YLL047W
REC114
PDS5
YOR261C
NAB4
SPO16
YPL267W
YGR053C
POP4
EXO1
CDC14
YDR355C
KEL2
DIN7
YNL013C
YML034W
ORM1
YGR226C
YUH1
YPL034W
YDR117C
YMR184W
CLB1
Early M
Middle
Mid Late
L
SPO16
YPL267W
YGR053C
POP4
EXO1
CDC14
YDR355C
KEL2
DIN7
YNL013C
YML034W
ORM1
YGR226C
YUH1
YPL034W
YDR117C
YMR184W
CLB1
DBF20
APC4
CCC1
YLR368W
YKL107W
MET32
YBL078C
YJR036C
HST4
GFA1
SPS18
MUD13
YPR078C
CDA2
CDC10
YGR276C
YER085C
YNL019C
YOL015W
YJR119C
CDC3
ORC1
PES4
YJL038C
YGL170C
YOL047C
YLR102C
YFL012W
SPR3
SPR28
YGL015C
YDL115C
SPS1
YDR147W
YDR104C
REV7
YOL024W
CDC20
YCK3
TEP1
YLR213C
YLR013W
YLR341W
MRPL37
HXT14
APC11
YIL112W
YBR064W
YDR370C
YOL132W
NDT80
YDL114W
YNL034W
YGL138C
SRD2
YNL205C
ISC10
YOR081C
YDR380W
YLR012C
YLL029W
YHR151C
YHL028W
YJL017W
YAL055W
YBL042C
YOR114W
YNL155W
DIE2
YBR168W
YBR025C
DIT2
SHC1
YDL024C

"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

"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

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, …

More on Computation

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

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

Critique

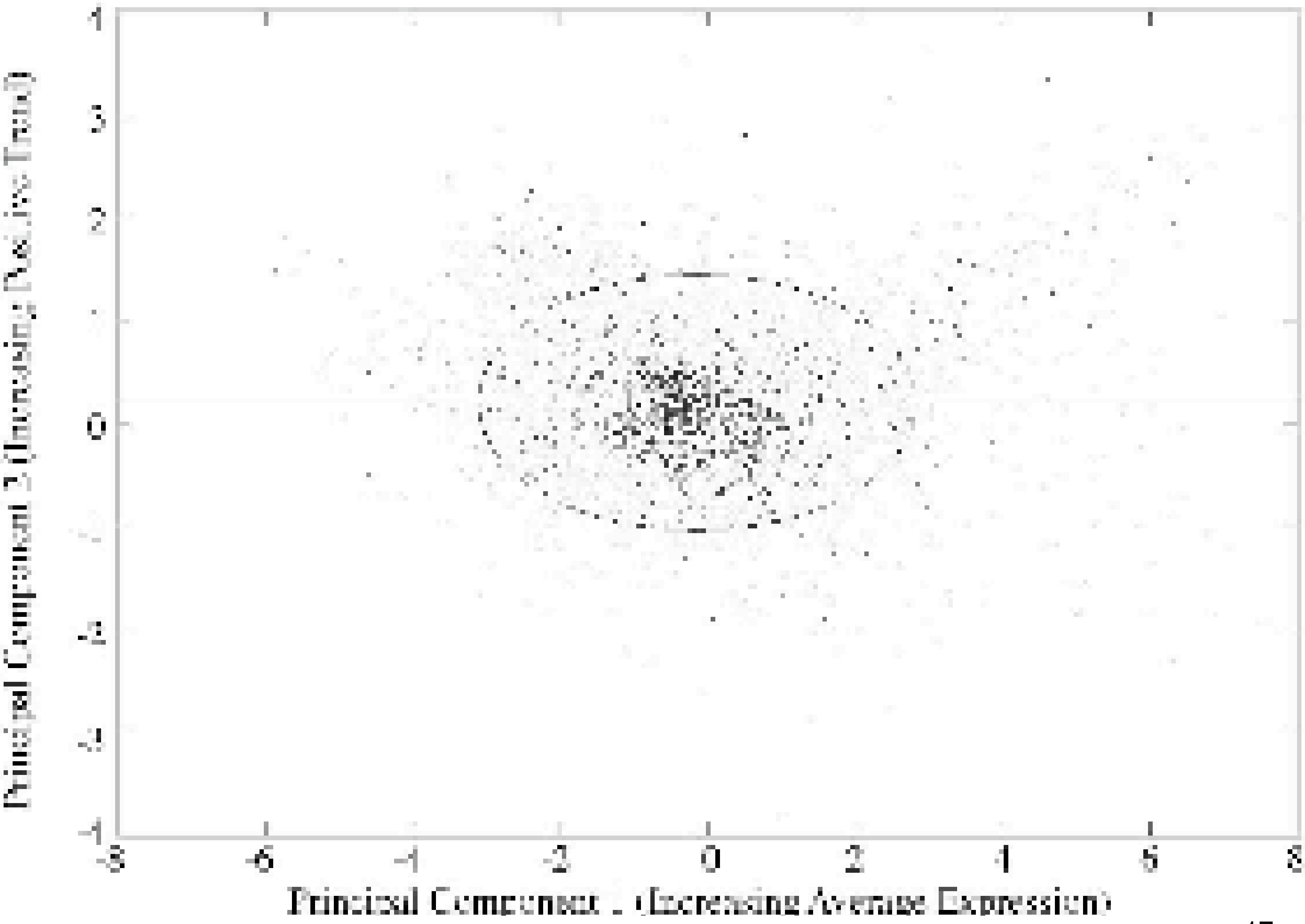
+

-

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.


4
3
2
1
0
-1
-2
-3
-4
-8
-6
-4
-2
0
2
4
6
8
Principal Component 1 (Increasing Average Expression)
Principal Component 2 (Increasing Positive Trend)

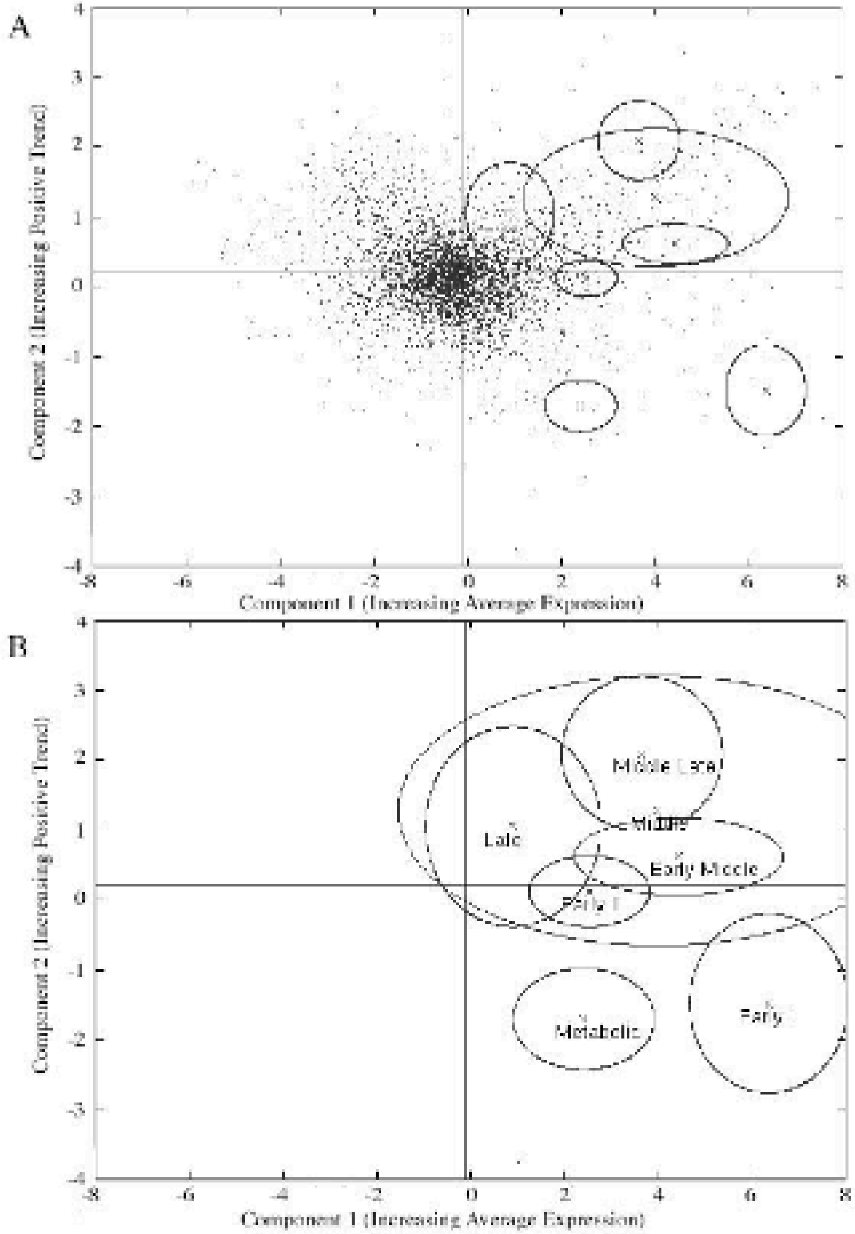
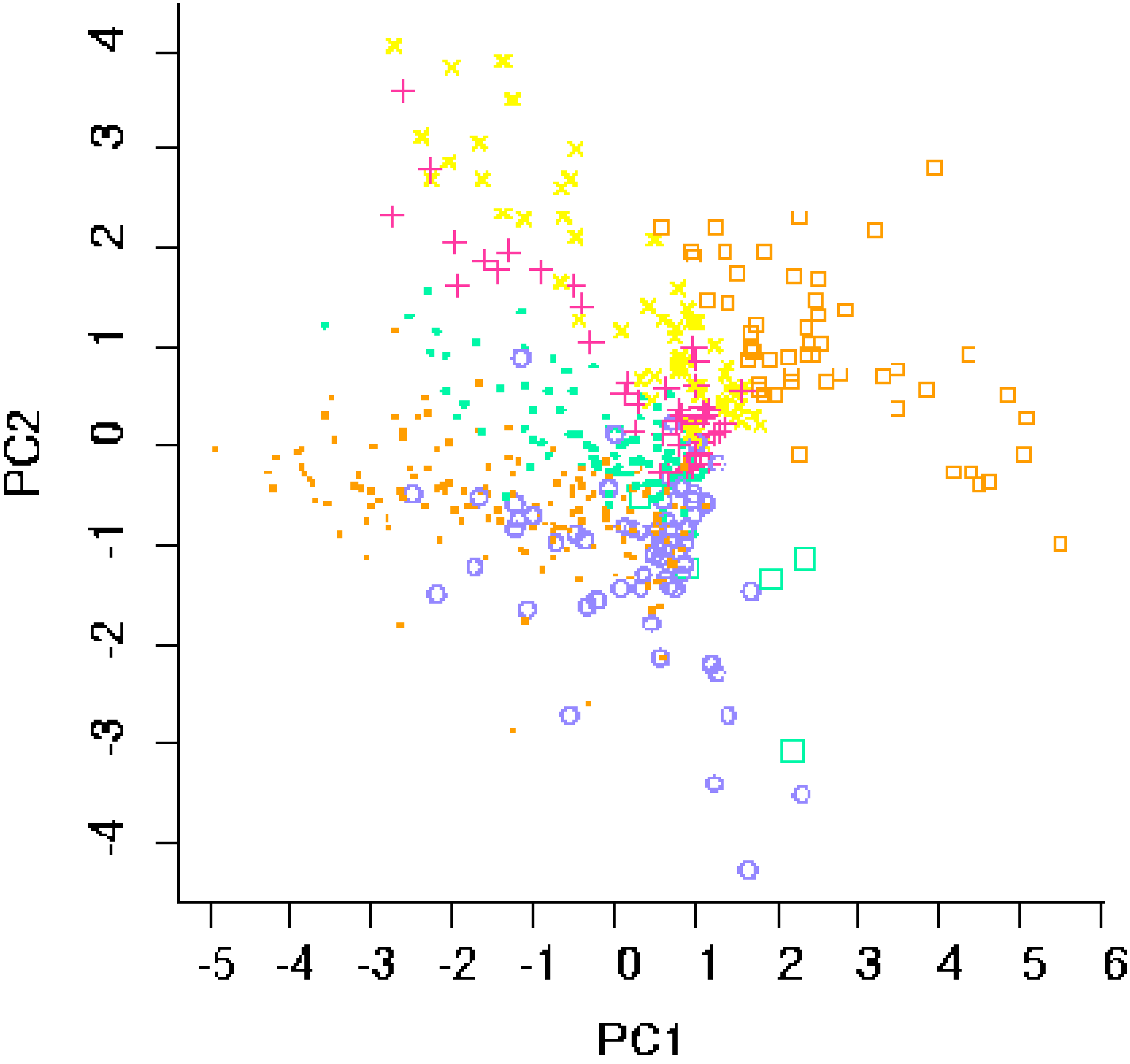
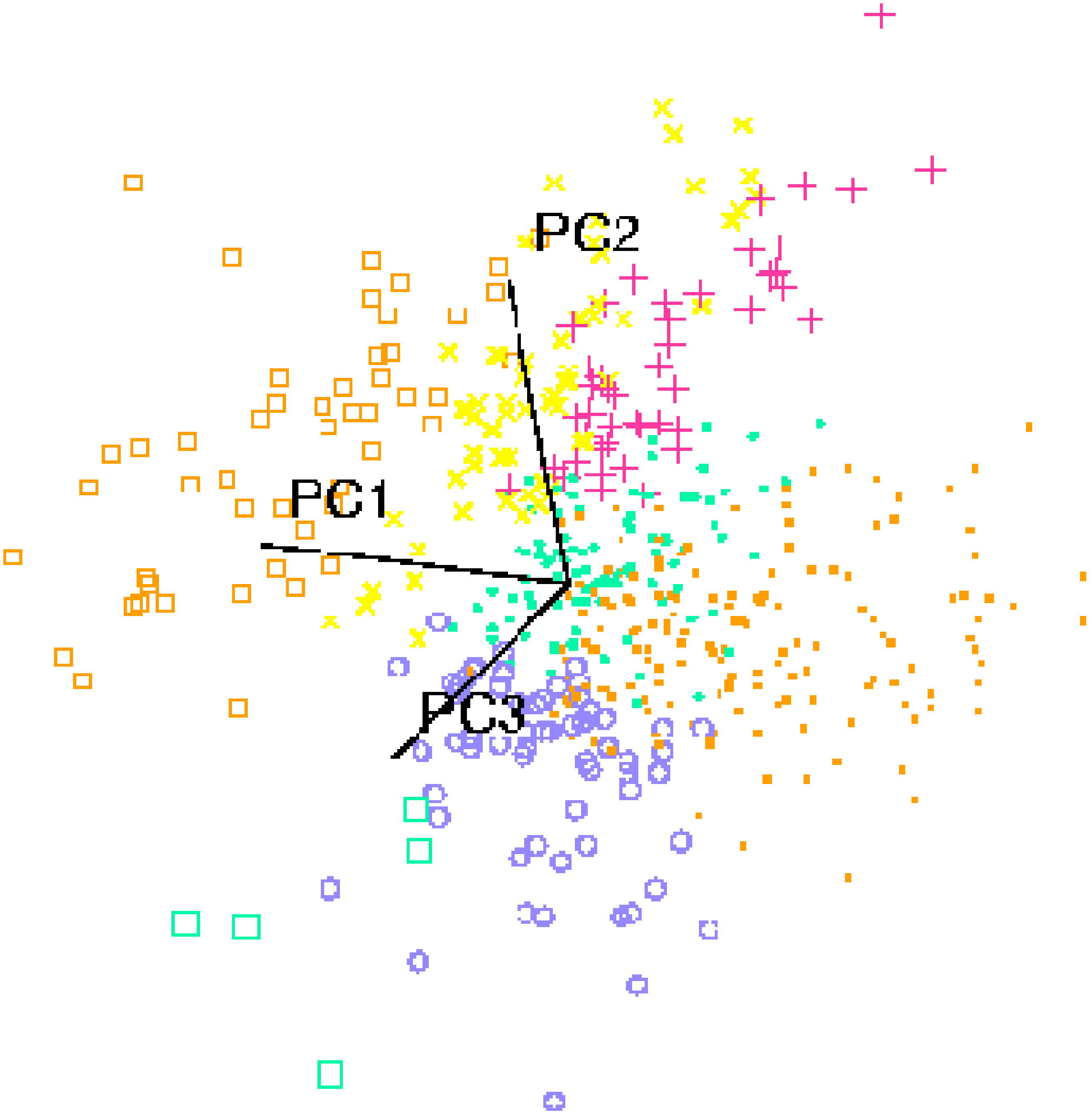


Figure 4. A. All genes plotted with respect to first and second principal components. Ellipses represent clusters identified in the original publication of the sporulation data. Ellipses are drawn to include 68% of the genes in the cluster. B. Ellipses are labelled using labels reported by the original investigators (Chu et al. 1998) and drawn to include 95% of genes in the cluster.


PC2
4
3
2
1
0
-1
-2
-3
-4
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
PC1


PC2
PC1
PC3