



A Hidden Markov Model for Microarray Time Course Data

Christina Kendziorski and Ming Yuan

Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

Full references available at <http://www.biostat.wisc.edu/~kendzior>

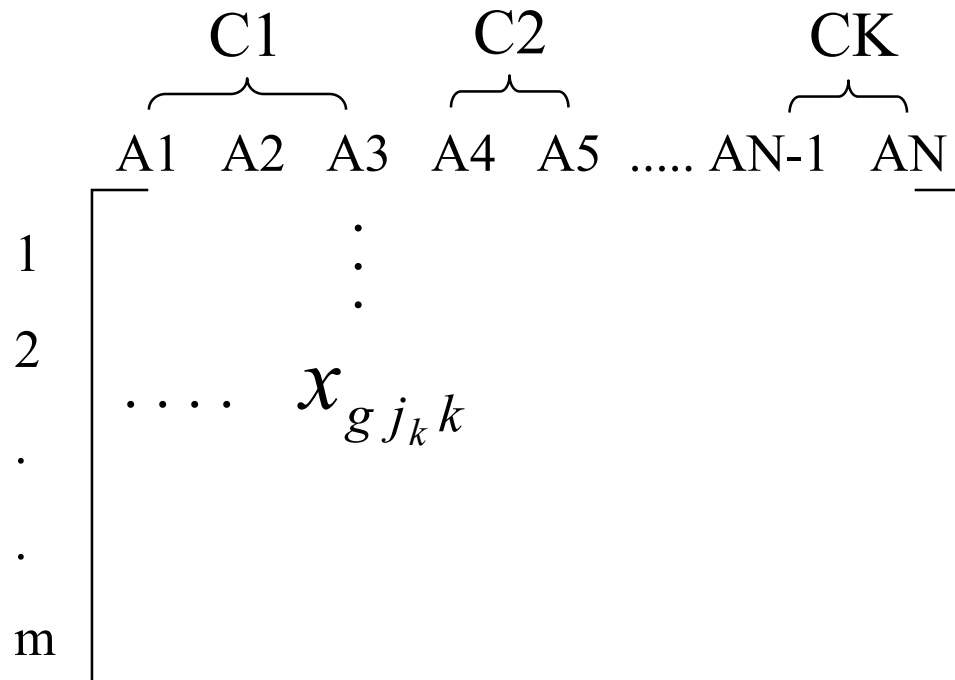


Case Studies

- Dioxin exposure in the liver (Bradfield lab, UW Madison)
Comparison of three dioxin doses in mice.
1, 2, 4, 8, 16, 32, and 64 days after treatment.
168 cDNA arrays (132 were used).
- Oxidative stress on the heart (Prolla lab, UW Madison)
Comparison of young and old mice.
Baseline and 1, 3, 5, and 7 hours after stress induction.
30 MG-U74A Affymetrix chips.
- Type II diabetes on the kidney (Park lab, LSU)
Comparison of lean and obese rats.
2 hours, 1 day, 3 days and 7 days after treatment.
18 Affymetrix Rat 230A chips.



Data Structure



- $X = [X_1, X_2, \dots, X_T]$
- $g = 1, 2, \dots, m$ genes; $k = 1, 2, \dots, K$ conditions; $j = 1, 2, \dots, j_k$ replicates
- $t = 1, 2, \dots, T$ time points.



Important Tasks

- Identify genes differentially expressed at each time.
- Identify a gene's expression pattern over time.
- Cluster genes within one biological condition
 - clustering, id of cyclic patterns, HMMs



Outline

- Motivation: demonstrated using one case study.
- Empirical Bayes approach for identifying DE genes at a single time point.
- Accounting for dependence via HMMs.
- Three case studies.



EBarrays at each time (Oxidative Stress on Heart)

	Baseline	1 hour	3 hours	5 hours	7 hours
DE	1005	701	499	380	637

1 Hour

		DE	EE
Baseline	DE	494	511
	EE	207	8831

3 Hours

		DE	EE
1 Hour	DE	378	323
	EE	121	9221

.....

7 Hours

		DE	EE
5 Hours	DE	278	102
	EE	359	9304

$$P(\text{DE}|\text{DE}) = 0.49$$

$$P(\text{DE}|\text{EE}) = 0.03$$

$$P(\text{DE}|\text{DE}) = 0.54$$

$$P(\text{DE}|\text{EE}) = 0.01$$

$$P(\text{DE}|\text{DE}) = 0.73$$

$$P(\text{DE}|\text{EE}) = 0.04$$



Accounting for time (Oxidative Stress on Heart)

	Baseline	1 hour	3 hours	5 hours	7 hours
DE-marg	1005	701	499	380	637
DE-HMM	1531	1029	972	944	959
Increase	52%	47%	95%	148%	51%
In common	100%	95%	97%	97%	91%

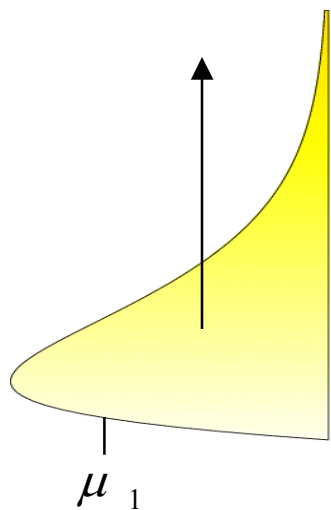


Empirical Bayes approach for identifying DE genes
at a single time point.

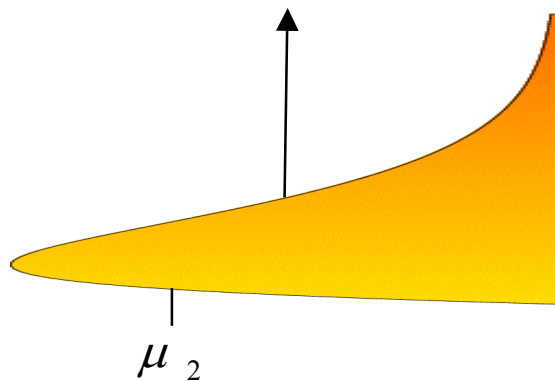


Hierarchical Model for Expression Data (One condition)

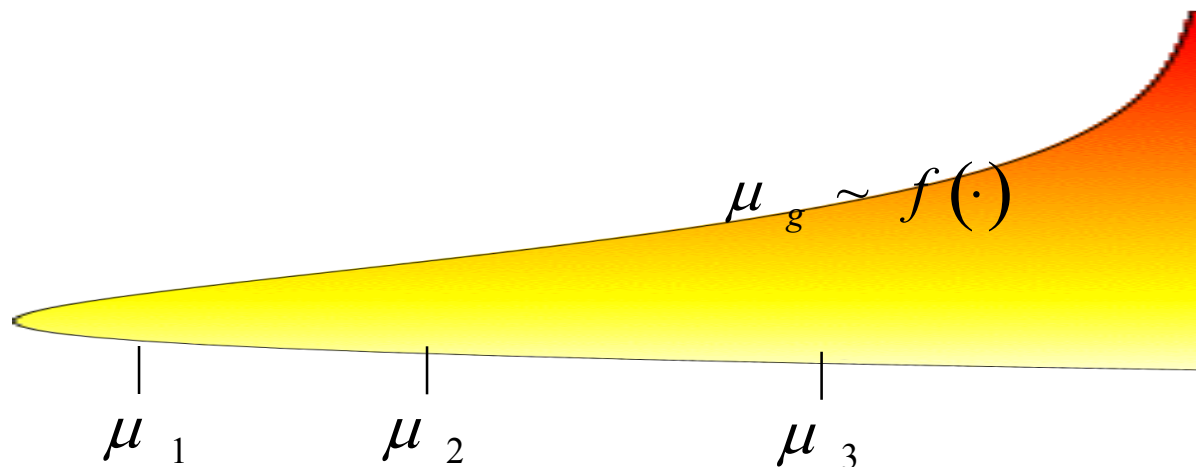
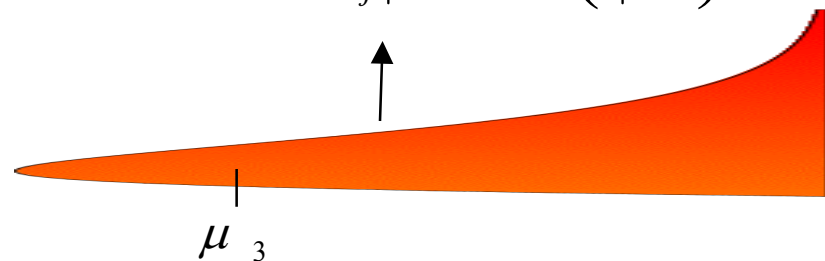
$$x_{1j} | \mu_1 \sim f(\cdot | \mu_1)$$



$$x_{2j} | \mu_2 \sim f(\cdot | \mu_2)$$



$$x_{3j} | \mu_3 \sim f(\cdot | \mu_3)$$



Hierarchical Model for Expression Data (Two conditions)

- Let $x = [x_{c1}, x_{c2}]$ denote data (one gene) in conditions C1 and C2.
- Two **patterns of expression**:

$$\text{P0 (EE)} : \mu_{c1} = \mu_{c2}$$

$$\text{P1 (DE)} : \mu_{c1} \neq \mu_{c2}$$

- For P0, $x \sim \int f(x|\mu)f(\mu)d\mu \equiv f_0(x)$

- For P1, $x \sim \int f(x|\mu_{c1}, \mu_{c2})f(\mu_{c1}, \mu_{c2})d\mu_{c1}d\mu_{c2}$

$$\equiv \underbrace{\int f(x_{c1}|\mu_{c1})f(\mu_{c1})d\mu_{c1}}_{f_0(x_{c1})} \underbrace{\int f(x_{c2}|\mu_{c2})f(\mu_{c2})d\mu_{c2}}_{f_0(x_{c2})} \equiv f_1(x)$$



Hierarchical Mixture Model for Expression Data

- Two conditions:

$$x \sim p_0 f_0(x) + p_1 f_1(x) \Rightarrow p(P1|x) = \frac{p_1 f(x|P1)}{p_0 f(x|P0) + p_1 f(x|P1)}$$

- Multiple conditions:

$$x \sim \sum_{k=1}^K p_k f_k(x) \Rightarrow p(Pk'|x) = \frac{p_{k'} f(x|Pk')}{\sum_{k \neq k'} p_k f(x|Pk)}$$

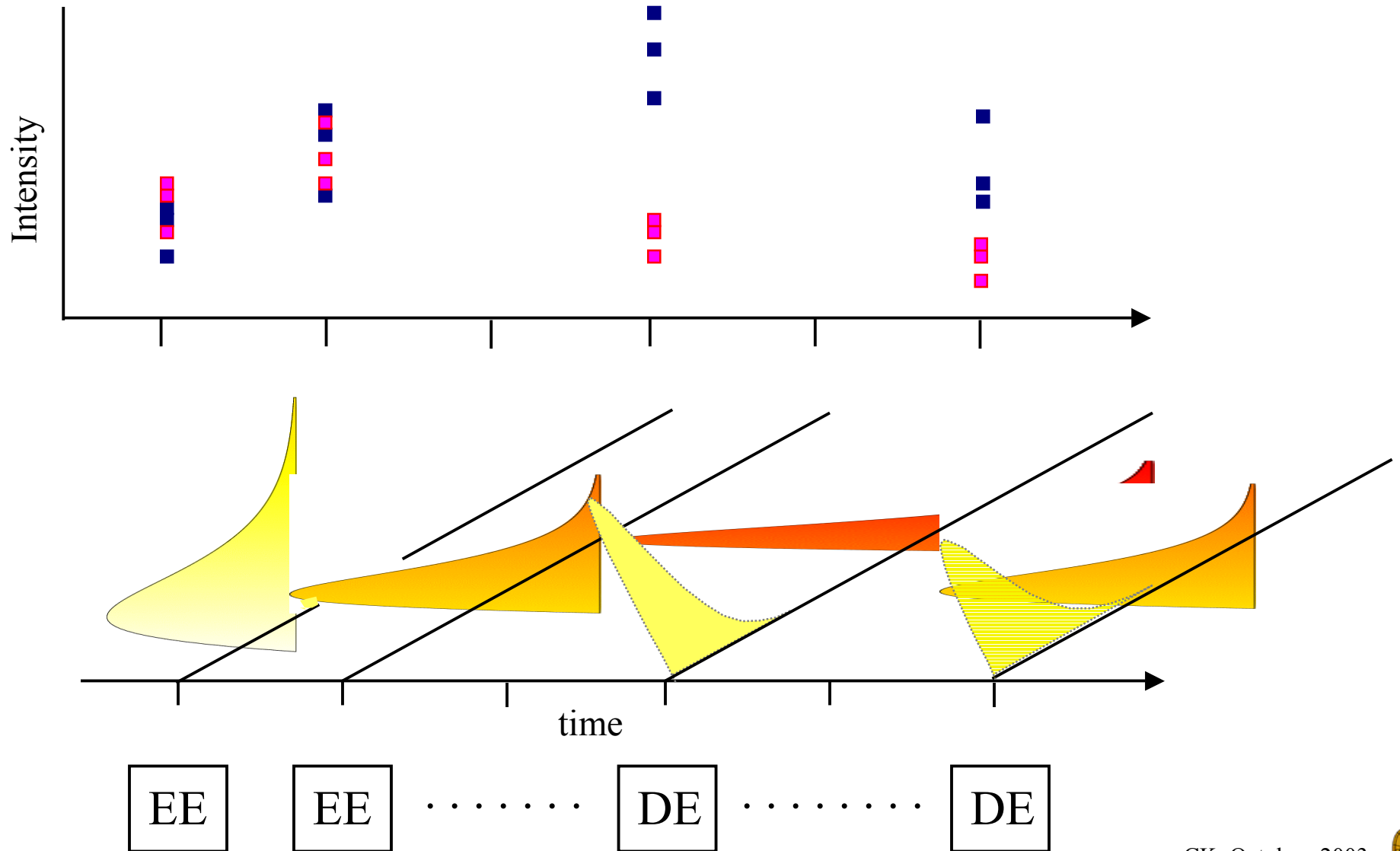
- Parameter estimates via EM
- Bayes rule determines threshold here; could target specific FDR.



Accounting for time dependence via HMMs



Hidden Markov Model (Two conditions)



Hidden Markov Model (Assumptions)

- Expression pattern processes is described by initial probability distribution and transition probability matrix $A(t)$.
 $A(t)$ might depend on time.

- Observed expression vector is characterized by:

$$x_t | s_t = k \sim f_{kt}(x_t)$$

- Temporal correlation in the data can be completely described by the pattern process:

$$x | s_1 = k_1, s_2 = k_2, \dots, s_T = k_T \sim \prod_{t=1}^T f_{k_t t}(x_t)$$



Hidden Markov Model: How does it help ?

■ Without HMM:
$$\frac{p(s_t = P1|x_t)}{p(s_t = P0|x_t)} = \frac{p(s_t = P1)f_{P1}(x_t)}{p(s_t = P0)f_{P0}(x_t)} > 1$$

■ With HMM:
$$\frac{p(s_t = P1|x_t, x_{-t})}{p(s_t = P0|x_t, x_{-t})} = \frac{p(s_t = P1|x_{-t})f_{P1}(x_t)}{p(s_t = P0|x_{-t})f_{P0}(x_t)} > 1$$



Does it work ?

One Simulation Study - and see Ming Yuan

Case Studies



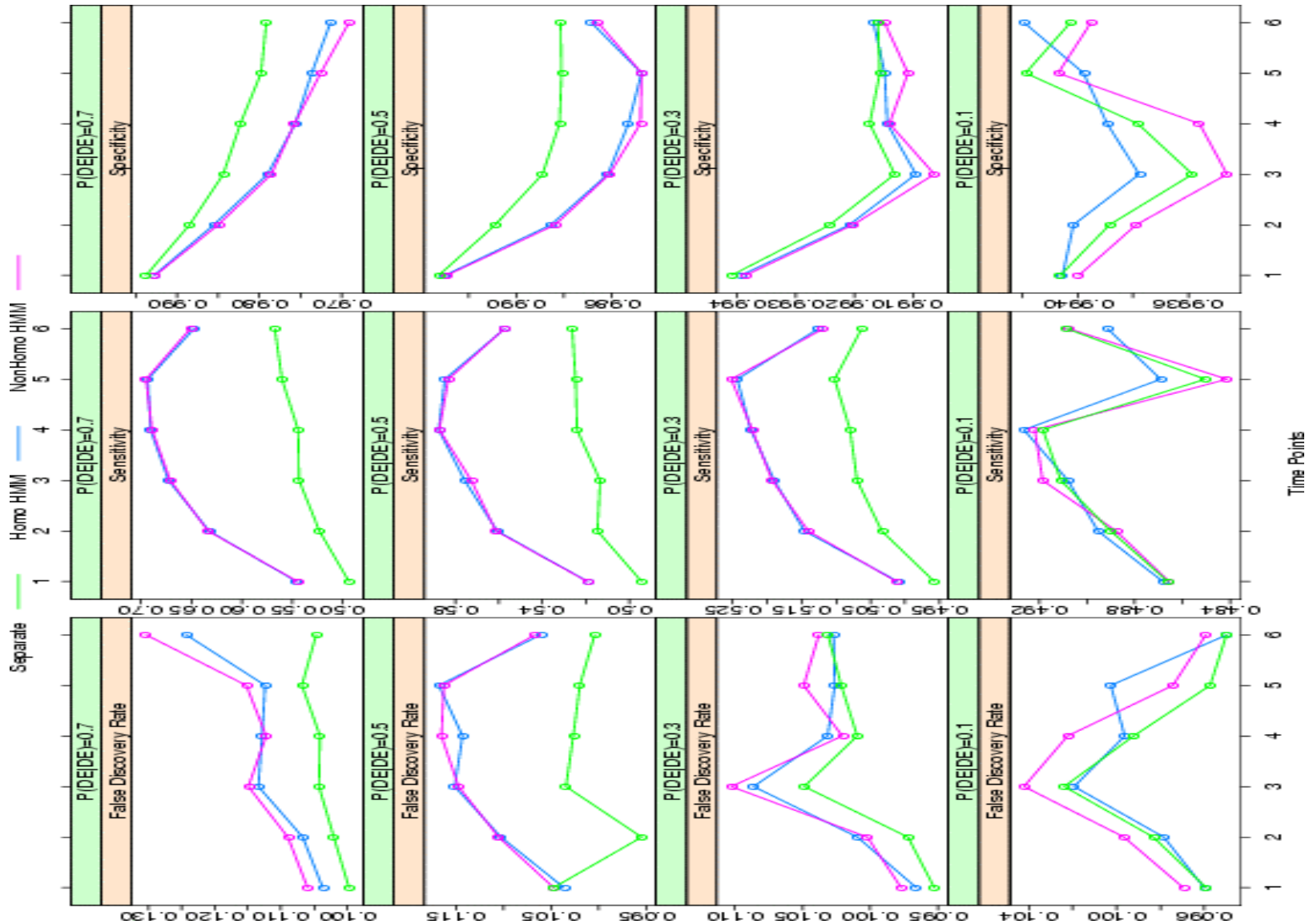
One Simulation Study

- HMM with 1500 genes, 2 conditions, 6 time points, no reps
- 10% DE at first time point. $P(\text{DE}|\text{EE})=0.1$; $P(\text{DE}|\text{DE})$ varies.
- Compare marginal, homogeneous and non-homogeneous HMM.

P(DE DE)	Method	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
0.1	I	81.65	82.33	85.52	82.39	80.01	82.41
	II	81.69	82.18	82.15	82.29	80.50	81.88
	III	81.75	82.43	82.79	82.74	80.04	82.50
0.3	I	82.15	100.78	106.01	105.49	106.30	105.40
	II	83.14	103.22	108.67	108.53	108.97	106.43
	III	83.29	103.24	109.23	108.61	109.73	106.59
0.5	I	84.05	120.72	134.05	142.61	144.40	145.32
	II	88.12	133.40	151.95	161.68	163.42	154.50
	III	88.27	133.80	151.42	162.28	163.07	154.93
0.7	I	82.58	140.76	178.48	197.66	216.63	225.55
	II	91.68	170.75	222.93	252.92	269.49	262.92
	III	91.75	171.75	223.06	252.39	271.41	266.83



Simulation Study Results



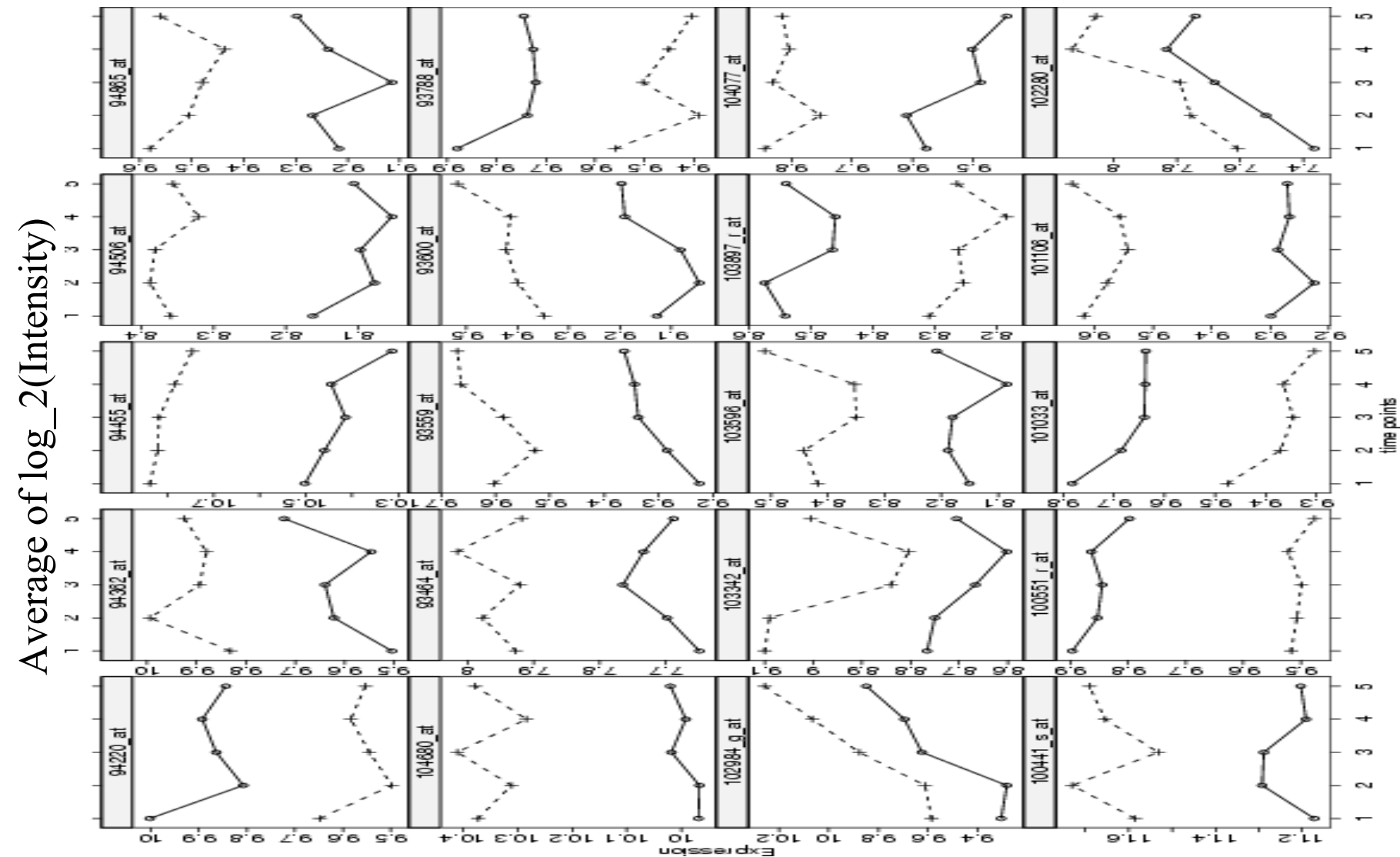
Oxidative Stress on Heart - Prolla Lab

	Baseline	1 hour	3 hours	5 hours	7 hours
DE-marg	1005	701	499	380	637
DE-HMM	1531	1029	972	944	959
Increase	52%	47%	95%	148%	51%

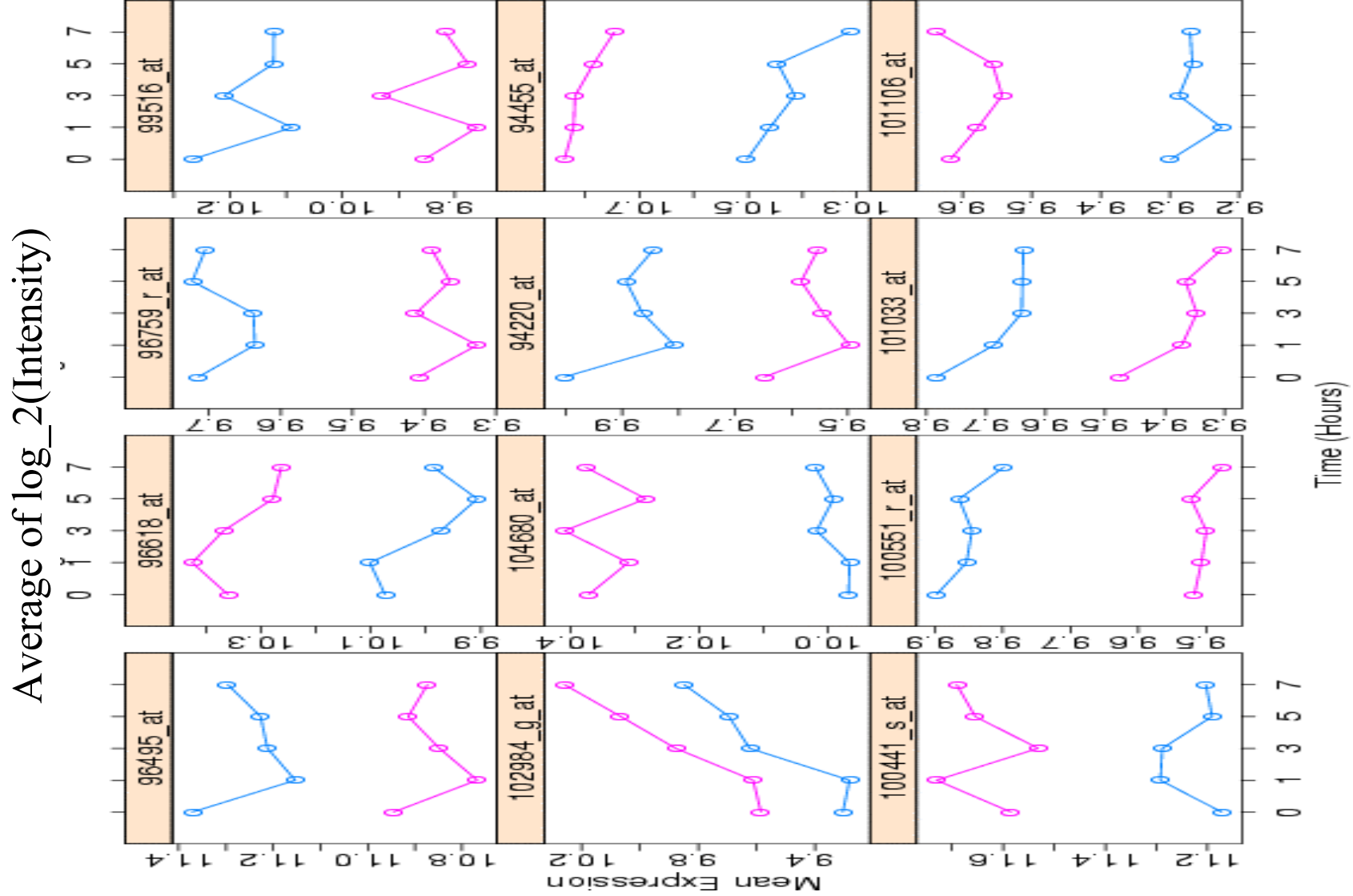
- 110 genes called EE at every time by marginal analysis, but DE at every time by HMM-EBarrays (HMME).



ALL EE by marginal analyses; All DE by HMME



ALL EE by marginal analyses; All DE by HMME



EBarrays at each time (Type 2 Diabetes on Kidney)

	2 Hours	1 day	3 days	7 days
DE	50	118	730	55

1 Day

DE EE

DE	10	40
EE	108	15765

3 Days

DE EE

DE	59	59
EE	671	15134

7 Days

DE EE

DE	37	693
EE	18	15175

$$P(\text{DE}|\text{DE}) = 0.2$$

$$P(\text{DE}|\text{EE}) = 0.007$$

$$P(\text{DE}|\text{DE}) = 0.5$$

$$P(\text{DE}|\text{EE}) = 0.042$$

$$P(\text{DE}|\text{DE}) = 0.051$$

$$P(\text{DE}|\text{EE}) = 0.001$$

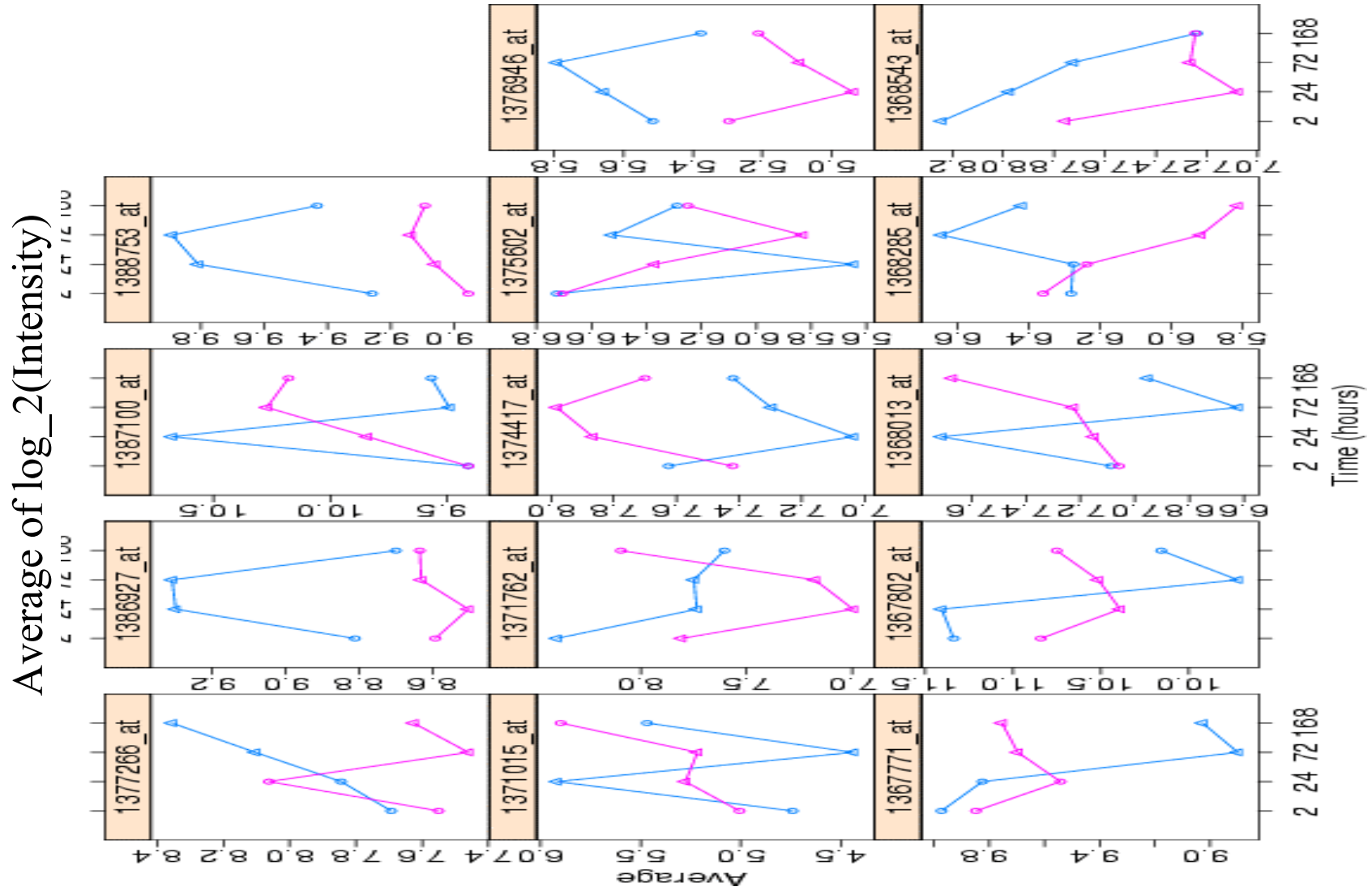


EBarrays vs. HMME (Type 2 Diabetes on Kidney)

	2 Hours	1 day	3 days	7 days
DE-marg	50	118	730	55
DE-HMM	72	218	717	112
Increase	44%	85%	-2%	104%
In common	98%	88%	88%	95%



EE by marginal analyses; DE by HMME (2x)



EBarrays vs. HMME (Dioxin on Liver)

Day	1	2	4	8	16	32	64
DE-marg	57	143	177	376	1081	169	111
DE-HMM	142	212	296	543	904	211	149
In common	77%	76%	68%	77%	75%	79%	77%



Summary

- Correlation in expression patterns over time exists.
- Most methods analyze time course data within condition.
- HMME approach identifies temporal expression patterns.
- HMME increases sensitivity.
- Pattern information is provided at each time point.
- RT-PCR results on the way.



Hierarchical Model for Expression Data

$$l(x_{1j}) | \mu_1 \sim f(\cdot | \mu_1)$$

$$l(x_{2j}) | \mu_2 \sim f(\cdot | \mu_2)$$

$$l(x_{3j}) | \mu_3 \sim f(\cdot | \mu_3)$$

