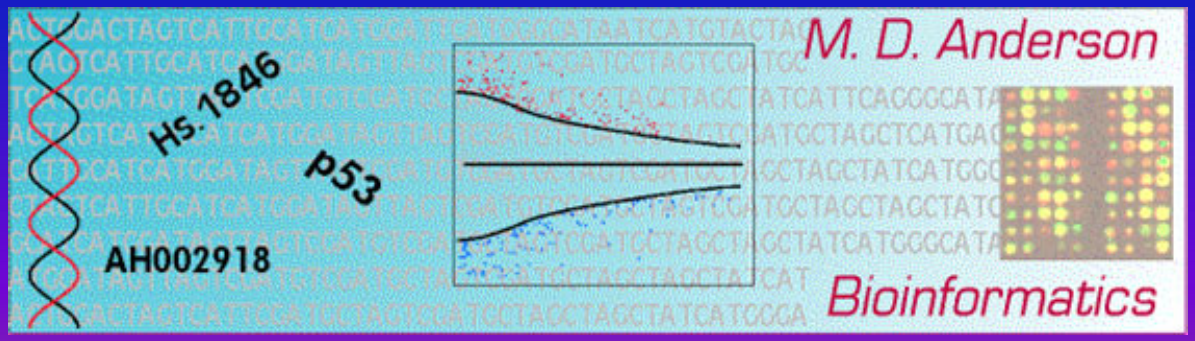
Organ-Specific Differences in Gene Expression and UniGene Annotations Describing Source Material

DN Stivers, J Wang,
GL Rosner, KR Coombes



Background Reference

Project normal: defining normal variance in mouse gene expression.

Pritchard, Hsu, Delrow and Nelson.

PNAS 98 (2001) 13266-13271.

Experimental Design

- Eighteen samples
  - Six C57BL6 male mice
  - Three organs: kidney, liver, testis
- Reference material
  - Pool all eighteen mouse organs
- Replicate microarray experiments using two-color fluorescence with common reference
  - Four experiments per mouse organ
  - Two red samples, two green samples

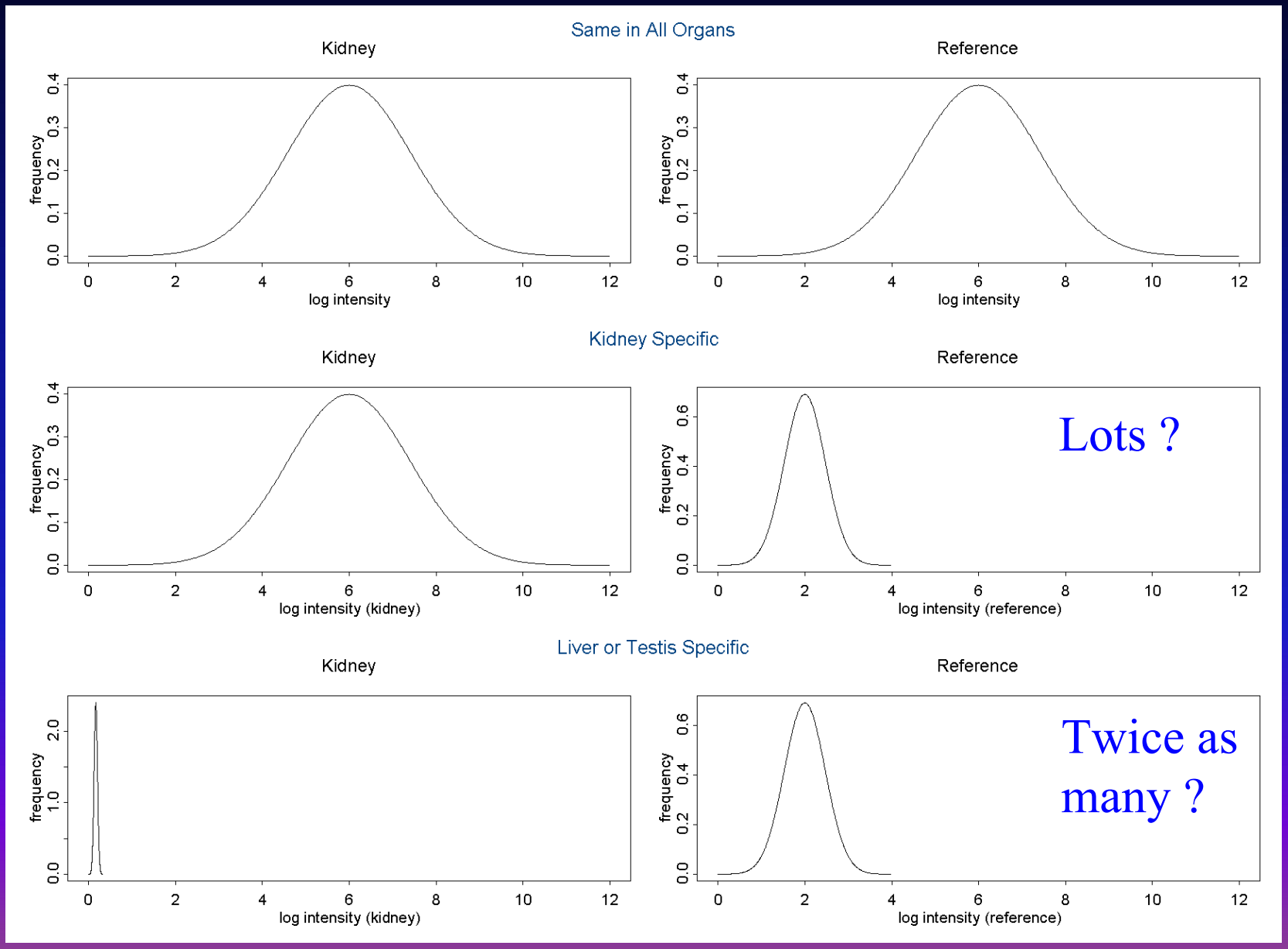
Their Analysis

- Print-tip specific intensity dependent loess normalization
- Scale adjusted (MAD)
- Use log ratios for further analysis
  - Log(experimental/reference)
- Perform F-test for each gene to see if mouse-to-mouse variance exceeds the array-to-array variance

Why Loess Normalization?

- Normalization methods assume:
  - Distributions of intensities are the same in the two channels
  - Most genes do not change expression
  - The number of overexpressed genes is about the same as the number of underexpressed genes
- Loess normalization tries to force the distributions in the two channels to match, believing that differences are attributable to technology.

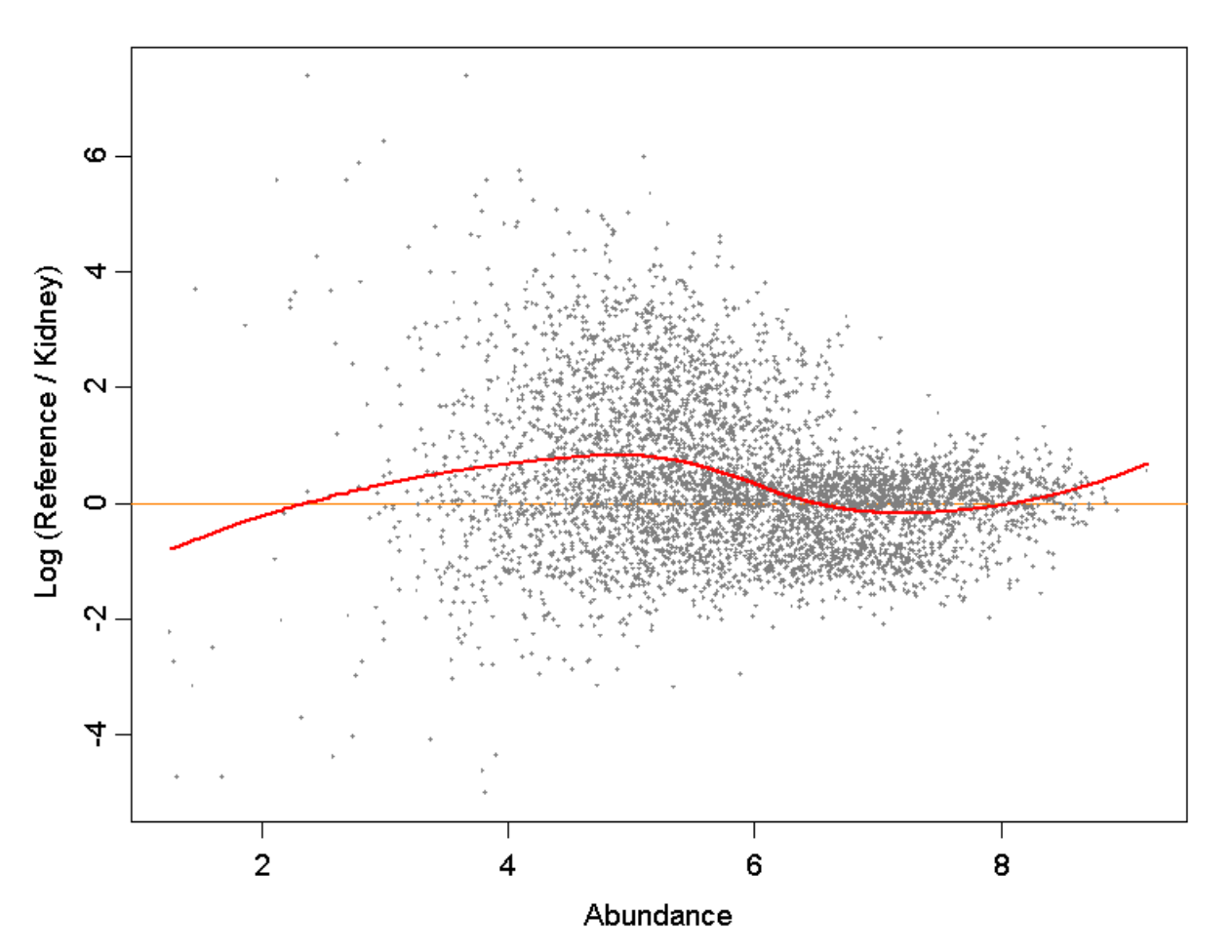
Theoretical Distribution



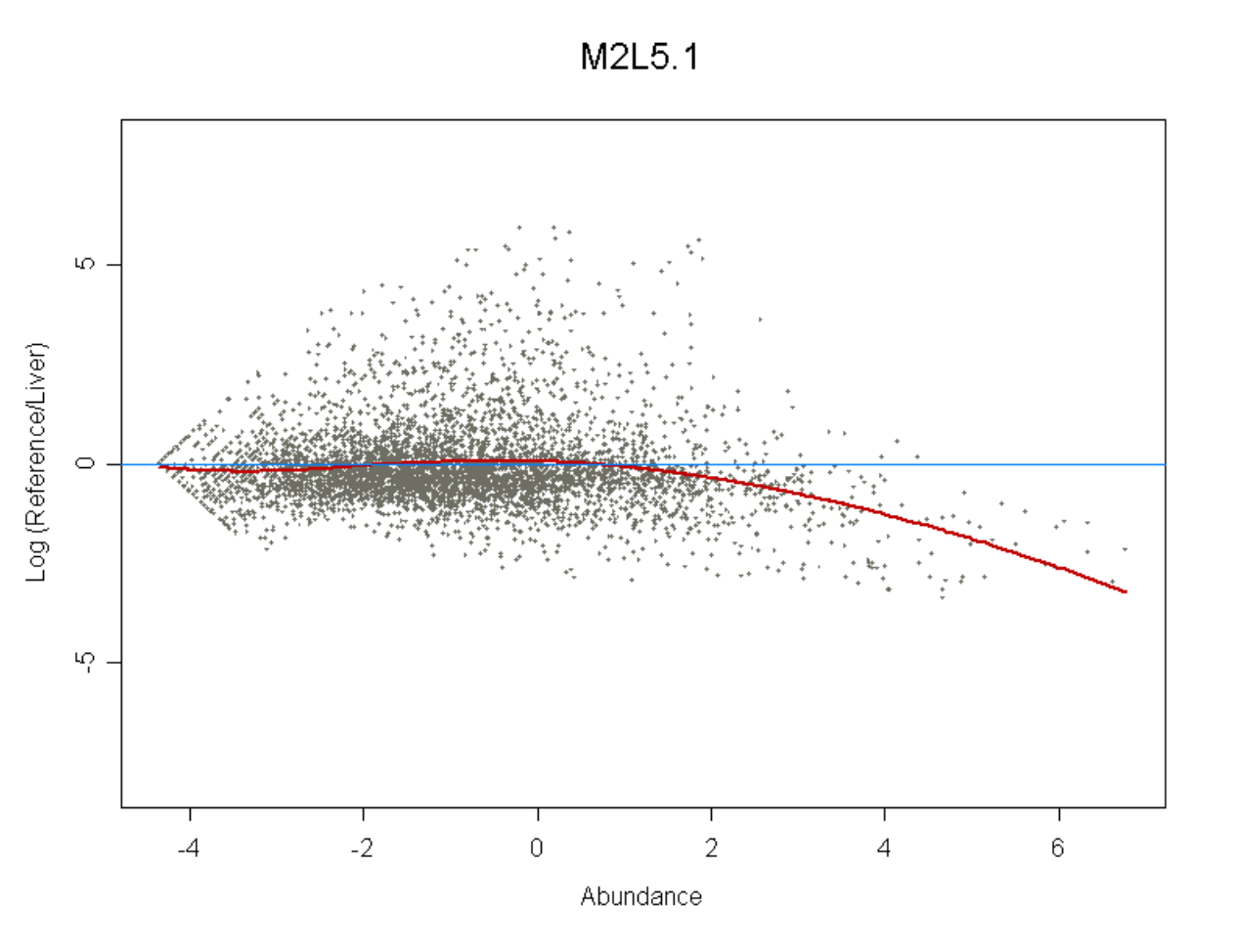
Our Data Processing: Keep It Simple

- Normalize channels separately
- Divide by 75th percentile
  - A magic number, but it avoids division by nominal zero
- Multiply by 10
  - A completely arbitrary number that has no effect on any of the analysis
- Set threshold at 0.5
  - More magic, chosen as five percent of the previous scaling factor
- Log transform

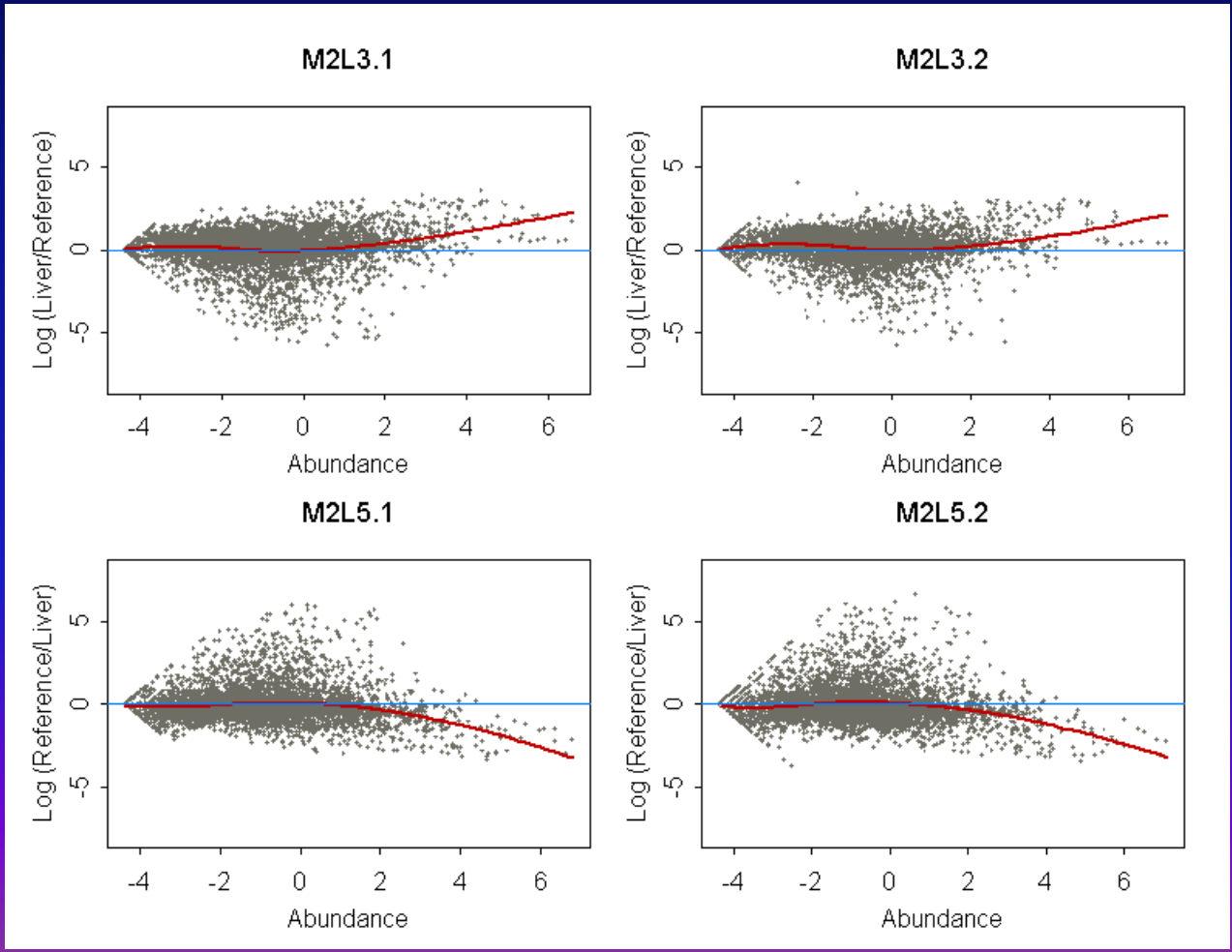
Comparison Between Channels Simulated From This Mixture



Real Data



Real Data


M2L3.1
Log (Liver/Reference)
5
0
-5
-4
-2
0
2
4
6
Abundance
M2L3.2
Log (Liver/Reference)
5
0
-5
-4
-2
0
2
4
6
Abundance
M2L5.1
Log (Reference/Liver)
5
0
-5
-4
-2
0
2
4
6
Abundance
M2L5.2
Log (Reference/Liver)
5
0
-5
-4
-2
0
2
4
6
Abundance

Interpretation

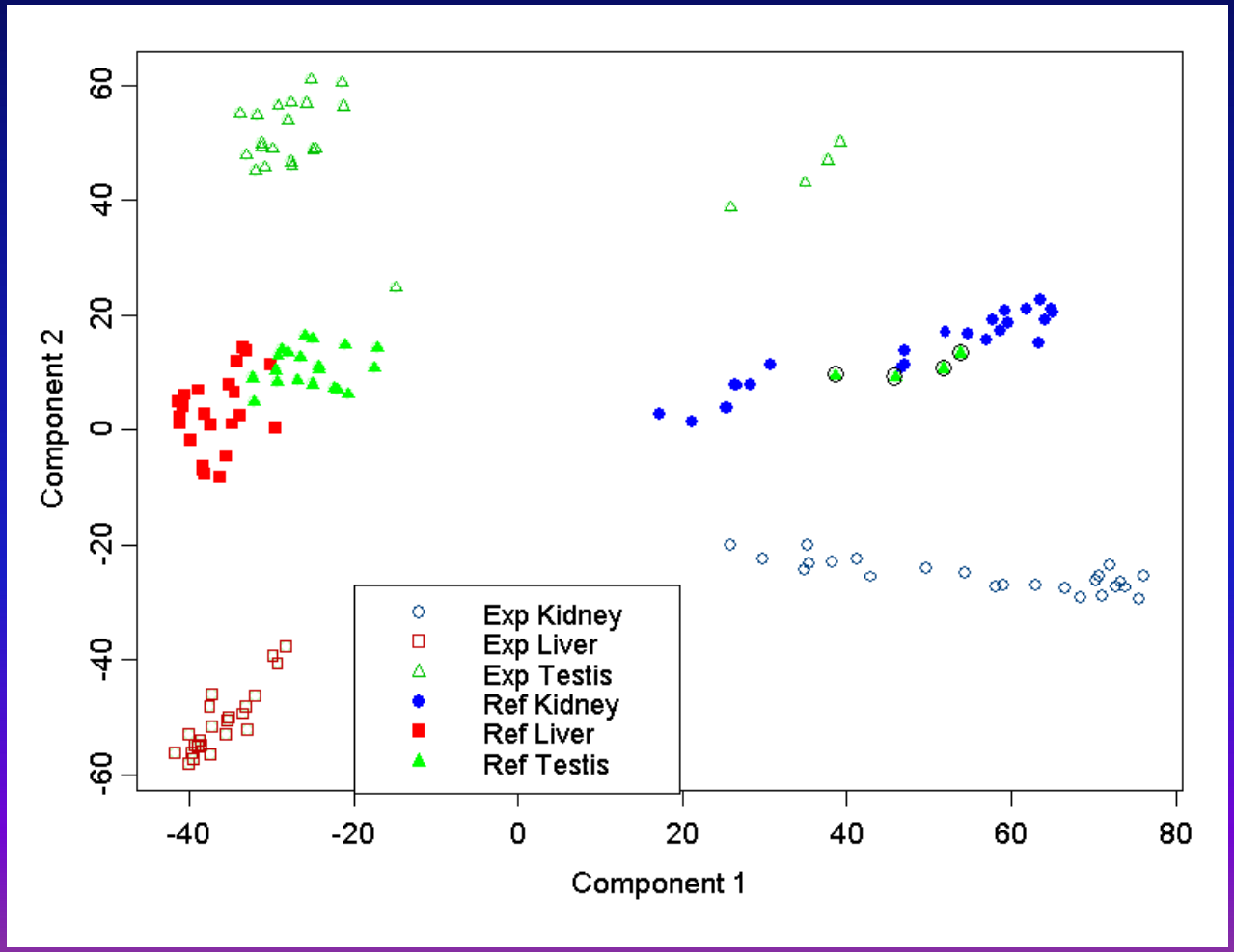
- Distributions of intensities are different in the two channels.
- Difference is NOT caused by arrays, dyes, or technology.
- Difference is inherent in the choice of reference material.

A Question

- Can we determine from this data set which genes are specifically expressed in each of the three organs?

- This question will become more important very soon…

Principal Components



When Bad Things Happen to Good Data

- Data was supplied in three files, one for each organ

- kidney.txt

```
Line# Unigene ID  Gene Name
589   Mm.4010     villin
```

- liver.txt

```
Line# Unigene ID  Gene Name
589   Mm.4010     villin
```

When Bad Things Happen to Good Data

- Data was supplied in three files, one for each organ

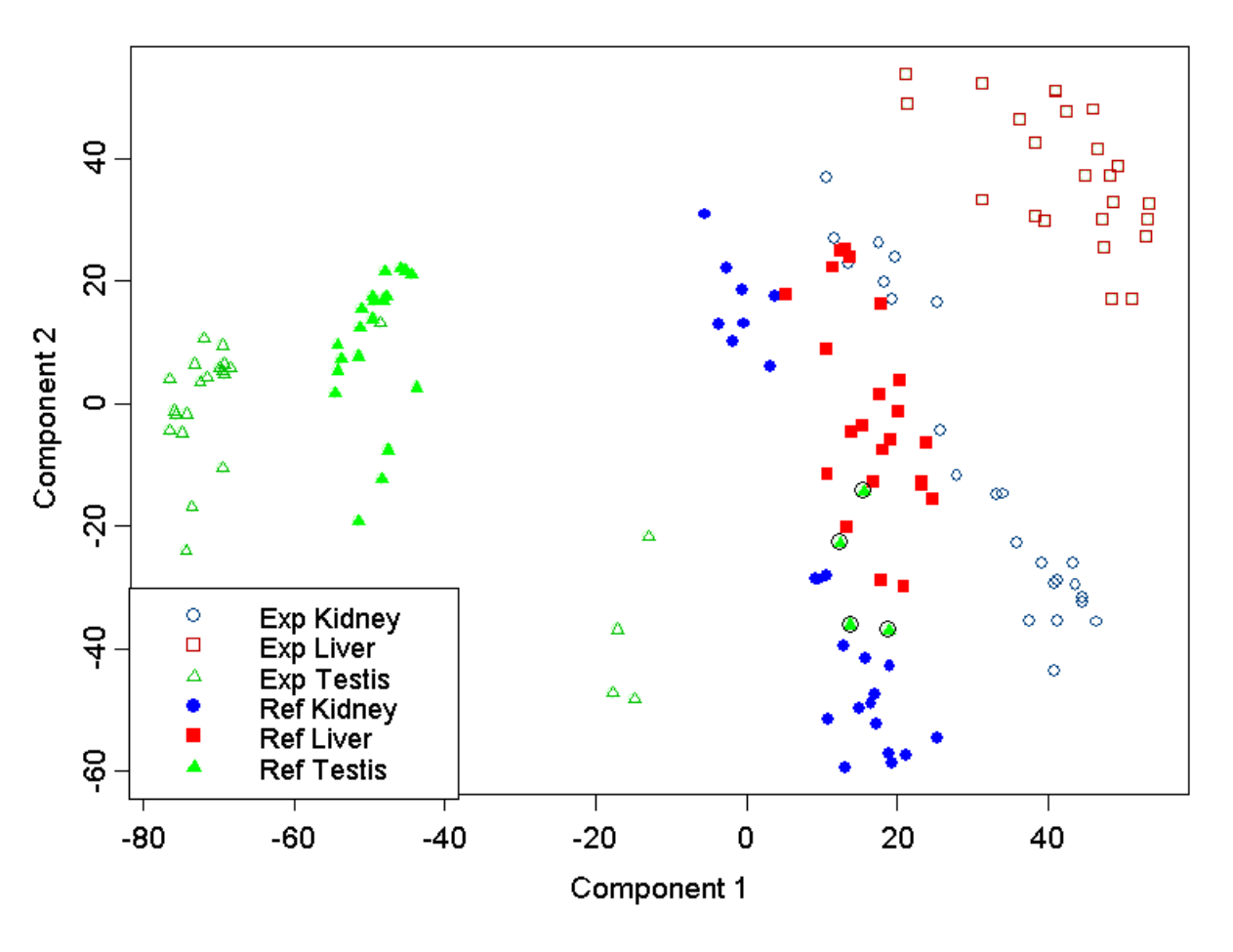
- kidney.txt

Line#	Unigene ID	Gene Name	Block	Column	Row
589	Mm.4010	villin	2	17	5

- liver.txt

Line#	Unigene ID	Gene Name	Block	Column	Row
589	Mm.4010	villin	4	17	5

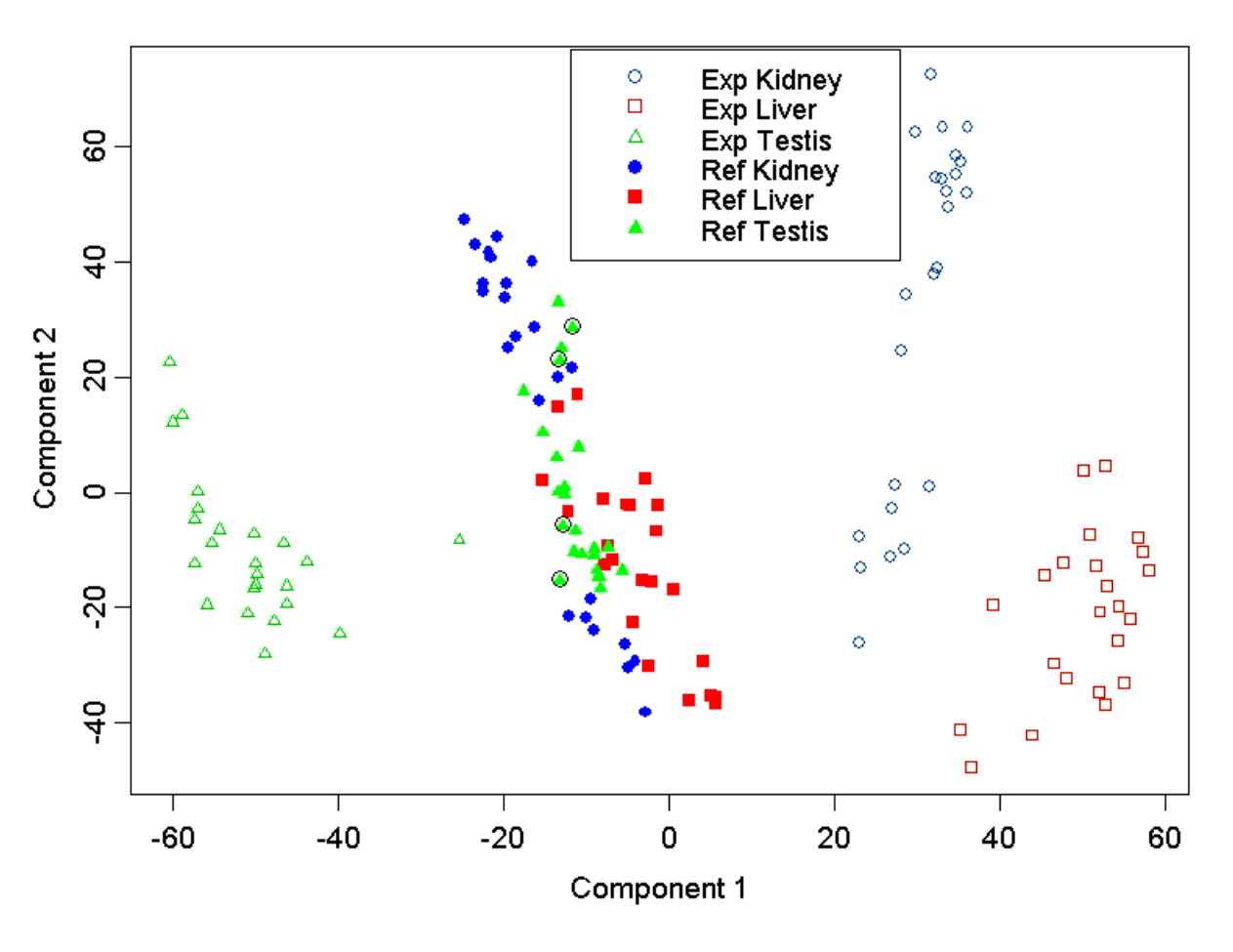
Principal Components (Take Two)



When Really Bad Things Happen to Good Data

- When the gene annotations match
  - Liver ref is close to 20 testis ref
  - Kidney ref is close to 4 testis ref
- When location annotations match
  - Kidney, liver and 4 testis ref are close
  - Other 20 testis ref are far away
- Conclusion: a data processing error occurred partway through the testis experiments

Principal Components (Take Three)



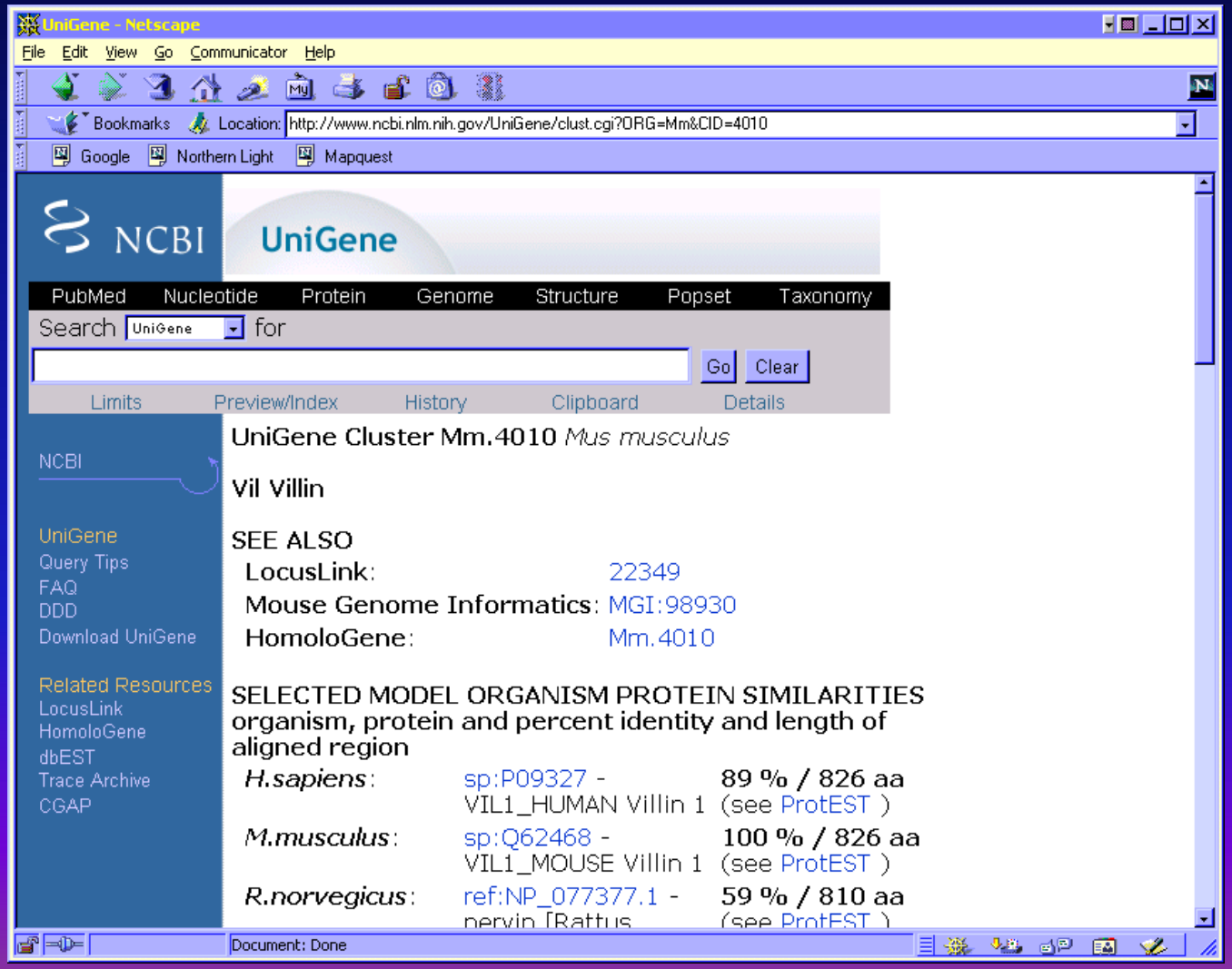
Every Solution Creates a New Problem

- Solution: After reordering all liver experiments and twenty testis experiments by location
  - Can distinguish the three organs
  - Reference samples cluster together
- New Problem: There are now two competing ways to map locations to genetic annotations (one from kidney.txt, one from liver.txt). Which is correct?

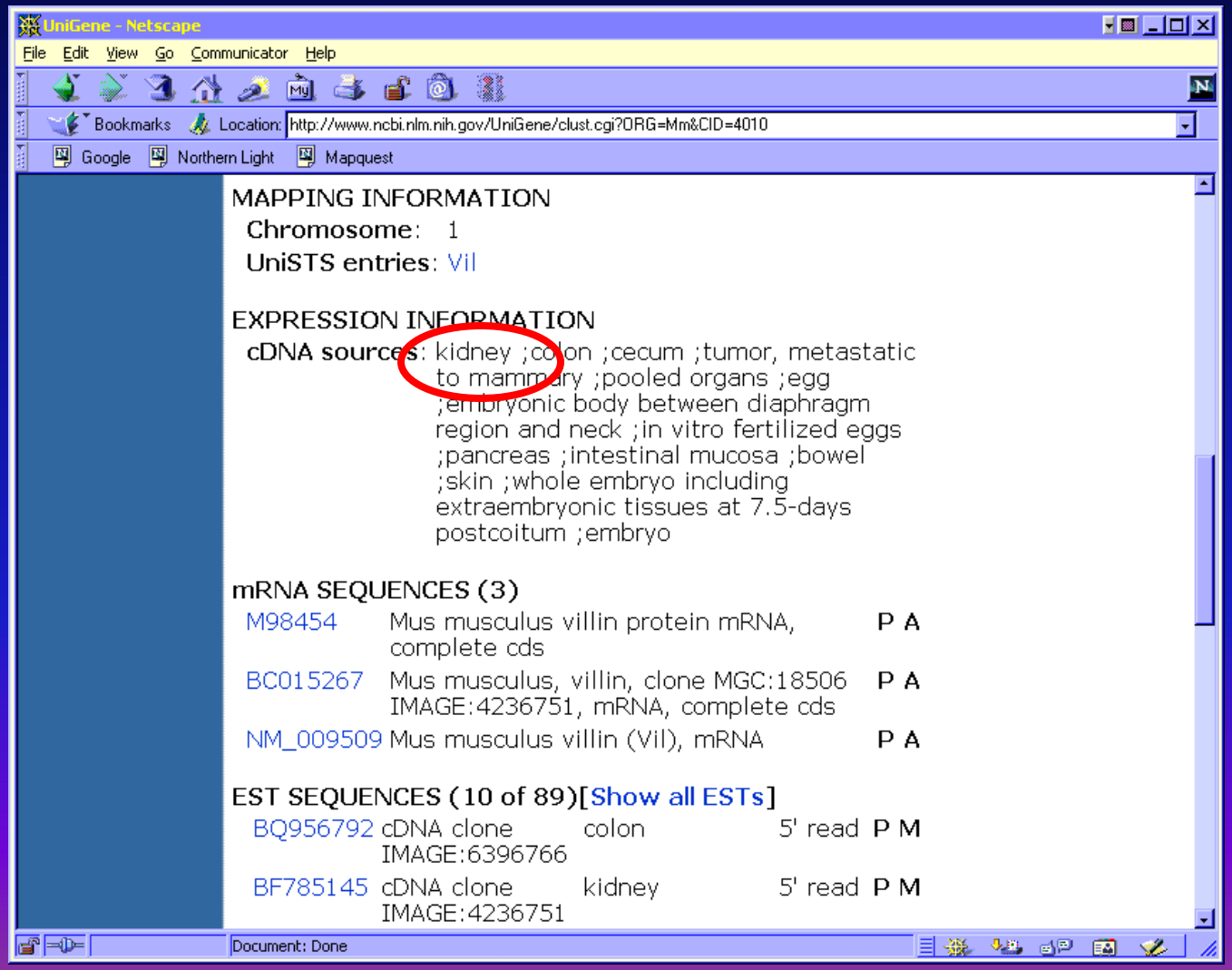
How Big is the Problem?

- Microarray contains 5304 spots
- Only 3372 (63.6%) spots have UniGene annotations that are consistent across the files
- So, 1932 (36.4%) spots have ambiguous UniGene annotations

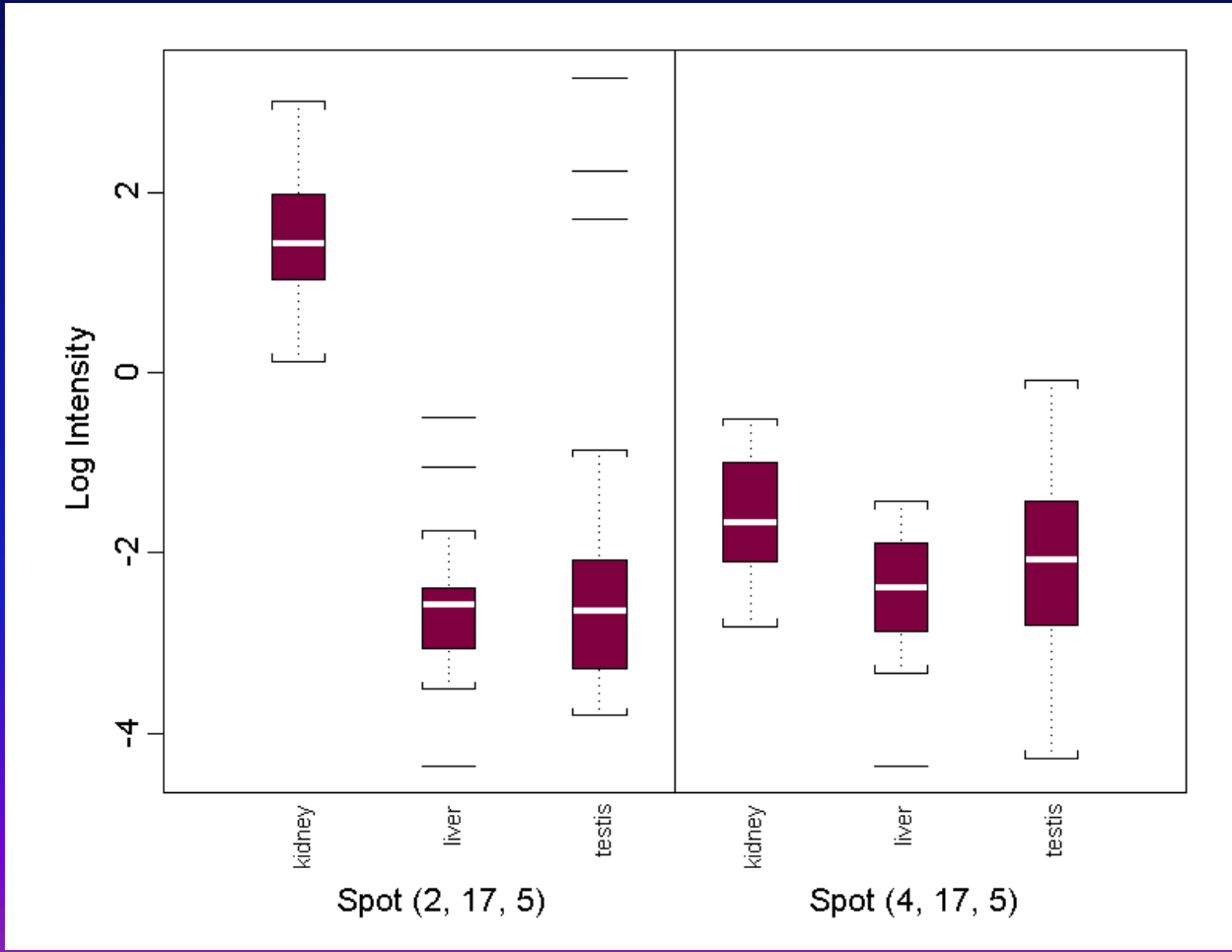
Example: Villin



Example: Villin



Example: Villin



Definition of Abundance

- If the UniGene database entry for "expression information" says that the sources of the clones found in a cluster included "kidney", then we will say that the gene is abundant in kidney.
- Similar definitions obviously apply for liver, testis, or other organs.

Abundance by Consistency

Abundance	All UniGene	Consistent	Ambiguous
None	409	237	172
Kidney	129	76	53
Liver	284	169	115
Testis	372	231	141
Kidney, Liver	126	69	57
Kidney, Testis	226	146	80
Liver, Testis	960	609	351
All	2798	1835	963

Combining UniGene Abundance with Microarray Data

- For each gene
  - Let **I** = (K,L,T) be the binary vector of its abundance in three organs as recorded in the UniGene database
  - Let **Y** = (*k, l, t*) be the measured log intensity in the three organs
- Model as 3D multivariate normal

$$\mathbf{Y} \mid \mathbf{I} = N_3(\mu_I, \Sigma_I)$$

Implementation Note

- We need a natural way to collect data from separate microarray experiments into measurement triples
  - Average replicate experiments from same mouse using same dye color
- Use consistently annotated genes to estimate model parameters

Estimated mean log intensity

Abundance	μ_K	μ_L	μ_T
None	2.027	2.129	2.012
Kidney	2.445	1.880	1.822
Liver	1.911	2.909	1.743
Testis	1.734	1.809	2.872
Kidney, Liver	3.282	3.051	1.961
Kidney, Testis	2.410	2.129	2.521
Liver, Testis	2.438	2.563	2.526
All	3.202	3.121	2.958

Distinguishing Between Competing Sets of Annotations

- Use parameters estimated from genes with consistent annotations
- At ambiguous spots, can compute log-likelihood of observed data for each possible triple of abundance annotations
- Given a complete set of annotations, can sum log-likelihood values over all genes

Distinguishing Between Competing Sets of Annotations

- Log-likelihood that kidney file contains correct annotations is equal to –52241
- Log-likelihood that liver file contains correct annotations is equal to –60183

Scrambled Rows

- We think the annotation problem was caused by reordering data rows
- We permuted the rows 100 times to obtain empirical p-values for the log-likelihoods:

$$P(\text{kidney}) < 0.01$$

$$P(\text{liver}) = 0.57$$

Future Directions

- The log-likelihood of the kidney file annotations was not close to the maximum of –33491
- Suggests that we can combine the microarray data with the UniGene expression data to refine the notion of abundance (more highly expressed in specific organs) on a gene-by-gene basis.