DNA Microarrays
Introduction Part II

Todd Lowe

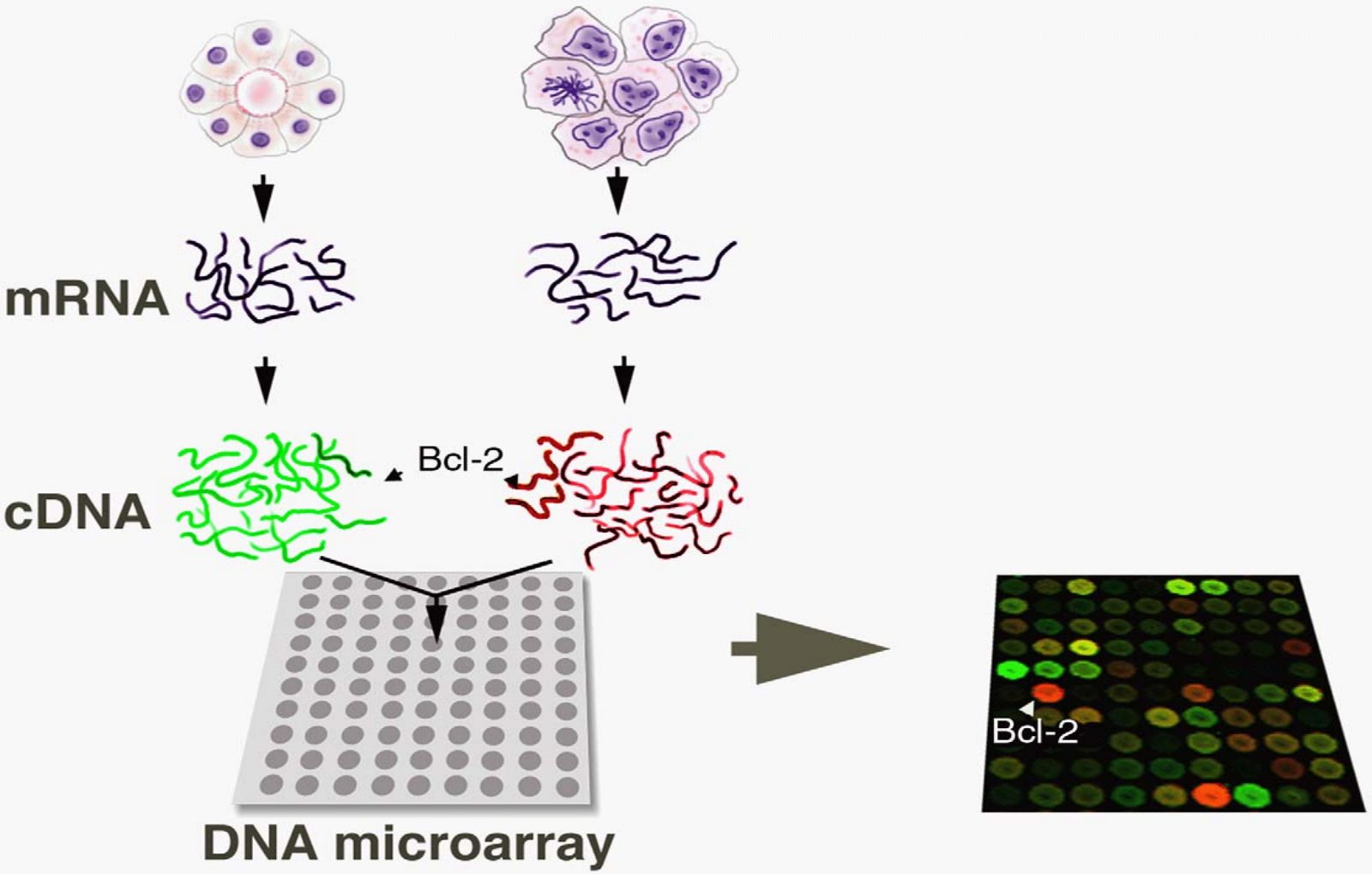
Bio 210

Jan 8, 2003

Readings Assigned

- Text: Chapters 3 + 5
- Be ready to discuss Spellman Cell Cycle paper next class

mRNA Expression Profiling



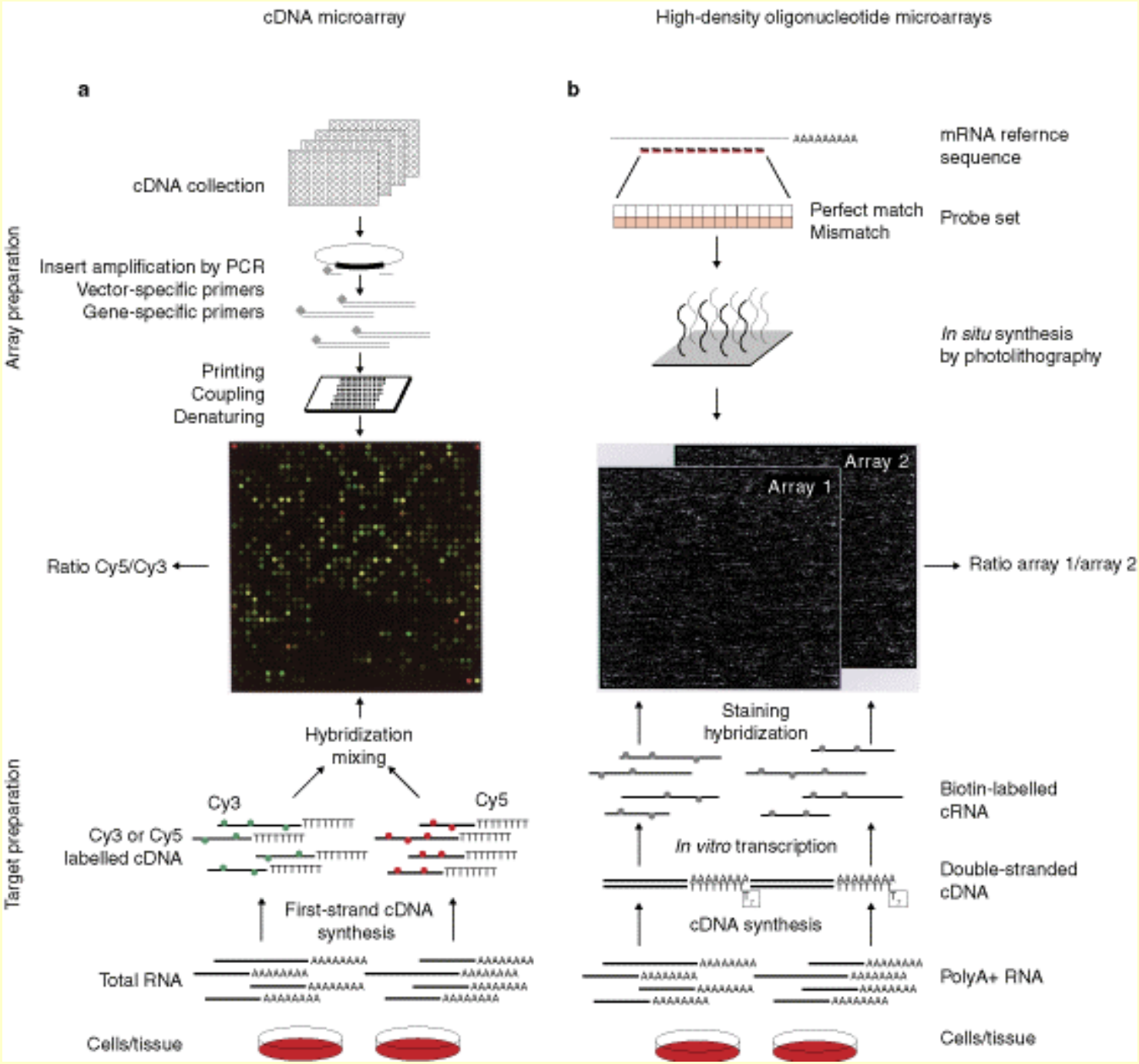
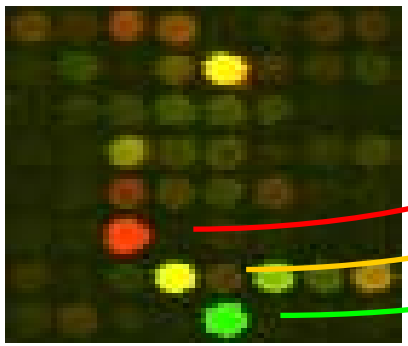
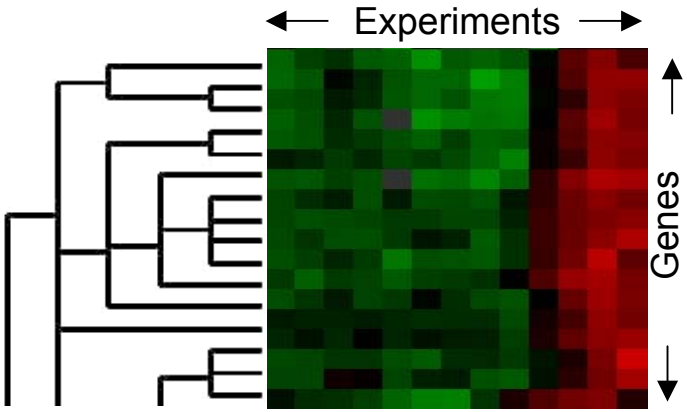

cDNA microarray
High-density oligonucleotide microarrays
a
b
Array preparation
cDNA collection
Insert amplification by PCR
Vector-specific primers
Gene-specific primers
Printing
Coupling
Denaturing
Ratio Cy5/Cy3
Hybridization
mixing
Cy3
Cy5
Cy3 or Cy5
labelled cDNA
First-strand cDNA
synthesis
Total RNA
Cells/tissue
Target preparation
mRNA refernce
sequence
Perfect match
Mismatch
Probe set
In situ synthesis
by photolithography
Array 2
Array 1
Ratio array 1/array 2
Staining
hybridization
Biotin-labelled
cRNA
In vitro transcription
Double-stranded
cDNA
cDNA synthesis
PolyA+ RNA
Cells/tissue

Image Analysis & Data Visualization

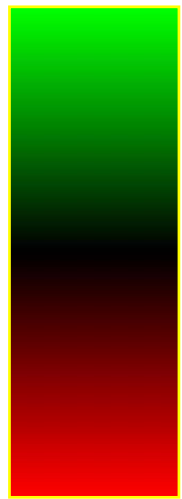


Cy3	Cy5	$\frac{Cy5}{Cy3}$	$\log_2\left(\frac{Cy5}{Cy3}\right)$
200	10000	50.00	5.64
4800	4800	1.00	0.00
9000	300	0.03	-4.91



Underexpressed

Overexpressed



8
4
2
fold
2
4
8

Genes involved in a specific biological process (i.e. heat shock)

- “Guilt by association” - assumption that genes with same pattern of changes in expression are involved the same pathway
- If we know what some genes in a group are doing, it hints that others with similar expression may be involved in same cellular process

Classification

- In some cases, don't need to know what any genes in a group are doing to be useful
- Total expression profile can be used as a "fingerprint" identifying cell type
  - example: Tumor classification - predict outcome / prescribe appropriate treatment based on clustering with "known outcome" tumors

Types of Analysis to Form Gene Associations

- Hierarchical clustering
- Partitioning of Data in Groups
  (supervised & unsupervised)
  - Self-organizing maps (SOM)
  - K-means clustering
  - Gene shaving
  - Support vector machines (SVM)
- Modeling
  - Singular value decomposition (SVD) / Principle component analysis (PCA)
- Ranking Tests

Data flow

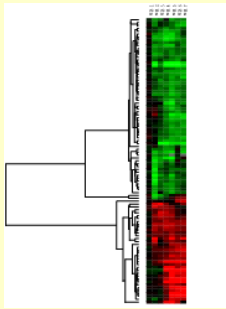
New Data → SA ScanAlyze

Database

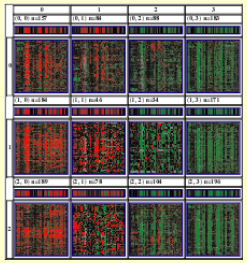
Data Selection

UID	NAME	GWEIGHT	spo0	spo30	spo2	spo5	spo7	spo9	spo11
EWEIGHT			1	1	1	1	1	1	1
YAL003W	EFB1	1	0.23	-1.79	-1.29	-1.56		-0.27	
YAL004W		1	0.41	-0.38	-0.89	-1.06	-1.6	-1.84	-1.6
YAL005C	SSA1	1	0.61	-0.07	-1.29	-1.29	-2	-1.84	-2.25
YAL010C	MDM10	1	0.16	-0.15	-0.76	-1.25	-1.89	-1.74	-1.6
YAL012W	CYS3	1	0.03	1.39	-0.84	-1.64	-2.84	-2.47	-2.4
YAL015C	NTG1	1	-0.18	-0.18	-0.62	-1.32	-1.69	-1.43	-1.79
YAL018C	YAL018C	1	-0.51	-0.62	-0.76	3.74	4.54	3.22	4.33
YAL025C	MAK16	1	-0.14	-3.32	-1.84	-1.12	-2.4	-1.03	-0.6
YAL034C	FUN19	1	0.19	-0.03	-1.03	-1.29	-1.84	-1.94	-1.74
YAL035W	FUN12	1	0.01	-1.47	-1.15	-0.69	-1.36	-1.64	-1.29
YAL036C	FUN11	1	-0.15	-2.74	-1.79	-1.32	-2.12	0.3	-0.89
YAL038W	CDC19	1	-0.06	-1.89	-1.69	-2.32	-2.4	-0.81	-1.6
YAL040C	CLN3	1	-0.17	-2.25	-1.69	-2.25	-2.56	-0.3	-2.4
YAL054C	ACS1	1	0.51	2.6	1.9	1.7	1.35	-0.03	-0.23
YAL055W	YAL055W	1	-0.32	0.83	0.58	0.82	1.4	2.05	2.24
YAL062W	GDH3	1	0.3	2.59	3	1.44	0.31	0.34	1.36
YAL067C	SEO1	1	-0.17	3.44	0.58	1.55	3.26	1.61	2.8
YAR003W	YAR003W	1	-0.29	0.54	0.6	1.08	1.42	1.86	1.42
YAR007C	RFA1	1	-0.14	1.74	2.41	2.1	2.04	0.57	0.84
YAR015W	ADE1	1	0.11	-1.51	-1.4	-1.36	-1.84	-1.89	-2
YAR027W	YAR027W	1	0.24	-1.06	-1.36	-1.56	-1.25	-0.94	-1.36
YBL009W	YBL009W	1	-0.01	0.62	1.04	1.3	2.52	2.15	2.24
YBL010C	YBL010C	1	0.01	0.21	0.7	1.45	2.25	1.77	1.24
YBL015W	ACH1	1	0.52	1.01	1.49	1.75	1.49	0.58	0.19

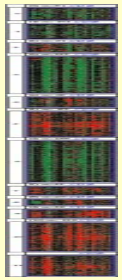
Complete Data Table (cdt)



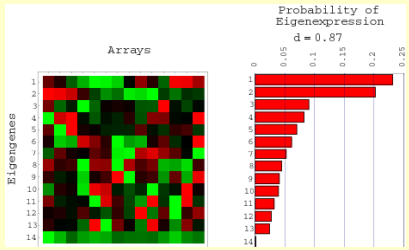
Cluster



SOM



K-means



SVD

Developing New Methods

- How do you know when your method performs better than a previous method?
- A “gold standard” test set for benchmarking array data doesn’t exist
- There is too much biology we don’t know: if a new method classifies a gene in the “wrong” gene group, is it recognizing new biology, or just getting it wrong??

Limitations of Arrays

- Do not necessarily reflect true levels of proteins - protein levels are regulated by translation initiation & degradation as well
- Generally, do not "prove" new biology - simply suggest genes involved in a process, a hypothesis that will require traditional experimental verification
- Expensive! $20-$100K to make your own / buy enough to get publishable data

Practical Problems

- Distinguishing background noise from low level, biologically important signals
- Difficult to independently verify so many data points, so success of array based on “anecodotal evidence”
- Multiple repetitions of array experiments is crucial, not always possible with limited samples
- Not easy to dissect specific from non-specific expression changes for a spec. experiment
  - Example: “common stress response” in yeast

Common Reference Problem

- Difficult to directly compare array experimental results from different labs / technologies
- For spotted arrays, all expression data are "relative" to a reference (ratios), not an absolute quantity
- Affy arrays are all absolute measurements, so must be converted to ratios for comparison to spotted arrays
- References should be consistent to facilitate comparisons, but vary a lot in practice
  - For human arrays, a pool of reference mRNA from 22+ cell lines is made to allow wide comparisons

Integrating Array Data with Sequence Analysis

One example: Promoter motif extraction

1. Cluster / classify genes with common response pattern
2. Align upstream promoter regions (Gibb's sampler) *or* count over-represented X-mers
3. Develop profile / motif from set & search genome for new candidates w/ motif
4. Return to array data, look for supporting evidence for new members
5. Carry out experiment to support hypothesis