GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

4 November 2004

Lecture 19: Applied Clustering

- Project Normal
  - Project Normal Clustering
  - Abnormal Behavior
  - Problems and Solution
- The NCI-60 cell lines

Project Normal

- Eighteen samples
  - Six C57BL6 male mice
  - Three organs: kidney, liver, testis
- Reference material
  - Pool RNA from all eighteen mouse organs
- Replicate experiments on two-color arrays with common reference
  - Four experiments per mouse organ
  - Dye swaps: two red samples, two green samples

Original analysis of Project Normal

Reference: Pritchard, Hsu, Delrow, and Nelson. (2001) *Project normal: defining normal variance in mouse gene expression*. PNAS **98**: 13266–13271.

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

First steps

We chose to process the data using a simple global normalization (to the 75^{th} percentile) instead of loess normalization, since we believed that the mixed reference RNA should have a different distribution of intensities than RNA from a single organ. We then transformed the intensities in each channel by computing their base-two logarithm.

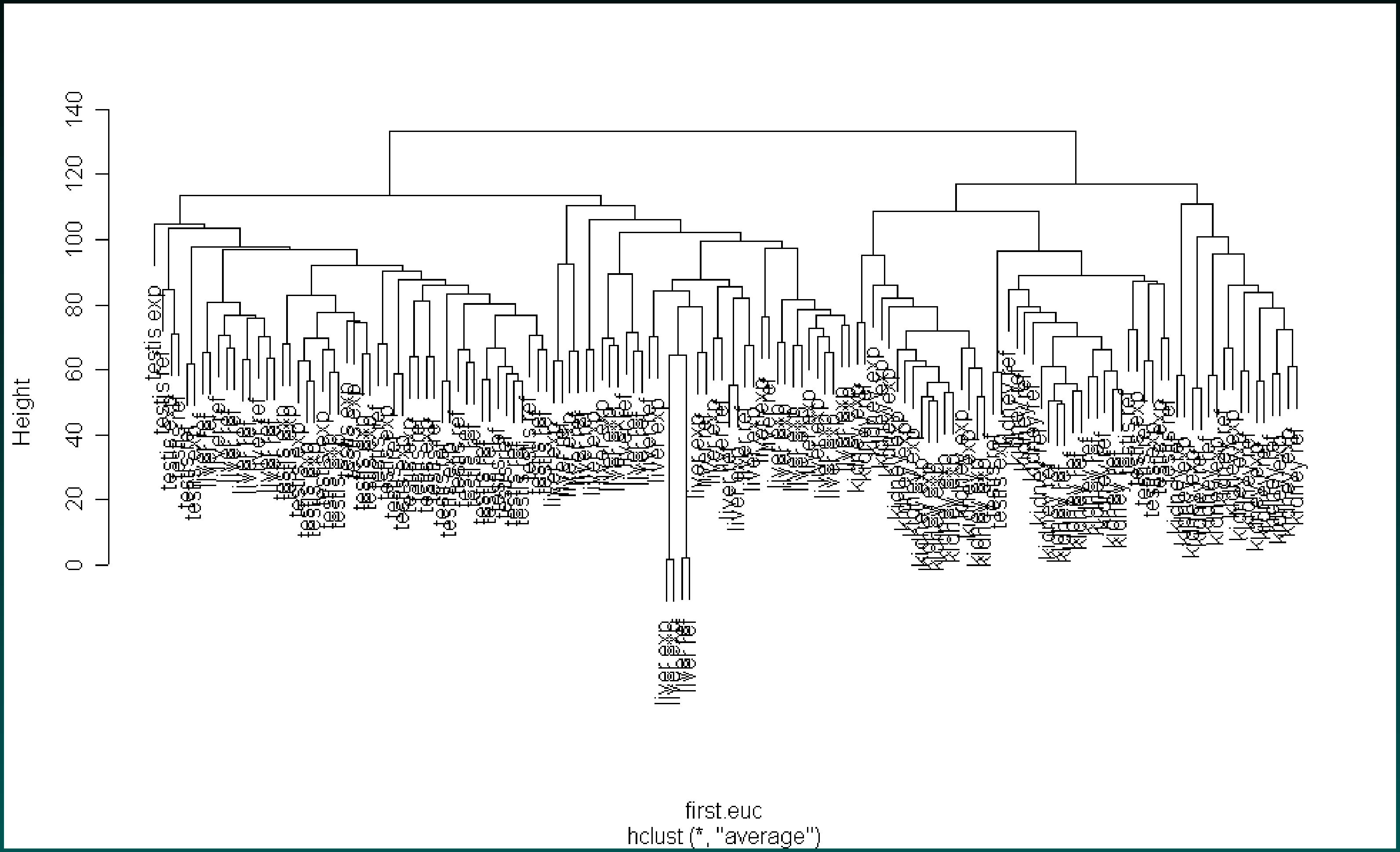
Main Question: Can we determine from the project normal data set which genes are specifically expressed in each of the three organs?

Clustering Methods

If we cluster the data, what should we expect to see? Which clustering method would be most appropriate for a first look at the data?

- Hierarchical clustering
- Partitioning around medoids
- K-means
- Multidimensional scaling
- Principal components analysis

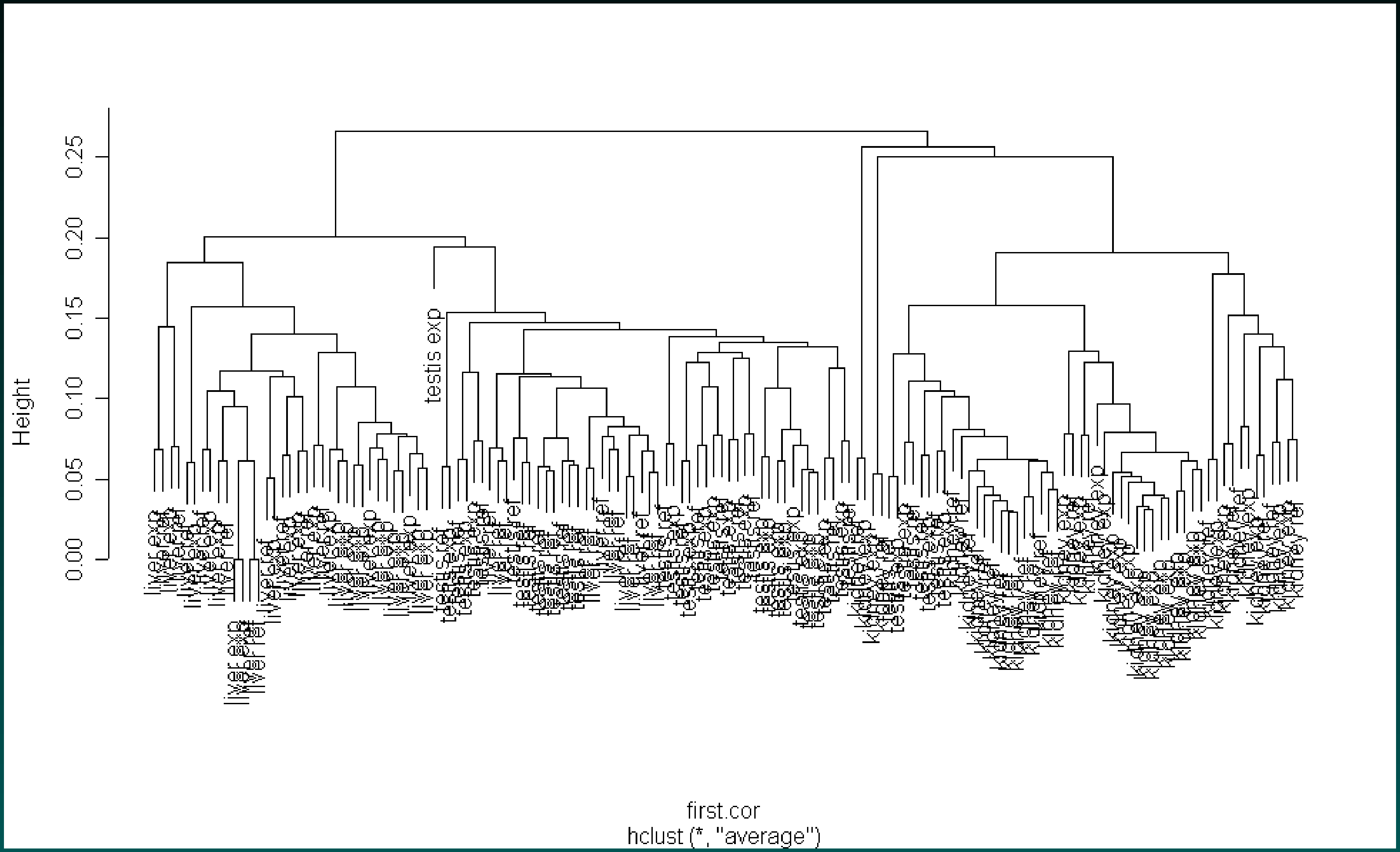
Hierarchical clustering



Euclidean distance, average linkage

Back to clustering methods

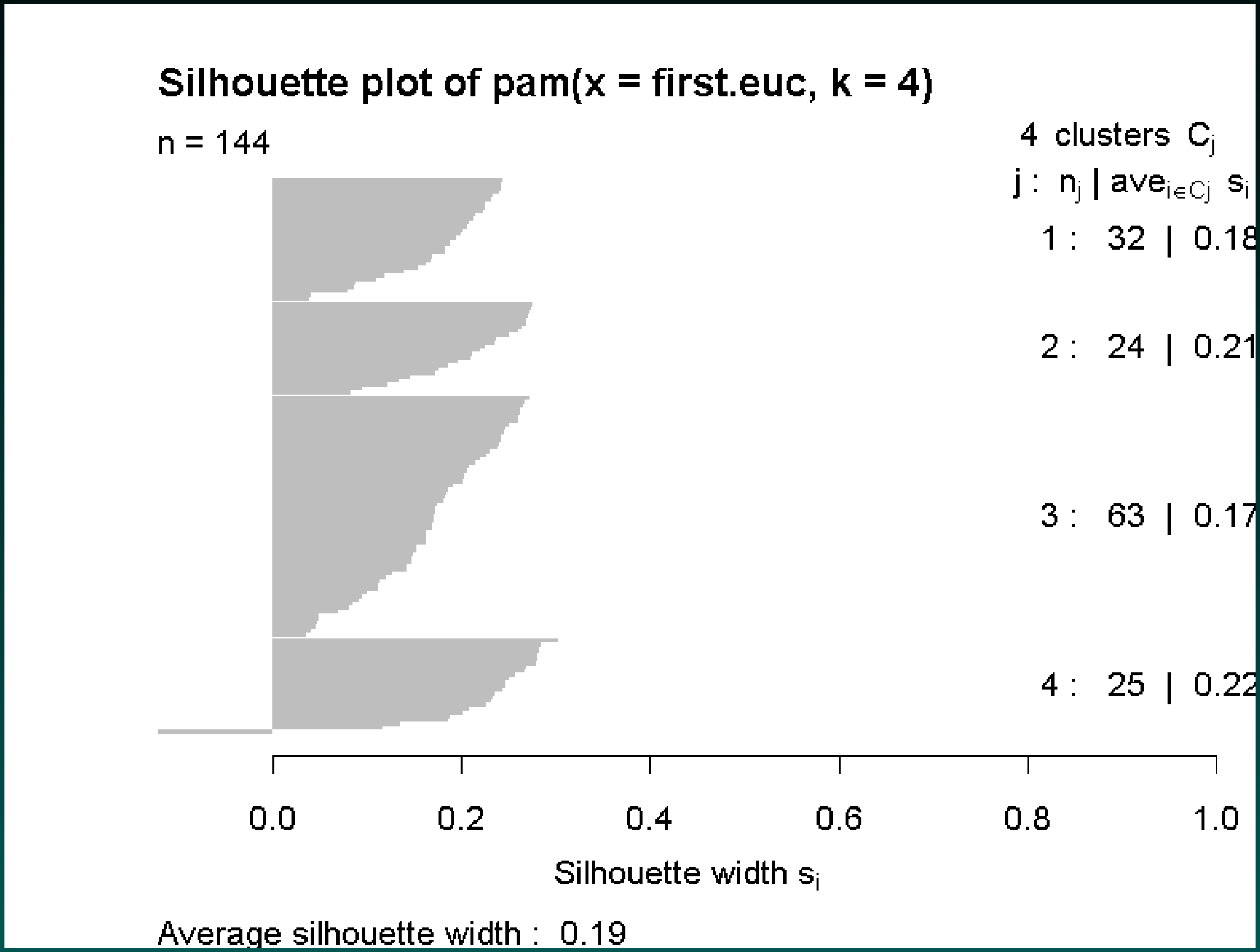
Hierarchical clustering



Correlation distance, average linkage

Back to clustering methods

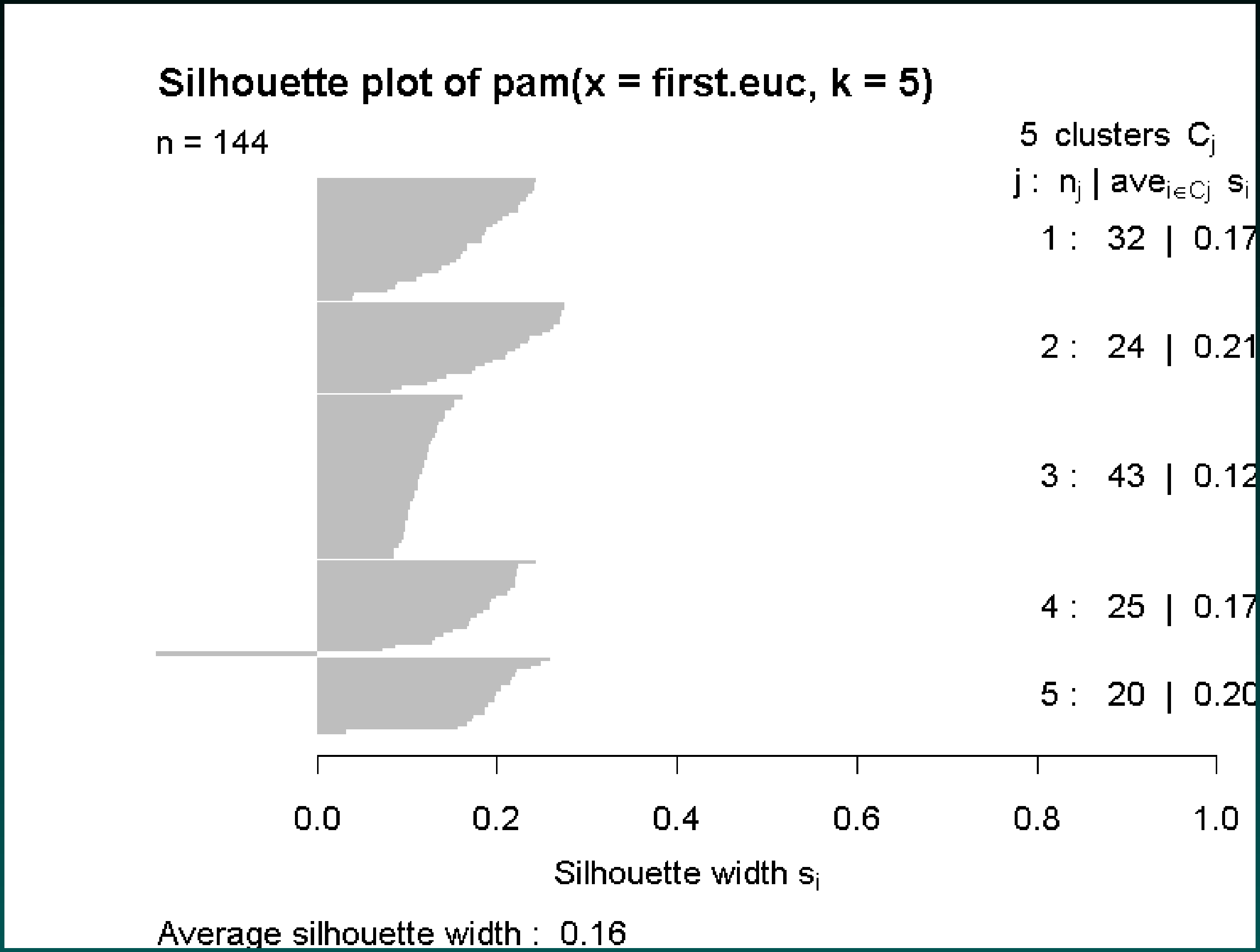
Partitioning Around Medoids



Euclidean distance, four clusters

Back to clustering methods

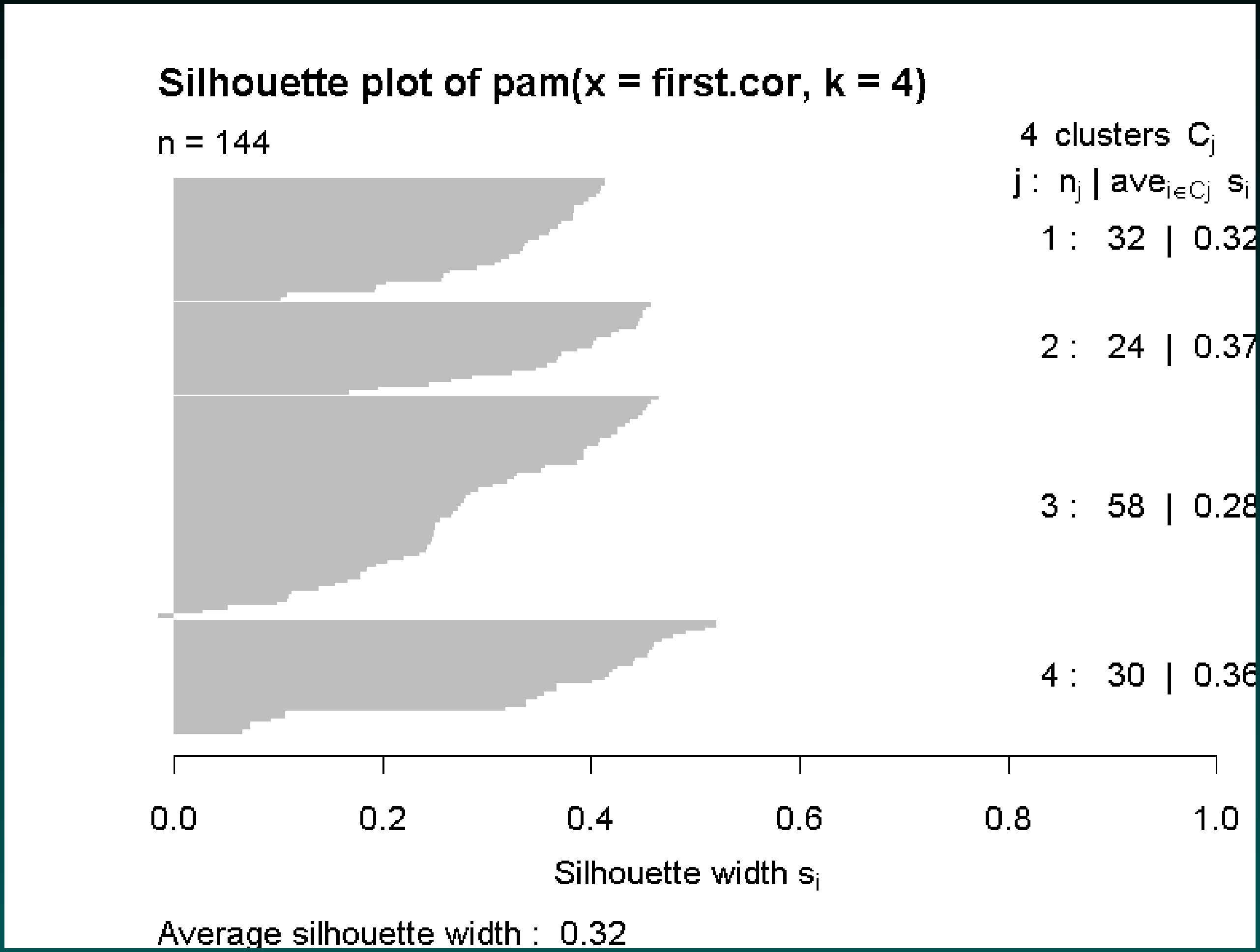
Partitioning Around Medoids



Euclidean distance, five clusters

Back to clustering methods

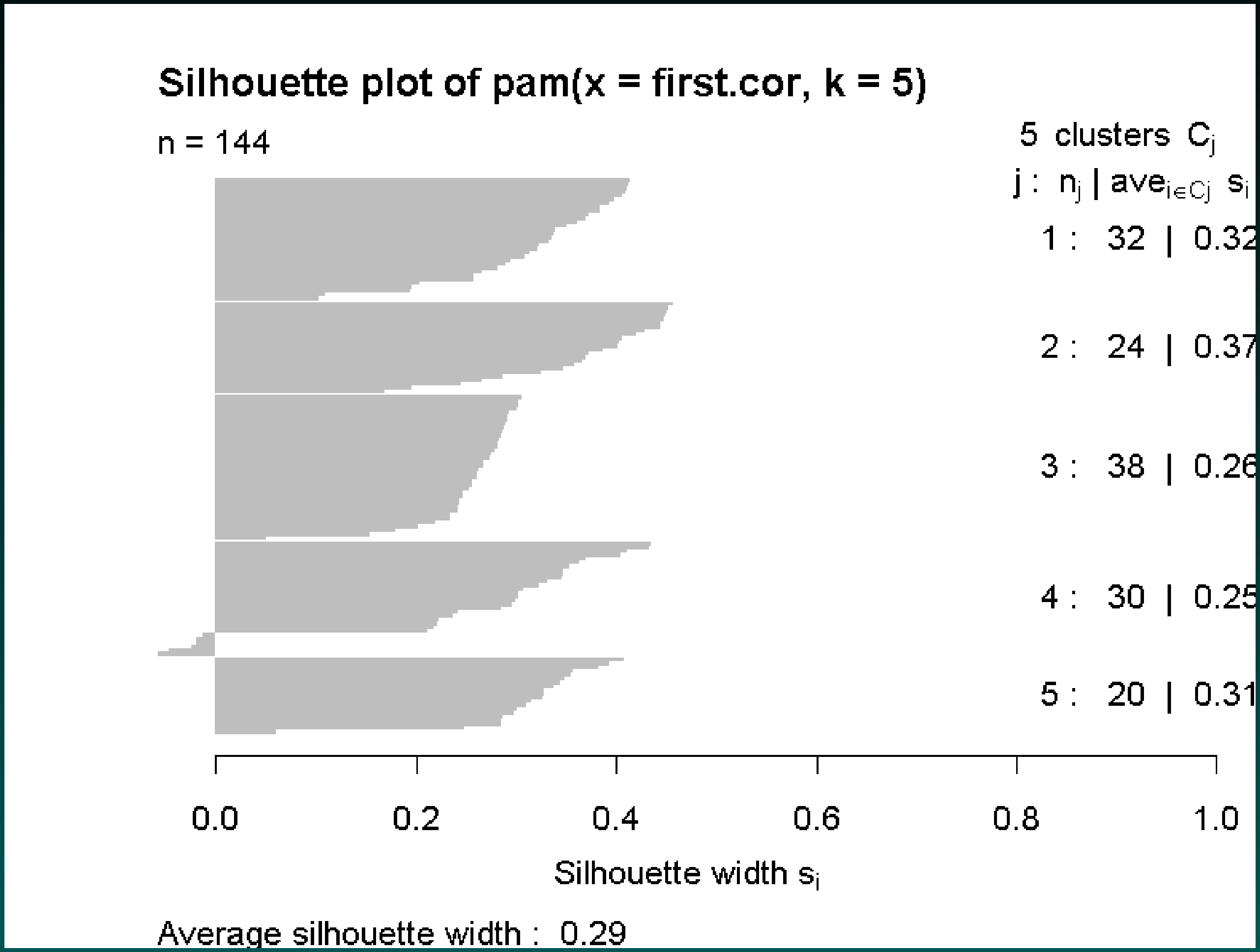
Partitioning Around Medoids



Correlation distance, four clusters

Back to clustering methods

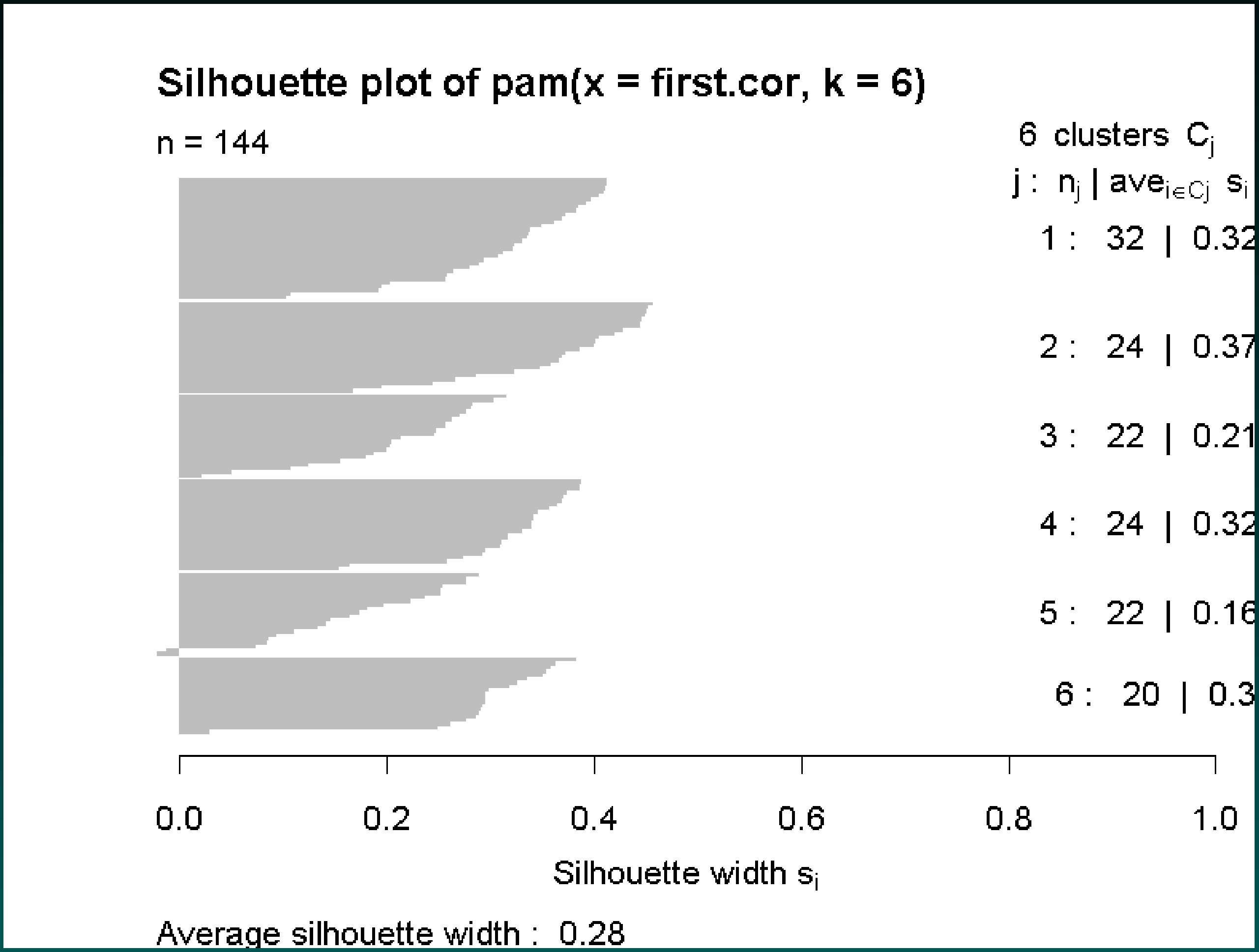
Partitioning Around Medoids



Correlation distance, five clusters

Back to clustering methods

Partitioning Around Medoids



Correlation distance, six clusters

Back to clustering methods

K-means

Number of channels in each cluster:

Channel	C1	C2	C3	C4
Experiment	4	24	20	24
Reference	28	0	44	0

Organ	C1	C2	C3	C4
Kidney	24	24	0	0
Liver	0	0	24	24
Testis	8	0	40	0

Best of 50 runs with four clusters

Back to clustering methods

K-means

Number of channels in each cluster:

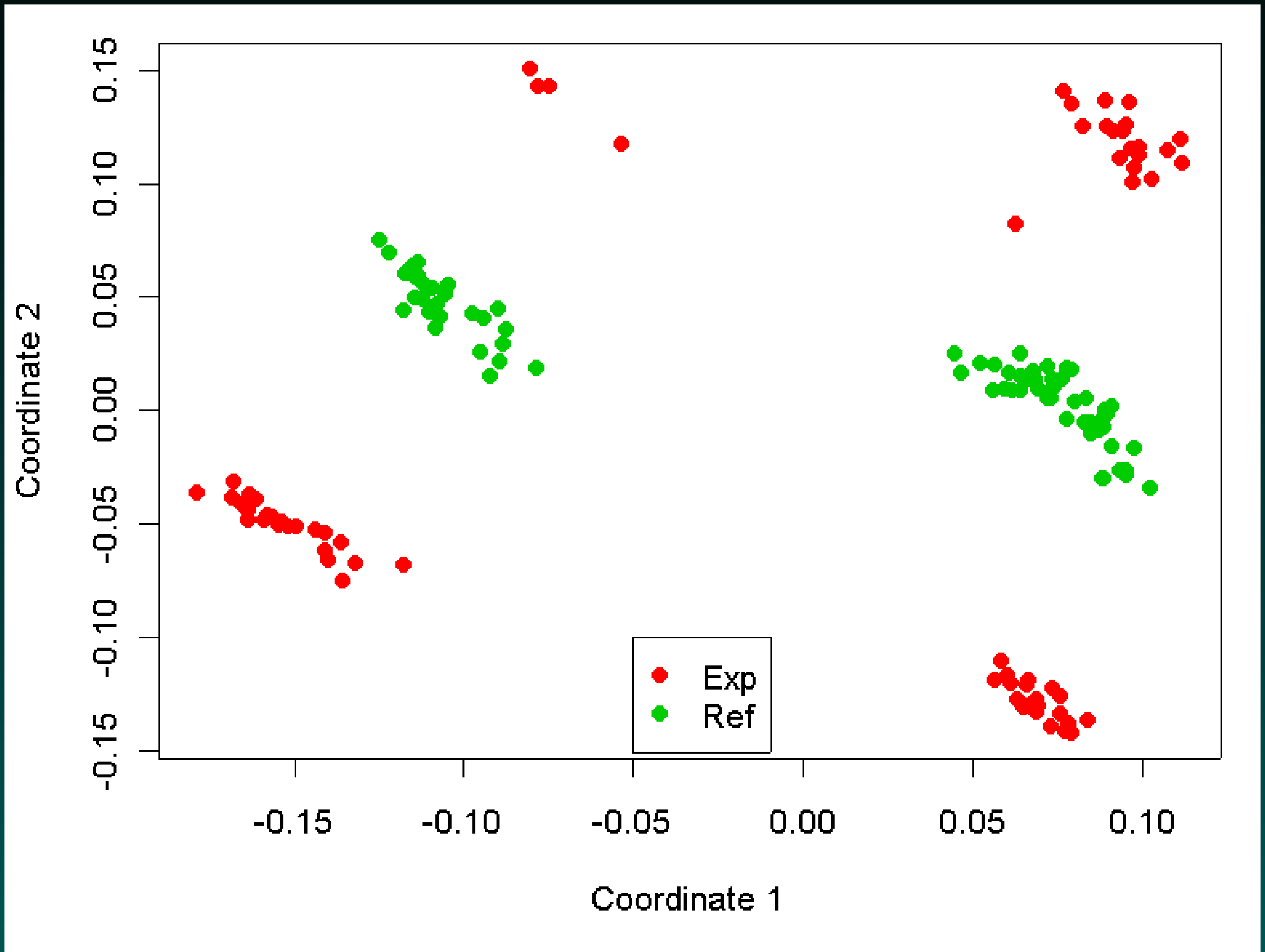
Channel	C1	C2	C3	C4	C5
Experiment	4	16	20	8	24
Reference	20	0	44	8	0

Organ	C1	C2	C3	C4	C5
Kidney	16	16	0	16	0
Liver	0	0	24	0	24
Testis	8	0	40	0	0

Best of 50 runs with five clusters

Back to clustering methods

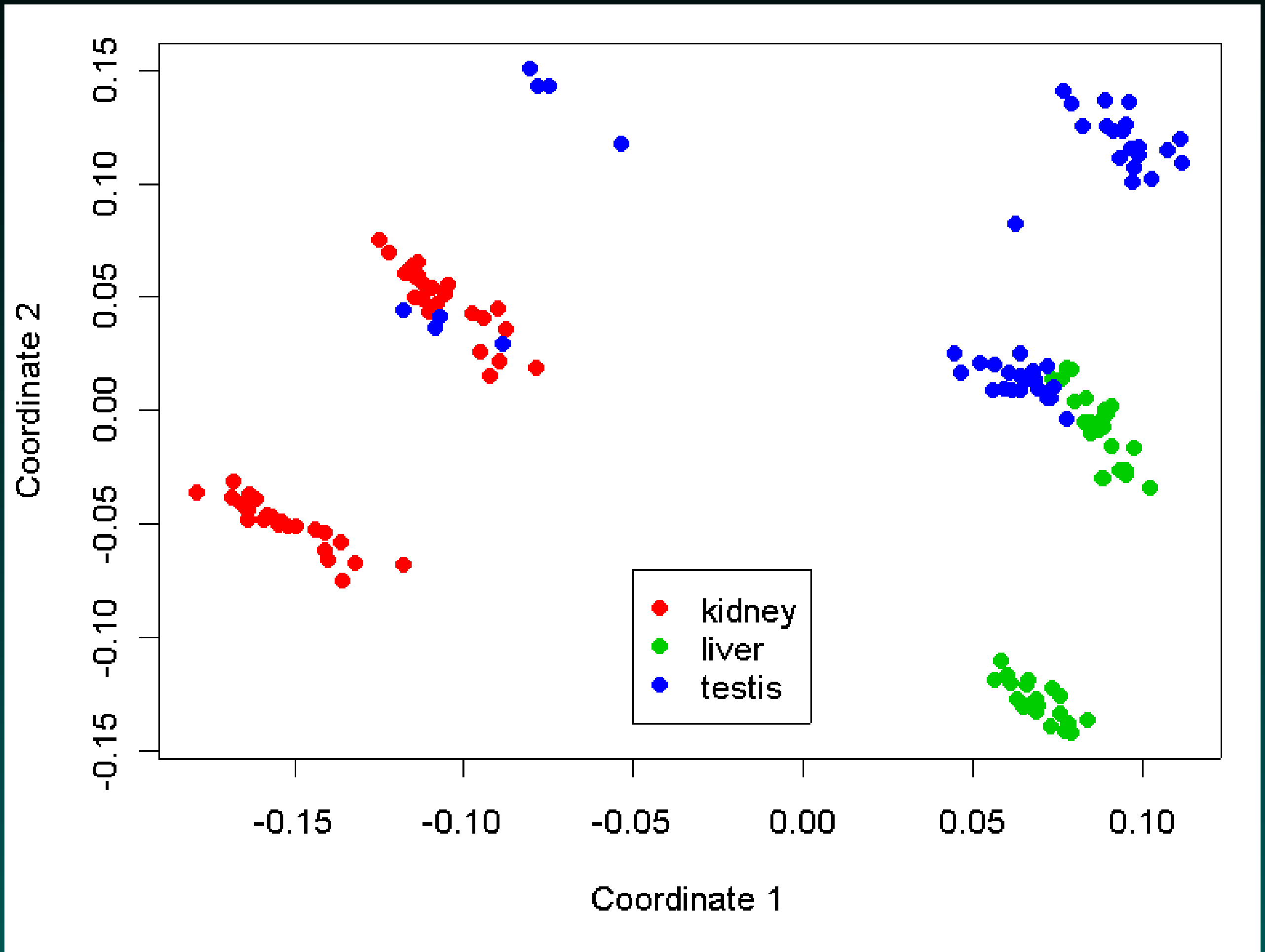
Multidimensional Scaling



Correlation distance, colored to indicate channel

Back to clustering methods

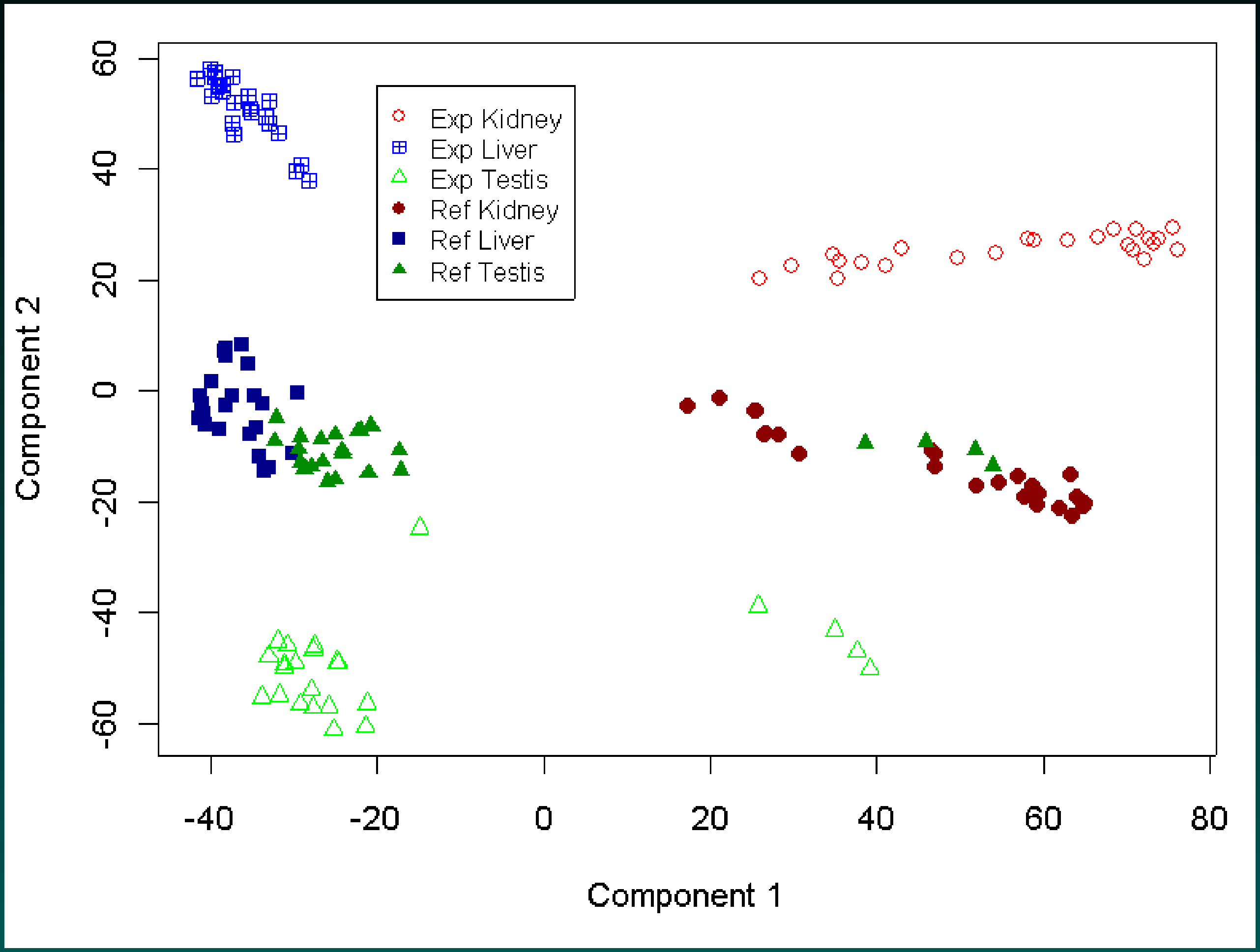
Multidimensional Scaling



Correlation distance, colored to indicate organ

Back to clustering methods

Principal components analysis



Euclidean distance, indicating channel and organ.

Back to clustering methods Forward to second PCA

Abnormal Behavior

Regardless of which exploratory method we use to look at the data, we see that sometihng strange is happening here.

We might not have noticed this behavior if we had immediately gone to the log ratios instead of clustering the separate channels.

What might explain the presence of two different kinds of reference channels? First thought: dye swaps. But this doesn't make sense, since then we would expect the experimental channels to split the same way (giving us eight clusters in total).

Data merging

- Data was supplied in three files, one each for kidney, liver and testis.

- Each row in each file contained two kinds of annotations:

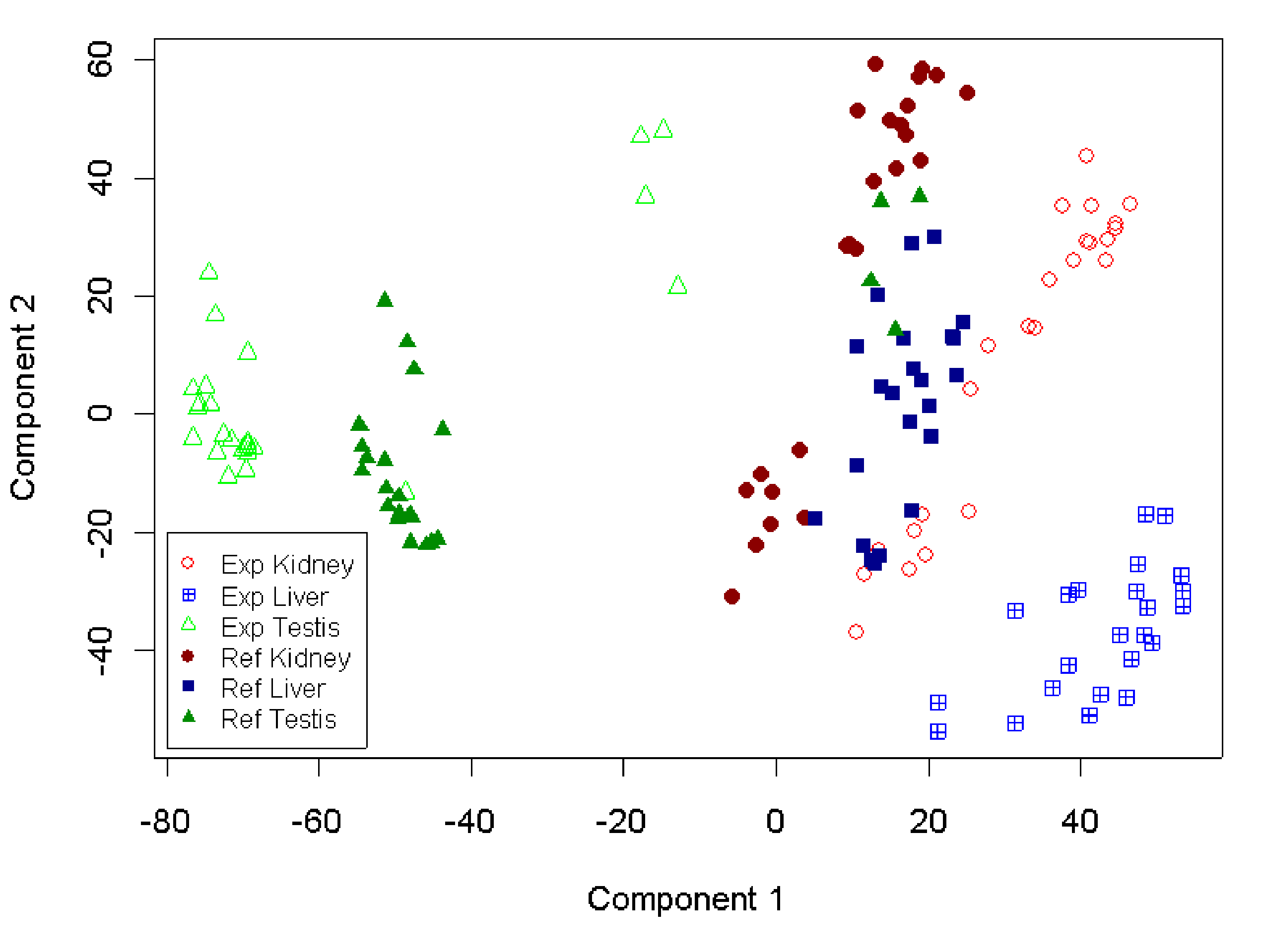
1. Location (block, row, and column)
2. Genetic material (IMAGE clone, UniGene ID)

- For our analysis, we merged the data using the annotations of genetic material.

- As it turns out, the locations did not agree

- So, we reordered data rows and merged on location...

PCA after merging data on location



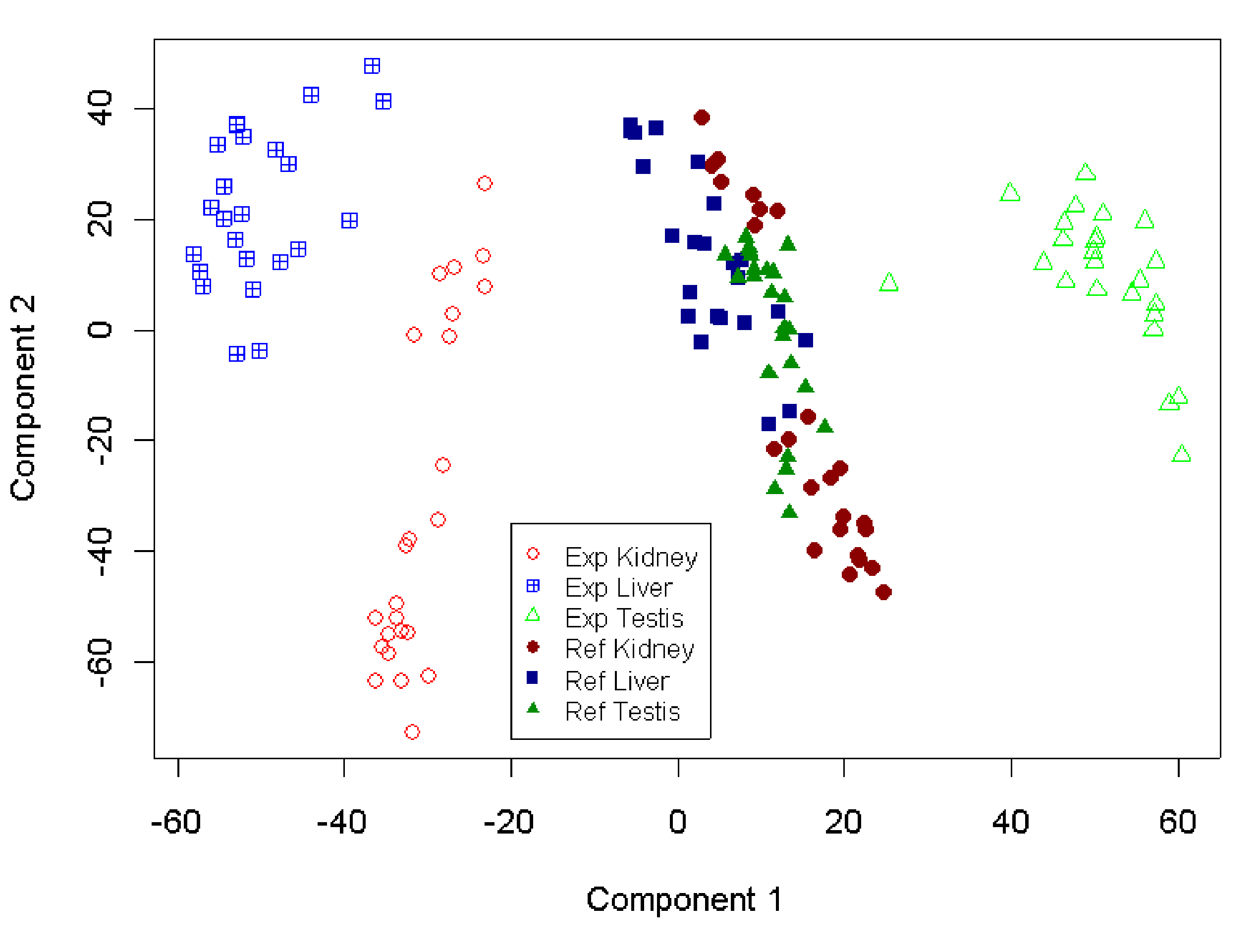
Yuck. So why are most of the testis references so weird?

Back to first PCA Forward to third PCA

Inspired guessing

- When the gene annotations are matched
  - Four of the testis reference channels are close to the kidney reference
  - Twenty of the testis reference are close to the liver reference
- When the location annotations are matched
  - Kidney, liver, and 4 testis references are close
  - The other 20 testis reference are off by themselves
- Conclusion: A data processing error occurred partway through the testis experiments.

Principal components, take 3



Finally, the picture we expected to start with!

Back to second PCA

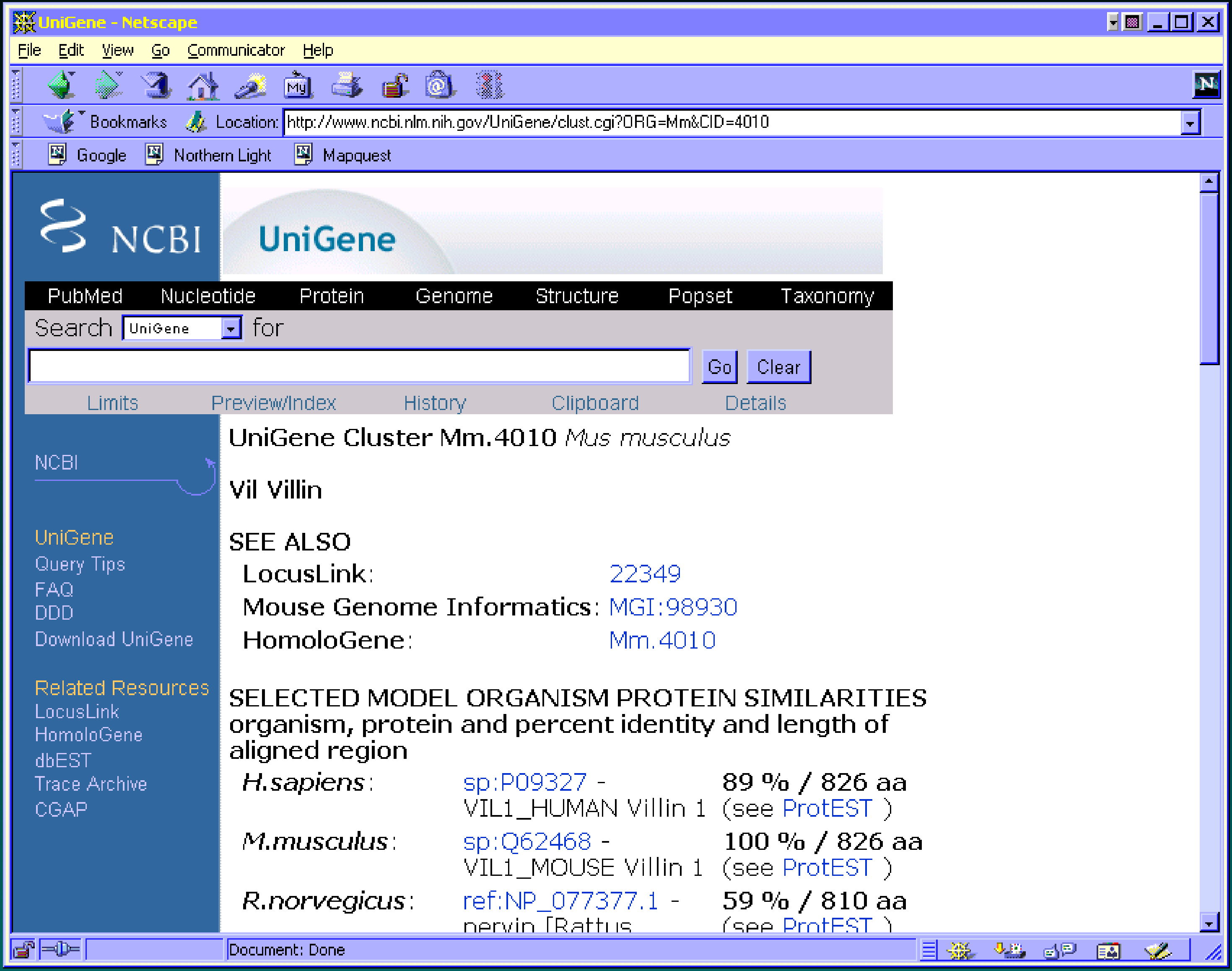
Every solution creates a new problem

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

UniGene Example



Location: http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Mm&CID=4010

UniGene Cluster Mm.4010 *Mus musculus*

Vil Villin

SEE ALSO

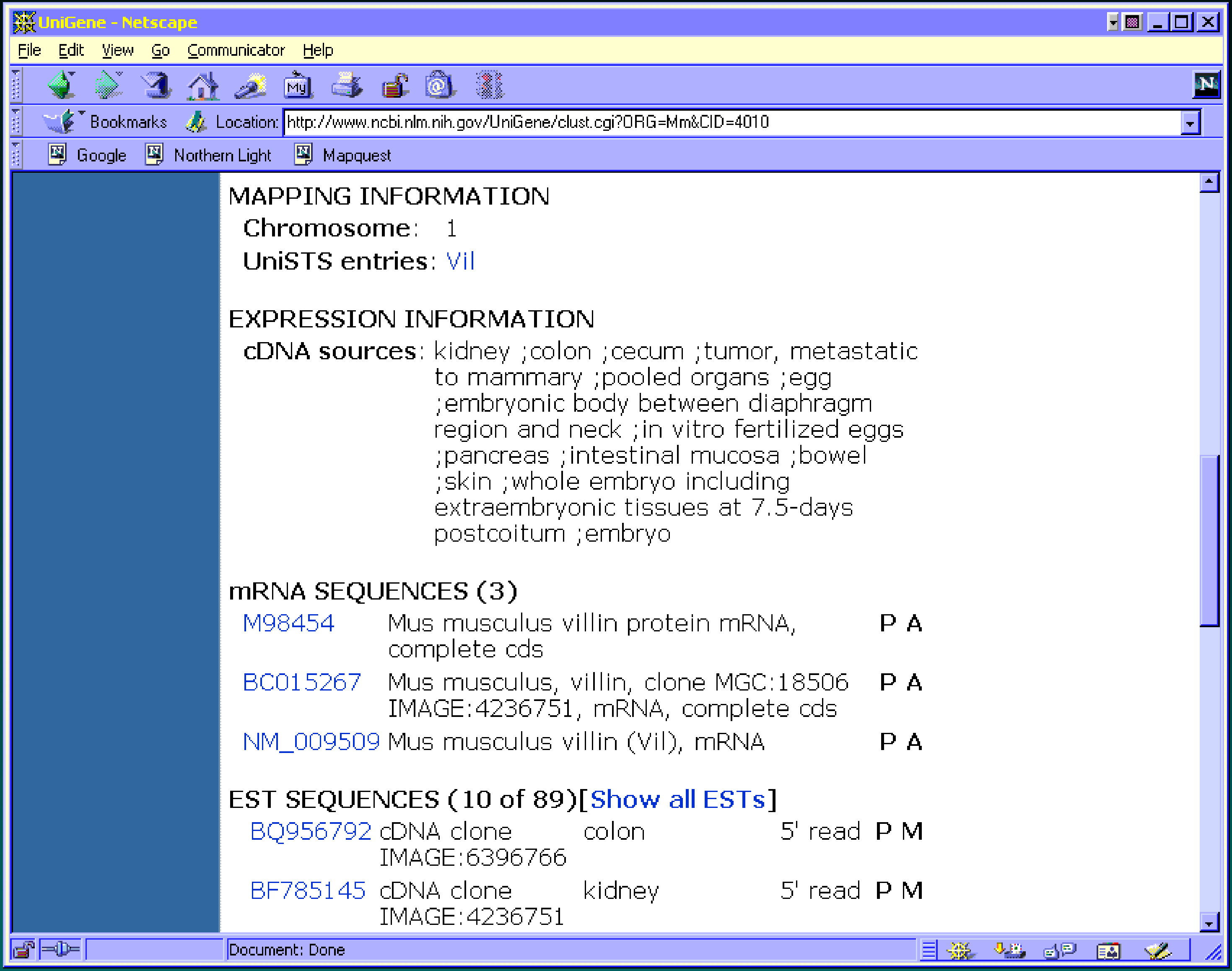
LocusLink:	22349
Mouse Genome Informatics:	MGI:98930
HomoloGene:	Mm.4010

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES
organism, protein and percent identity and length of aligned region

H.sapiens:	sp:P09327 - VIL1_HUMAN Villin 1	89 % / 826 aa (see ProtEST)
M.musculus:	sp:Q62468 - VIL1_MOUSE Villin 1	100 % / 826 aa (see ProtEST)
R.norvegicus:	ref:NP_077377.1 - nervin [Rattus	59 % / 810 aa (see ProtEST)

© Copyright 2004, Kevin R. Coombes and Keith A. Baggerly

UniGene Example



http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Mm&CID=4010

MAPPING INFORMATION

Chromosome: 1

UniSTS entries: Vil

EXPRESSION INFORMATION

cDNA sources: kidney ;colon ;cecum ;tumor, metastatic to mammary ;pooled organs ;egg ;embryonic body between diaphragm region and neck ;in vitro fertilized eggs ;pancreas ;intestinal mucosa ;bowel ;skin ;whole embryo including extraembryonic tissues at 7.5-days postcoitum ;embryo

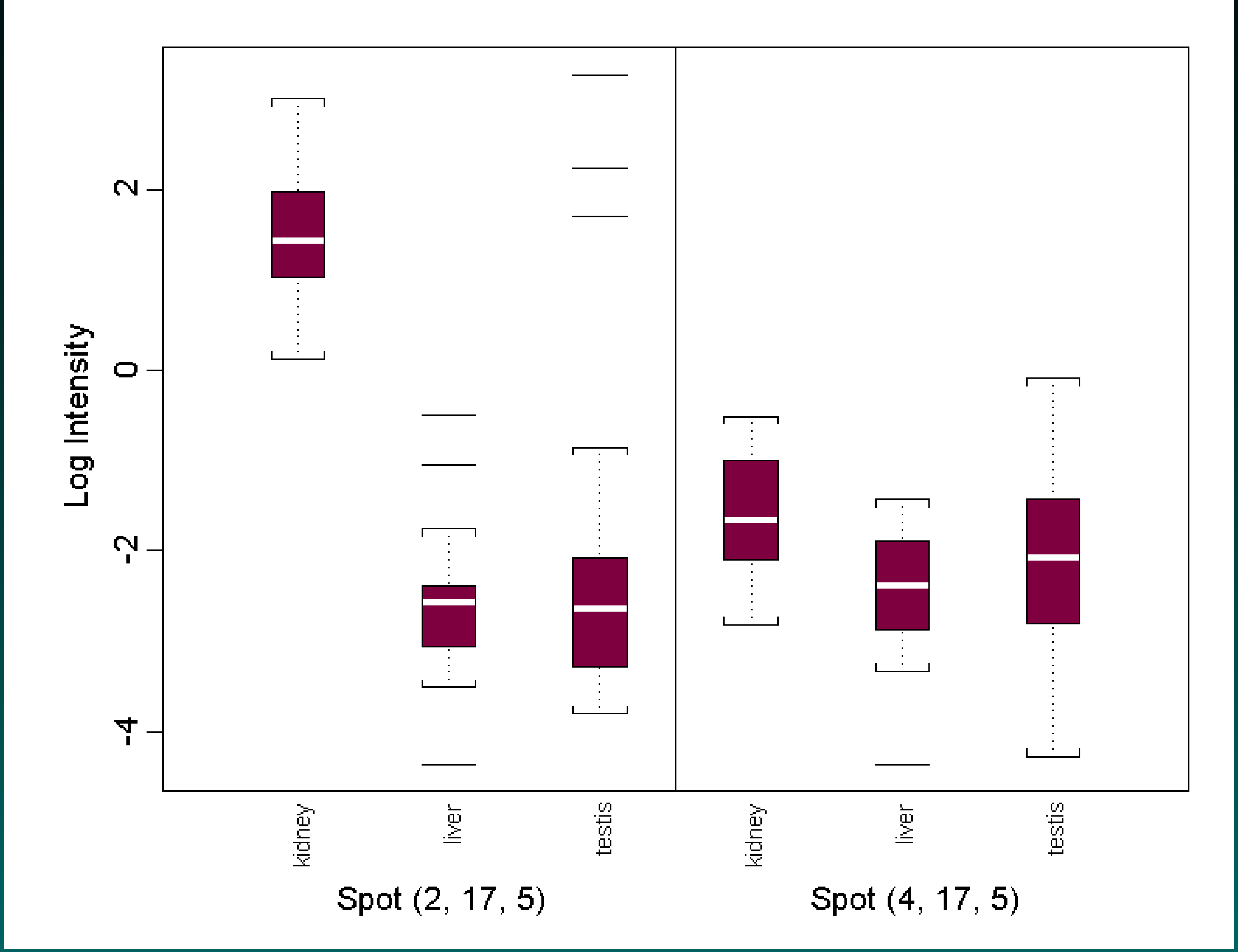
mRNA SEQUENCES (3)

M98454	Mus musculus villin protein mRNA, complete cds	P A
BC015267	Mus musculus, villin, clone MGC:18506 IMAGE:4236751, mRNA, complete cds	P A
NM_009509	Mus musculus villin (Vil), mRNA	P A

EST SEQUENCES (10 of 89)[Show all ESTs]

BQ956792	cDNA clone IMAGE:6396766	colon	5' read	P M
BF785145	cDNA clone IMAGE:4236751	kidney	5' read	P M

Villin Expression



Definition of abundance

- If the UniGene database entry for “gene expression” says that the cDNA sources of the clones found in a cluster included “kidney”, then we will say that the gene is abundant in kidney.
- Analogous 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	2789	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 te UniGene database.
  - Let $Y = (k, l, t)$ be the measured log intensity in the three organs.
- Model using a 3-dimensional multivariate normal distribution

$$Y|I = N_3(\mu_I, \Sigma_I)$$

- Average replicate experiments from same mouse with same dye to produce natural triplets of measurements.

Use consistently annotated genes to fit the model

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

The estimates support the idea that (UniGene) abundant genes are expressed at higher levels than (UniGene) “rare” genes.

Distinguishing between competing sets of annotations

- Use parameters estiomated from the genes with consistent annotations
- At the ambiguos spots, compute the log-likelihood of the observed data for each possible triple of abundance annotations
- Given a complete set of annotaiosn, sum the log-likelihood values over all genes
  - Log-likelihood that the kidney data file contains the correct annotations is equal to $-52,241$
  - Log-likelihood that the liver data file contains the correct annotations is equal to $-60,183$

Scrambled rows

- Our “inspired guess” earlier was motivated by the idea that the rows containing the annotations had somehow been reordered.

- We permuted the rows 100 times to obtain empirical p-values for the observed log-likelihoods
  - P(kidney is correct) < 0.01
  - P(liver is correct) $= 0.57$.

- The log-likelihood of the kidney file annotations was not particularly close to the maximum of $-33,491$. This suggest that we can use the array data to refine the notion of “abundance” on a gene-by-gene basis.

The NCI-60 cell lines

- Two-color fluorescence microarray experiments on the NCI-60 set of cancer cell lines.
- Original References
  - Ross et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet.* **24**: 227–35.
  - Scherf et al. (2000) A gene expression database for the molecular pharmacology of cancer. *Nat Genet.* **24**: 236–44.

Quality of UniGene Annotations

Number of Spots	UniGene Status (build 137)
294	None (controls)
128	Only 3’ – unknown to Unigene
1379	Only 3’ – known to Unigene
1	Only 5’ – unknown
6	Only 5’ – known
399	Both – unknown
763	Both – 3’ known, 5’ unknown
291	Both – 3’ unknown, 5’ known
646	Both known, but disagree
6093	Both known, and agree

We only trust the 7478 spots where the UniGene clusters are known and match.

Functional categories

Get functional categories of the genes on the microarray by mapping from UniGene to LocusLick to GeneOntology.

Function	Ann.	Spots	Function	Ann.	Spots
Oncogenesis	140	180	Cell shape and size	78	101
Apoptosis	128	138	Protein traffic	157	188
Physiological proc.	180	210	Transport	146	136
Perc. of ext. stimuli	238	150	Cell proliferation	197	249
Ectoderm devel.	129	152	Stress response	599	372
Mesoderm devel.	92	102	Radiation response	147	136
Cell adhesion	111	140	Cell cycle	494	283
Cell-cell signaling	137	166	Nucleic acid met.	695	595
Signal transduction	222	228	Protein metabolism	471	567
Intracell sig cascade	110	110	Lipid metabolism	146	156
Cell motility	120	153	Carbohydrate met.	103	97
Cell organization	98	118	Energy pathways	88	98

How well does a set of genes distinguish types of cancer?

- Three methods for assessment
  - Qualitative (MDS, PCA)
  - Quantiative (PCA + ANOVA)
  - Semi-quantitative (grading dendrograms)
    - A = cluster contains all and only one kind of cancer
    - B = all, with extras
    - C = all except one
    - D = all except one, with extras
    - E = all except two
    - F = all except two, with extras
- `http://bioinformatics.mdanderson.org/camda01.html`

Grading dendrograms by chromosome

ch	B	C	L	M	N	O	P	R	S	ch	B	C	L	M	N	O	P	R	S
1		B	A	D	F			D	B	13				D	E				
2		E	C	D		D	E	D	E	14		A	A		F				
3		C	E	D				E	F	15		C	B	C	F			C	
4			E	E			E	E		16									
5		A	A	D	F			E		17		A	A	D	F	E			E
6 .		C	A	D			E	E	D	18		E	D						
7		E	A	D		E		C	E	19				D		D			
8		E		C				D		20		E				C			
9		B	C	C		E	E	E		21									
10				D	E					22		A		E					E
11		E		C				C	D	X		B	A	D				E	D
12		B	C	C		E	E	E											

Heterogeneity

- Some cancers (colon, leukemia) are fairly easy to distinguish from others
- Some (breast, lung) are so heterogeneous as to be almost impossible to distinguish
- Some chromosomes (1, 2, 6, 7, 9, 12, 17) can distinguish many cancers.
- Some (16, 21) are essentially random

Can functional categories distinguish the origin of different types of cancer?

- Table for functional categories looks a lot like the table for chromosomes
- Some biological process categories (signal transduction, cell proliferation, cell cycle, protein metabolism) can distinguish many types of cancer
- Others (apoptosis, energy pathways) cannot