Analysis Techniques for Probe-level Microarray Data

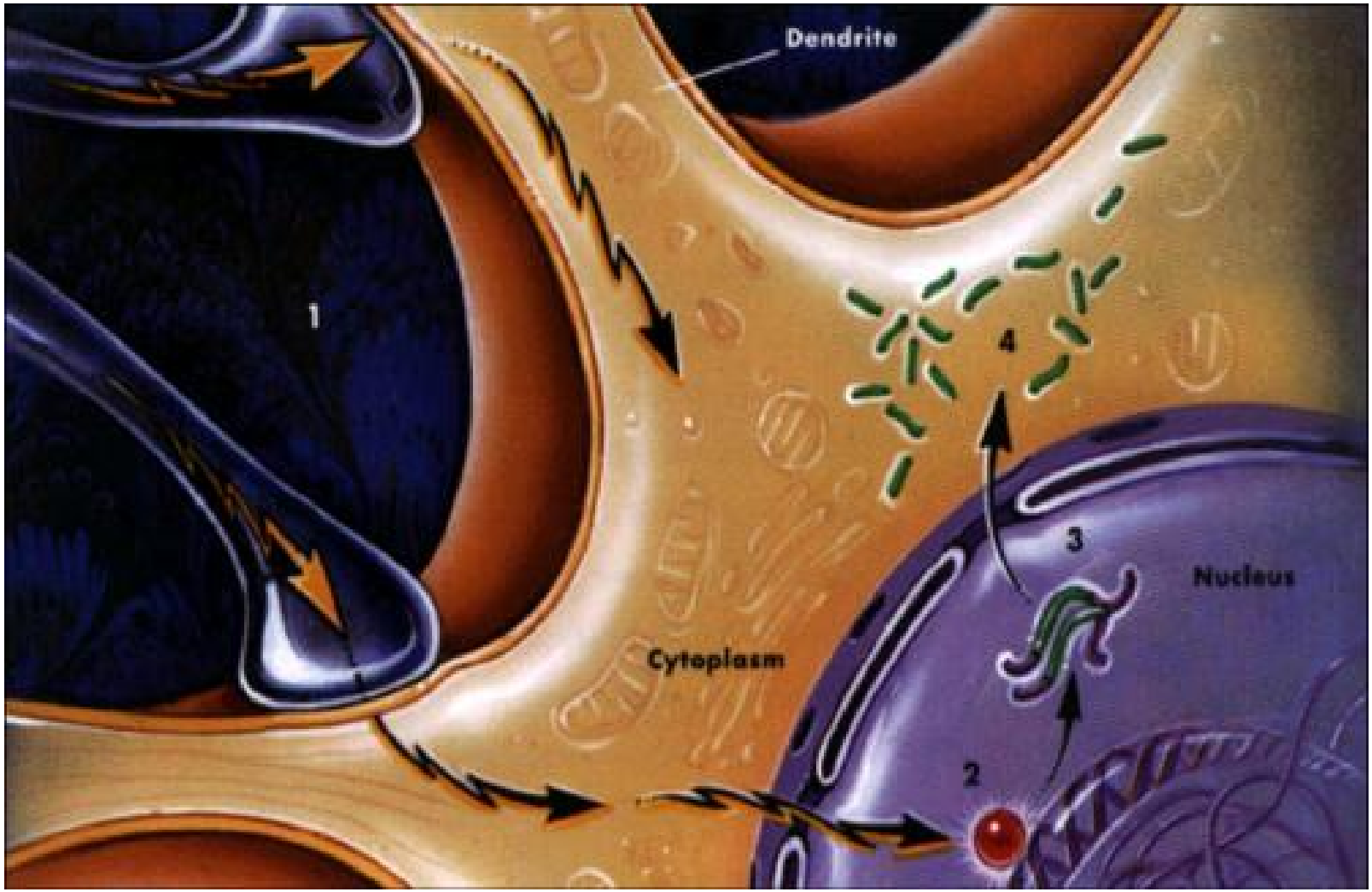
Humberto Ortiz Zuazaga

University of Puerto Rico
High Performance Computing facility

August 20, 2009

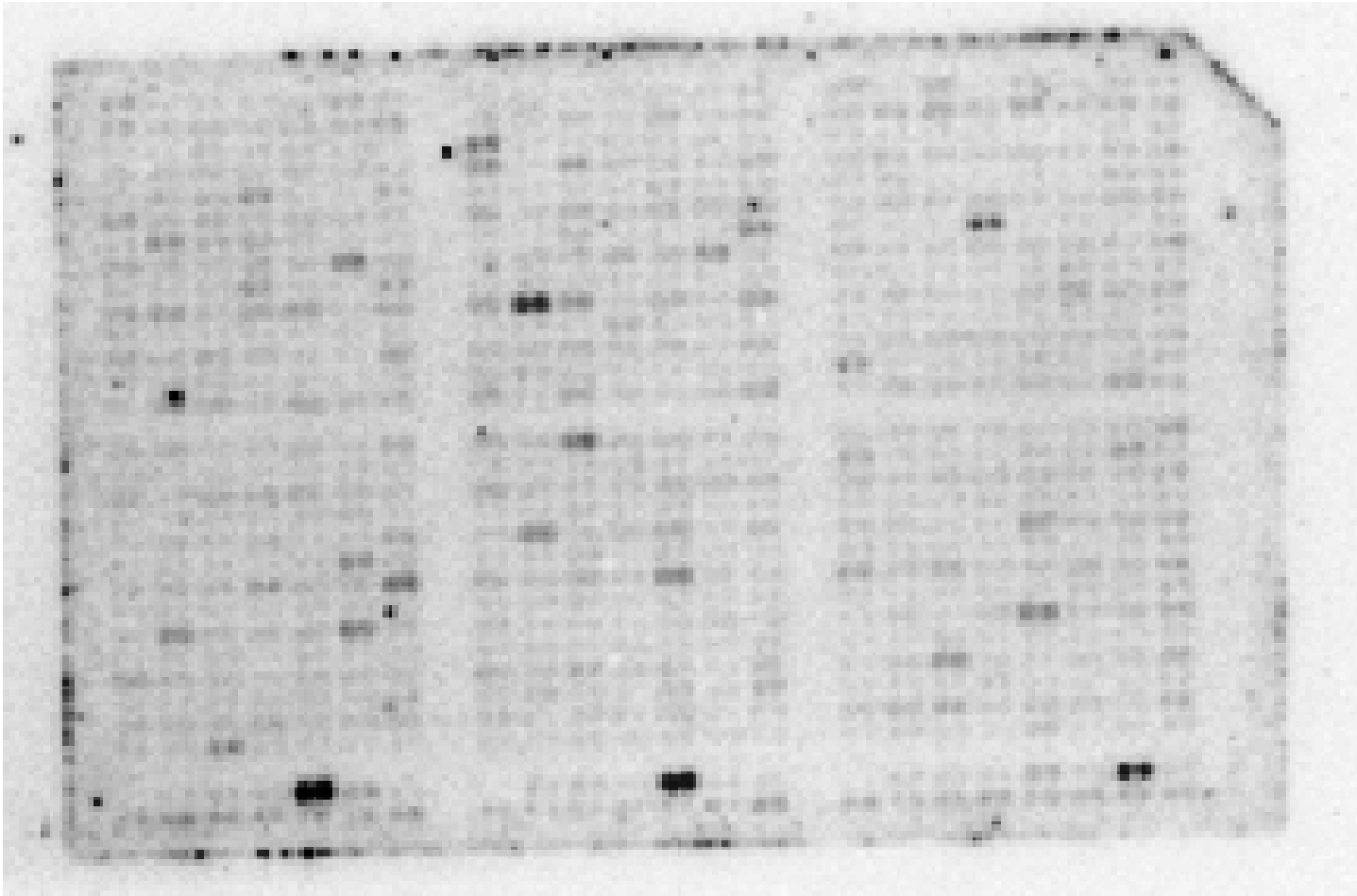
Outline

A Model Cell



Sample Array Image

or, "I've got all the genes, now what do I do with them?"



Outline

Conditioned taste aversion (CTA)

- associative aversive conditioning paradigm
- Animals are exposed to a novel taste, the *conditioned stimulus*
- An *unconditioned stimulus* induces malaise
- The animals develop a long lasting aversion to the conditioned stimulus

CTA Dataset

- two controls, the pre-treatment group and the one hour saline group
- four time points, 1, 3, 6, and 24 hours after conditioning
- 1185 genes on each spotted array
- 5 biological replicates of each array

Chiesa, R., Ortiz-Zuazaga, H. G., Ge, H. and Peña de Ortiz, S. (2000), Gene expression profiling in emotional learning with cDNA microarrays, *in* '40th meeting of the American Society for Cell Biology', San Francisco, California.

cAMP Responsive Element Binding protein (CREB)

- Transcription factor, known to regulate gene expression
- Plays important role in memory formation
- Binds to a DNA element called cAMP-response element (CRE)
- We will use this gene as an example of a canalyzing gene

R. Lamprecht, S. Hazvi, and Y. Dudai, "cAMP response element-binding protein in the amygdala is required for long- but not short-term conditioned taste aversion memory," *J. Neurosci.*, vol. 17, pp. 8443–8450, 1997.

Outline

Discretization

A01a glypican 1; HSPG M12; nervous system cell-surface heparan sulfate proteoglycan

Repetition	Pre	Sal	1 h	3 h	6 h	24h
1	0.172	0.099	0.176	0.142	0.062	0.152
2	0.274	0.168	0.126	0.114	0.104	0.276
3	0.003	0.119	0.552	0.178	0.193	0.114
4	0.114	0.139	0.6	0.311	0.179	0.181
5	0.04	0.006	0.172	0.103	0.036	-0.047
average	0.121	0.106	0.325	0.17	0.115	0.135
control	0.113					
diffs			0.212	0.057	0.002	0.022
epsilon						0.022
calls			+	+	0	0

Majority logic

Repetition	1 h	3 h	6 h	24h
1	+	0	−	0
2	−	−	−	+
3	+	+	+	+
4	+	+	+	+
5	+	+	0	−
consensus	+	+	?	+

Substituting averaged controls

Repetition	1 h	3 h	6 h	24h
1	+	+	−	+
2	0	0	0	+
3	+	+	+	0
4	+	+	+	+
5	+	0	−	−
cvac	+	+	?	+

Pruning extreme values

Repetition	Pre	Sal	1 h	3 h	6 h	24h
1	—	0.099	0.176	0.142	—	0.152
2	—	—	0.126	0.114	0.104	—
3	0.003	0.119	—	—	0.193	0.114
4	0.114	0.139	—	—	0.179	0.181
5	0.04	—	0.172	0.103	—	—
new average	0.052	0.119	0.158	0.12	0.159	0.149
new control	0.086					
new diffs			0.072	0.034	0.073	0.063
new epsilon	0.063					
new calls			+	0	+	0

Consistent calls

1. at least two of the above set of calls agrees in the last 4 columns of data (1 h, 3 h, 6 h, and 24h)
2. either the 1 h or the 24 h columns is a "0"
3. across the last 4 columns of data, the column exhibits the consecutive zeros property (*i.e.*, values do not oscillate between "0" and "+" or "−")

A01a is not consistent

	1 h	3 h	6 h	24h
average calls	+	+	0	0
consensus	+	+	?	+
cvac	+	+	?	+
new calls	+	0	+	0

Clustering

- Categorizing each timepoint for each gene into coarse divisions yields a clustering of genes
- In our current experiment there are $3^4 = 81$ possible clusters that a gene may fall into
- Longer time series or larger fields will allow finer grained division of the genes into clusters

Outline

CTA Dataset

- 127 consistent genes in CTA dataset
- Grouping genes with same calls in 1 h – 24 h timepoints yields 23 clusters

Clusters

Cluster	Gene coordinate
- 0 0 0	A05f, A12k, A14g, B07m, B07n
0 + + +	B01n, B04i, B05l, B06j, B06l, C10l, D01j, D10l, E01k, E03j, E09c, E10e, E13i, E13l, F01e
+ 0 0 0	B07e, D04f, D09l
- - 0 0	A02m, A14e, B03i, C08i, F08c, F13e
- - + 0	C14k, E03l
0 0 + +	B13n, D14g, E06i, E10m, E11l, E13e, E14d
- + + 0	B06m, E02i
0 0 - -	A05n, D02a, F12e
0 - + +	B04j, E10l, F11j
+ - - 0	C01f, F01a
0 - + -	A11d, C09m
- + - 0	D12a
0 + - +	D09i
- - - 0	A02l, A03h, A09c, B10l, C02m, C04d, C04f, C06e, D02b, F02b, F03c, F09n, F11e
+ - + 0	B13a, F02a
0 0 0 -	A10l, C08a, C14g, C14l, D13e
+ + 0 0	A08l, B08b, C08j, F11k
+ + - 0	D04l
0 0 0 +	A07i, B09h, C10c, D08n, E03m, E04i, E13h, E14f
0 - - +	C01e, F12j
0 - - -	B05b, C13a, C13i, C14n, E02e, F04l, F06a, F11l
- + 0 0	A12m
+ + + 0	B01i, B07f, B10c, B10d, B14h, D08f, D09e, E03i, E14k, F02n, F05k, G15

Genes equivalent to canalyzing genes

- A canalyzing gene is one that is required for proper function of the organism
- We hypothesize that the canalyzing gene does not directly affect the organism, but acts through canalyzed genes
- Our clustering algorithm groups canalyzed and canalyzing genes together

Genes equivalent to CREB

- We search for genes with similar patterns of expression to CREB
- Obtained upstream sequences for "000+" cluster (1020 bp, 800 bp before start of transcription) expression most similar to CREB

Coord.	Accession	Description
A07i	L24388	galactosyltransferase-associated protein kina (GTA); CDC2-related protein kinase (CDC2L1)
B09h	L10362	synaptic vesicle protein 2B
C10c	L33869	ceruloplasmin (CERP; CP); ferroxidase
D08n	X63255	N-methyl-D-aspartate receptor subtype 1 (NMDAR NR1); glutamate receptor subunit zeta 1 (GRIN1)
E03m	M29712	melanin-concentrating hormone (PMCH; MCH)
E04i	V01228	calcitonin
E13h	M20713	guanine nucleotide-binding protein G(K) alpha 3 su unit (G(I) alpha 3 (GNAI3)
E14f	X06890	ras-related protein RAB4A

Transcription factor sites

- Searched sequences for transcription factor binding sites with TESS [1]
- Found two very interesting genes: Pmch and Calca, both have CRE sites

[1] `http://www.cbil.upenn.edu/cgi-bin/tess/tess`

Pmch

- Cyclic neuropeptide
- Affects appetite or metabolism
- Induces hippocampal synaptic transmission

Varas, M., Perez, M., Ramirez, O. and de Barioglio, S. (2002), 'Melanin concentrating hormone increase hippocampal synaptic transmission in the rat', *Peptides* **23**(1), 151–155.

Calca

- Vasodilator
- May be involved in axonal regeneration
- May be involved in synaptogenesis

Li, X. Q., Verge, V. M., Johnston, J. M. and Zochodne, D. W. (2004), 'CGRP peptide and regenerating sensory axons', *J. Neuropathol. Exp. Neurol.* **63**(10), 1092–1103.

Outline

Summary

- We have developed an error correction procedure for microarray data.
- Each gene is described as upregulated, unchanged or downregulated, corresponding to a natural division of expression values.
- Additional heuristics reduce the set of genes being examined to a small group of candidate genes.
- The procedure results in genes being grouped into discrete clusters.
- We found two candidate genes, Pmch and Calca.
- Multiple sources of evidence support the theory that these genes are involved in biological processes related to memory.

Acknowledgments

Portions of this work were supported by a SCORE grant (S06GM08102), an INBRE grant (P20RR016470) and an IDeA Program grant (P20RR15565), from the National Institutes of Health.
Data provided by Dr. Sandra Peña de Ortiz.

Outline

General Introduction

Error Correction and Clustering
Introduction
Methods
Results
Conclusions

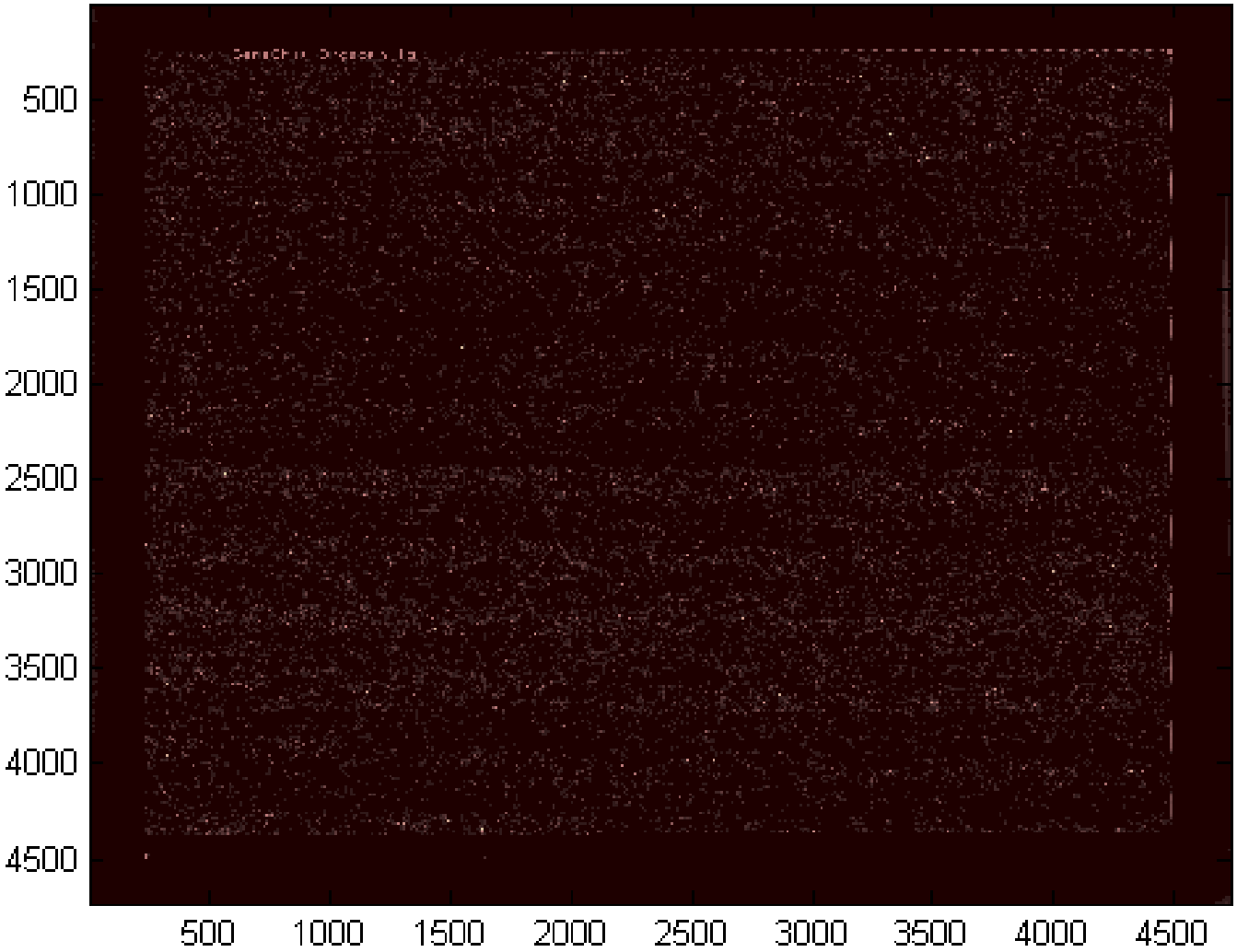
Probe-level data
Introduction
Methods
Results
Conclusions

Future work
Motivation
Methods
Preliminary results

Affymetrix probe-level data

- Microarray feature density continues to increase
- Oligo arrays have shorter probe sequences (cost)
- Many arrays utilize multiple probes per gene or feature
- probe sets are summarized to obtain a single value

Affymetrix array image



Drosophila data

- Experiment designed to test microarray analysis methods
- Two separate microarray technologies (Affymetrix, NimbleGen)
- Time course data, 0, 6 hours and 24 hours after training
- Massed and Spaced training
- Affymetrix: 14 probes per probeset, 14,000 probesets
- NimbleGen: about 10 probes per probeset, 12,000 probesets

Outline

Discretization

- Very similar to methods in prior chapter
- Take the mean over all repetitions
- Take average of controls
- Select discretization threshhold ϵ so that one of the remaining points is within ϵ of the control
- perform consensus, consensus vs mean, mean and trimmed mean analyses
- will define a new measure of "consistent" genes

Majority Logic Decoding Probes


```
[['0', '-', '0', '+'],
 ['0', '-', '+', '+'],
 ['0', '-', '0', '+'],
 ['0', '-', '+', '+'],
 ['0', '-', '+', '+'],
 ['-', '-', '0', '0'],
 ['0', '-', '0', '+'],
 ['0', '-', '0', '+'],
 ['-', '-', '-', '0'],
 ['0', '-', '0', '+'],
 ['-', '-', '0', '0'],
 ['0', '-', '0', '+'],
 ['0', '-', '0', '+'],
 ['0', '-', '+', '+']]

'mld': ['0', '-', '0', '+']
```

Weighted Mutual Information

We want to measure concordance and informativeness of probes.

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} w(x,y) p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Weights

We set the weights such that similar patterns of expression are given higher weights, and opposite expression is given lower weight:

$$w(x,y) = \begin{cases} 1.0 & \text{if } x = y,\ x, y \neq ? \\ 0.5 & \text{if } x = ? \text{ or } y = ? \\ 0.1 & \text{otherwise} \end{cases}$$

Example WMI


```
ac = ['0', '-', '0', '+']
nc = ['0', '-', '-', '-']
```

$$
\begin{aligned}
I(\mathtt{ac},\mathtt{nc}) &= w(-,-)p(-,-)\log\left(\frac{p(-,-)}{p(-)p(-)}\right) + \\
&\quad w(0,-)p(0,-)\log\left(\frac{p(0,-)}{p(0)p(-)}\right) + \\
&\quad w(+,-)p(+,-)\log\left(\frac{p(+,-)}{p(+)p(-)}\right) + \\
&\quad w(0,0)p(0,0)\log\left(\frac{p(0,0)}{p(0)p(0)}\right) \\
&= 1\cdot 1/4\cdot\log\left(\frac{1/4}{1/4\cdot 3/4}\right) + 0.1\cdot 1/4\cdot\log\left(\frac{1/4}{2/4\cdot 3/4}\right) + \\
&\quad 0.1\cdot 1/4\cdot\log\left(\frac{1/4}{1/4\cdot 3/4}\right) + 1\cdot 1/4\cdot\log\left(\frac{1/4}{2/4\cdot 1/4}\right) = 0.35
\end{aligned}
$$

SWAMI Score

- One Affymetrix probeset may map to several NimbleGen probesets
- Average WMI over all equivalent NimbleGen probesets
- Sum all average WMI scores over all Affymetrix probesets
- Obtain Summed Weighted Averaged Mutual Information or SWAMI score
- A single numerical value characterizing an analysis method
- Replaces consistency test from previous chapter

Outline

SWAMI Scores for several transformation and regression methods

Transformation	Regression	SWAMI
log2	Huber	182
log2	fair	186
log2	Cauchy	169
sqrt	Huber	212
sqrt	fair	230
sqrt	Cauchy	200
cuberoot	Huber	207
cuberoot	fair	216
cuberoot	Cauchy	202

SWAMI scores for error correction methods

Method	summarized	probes
trimmed mean	2753	3657
mean	2610	2535
consensus	1920	3058
consensus vs mean control	1888	1525

Outline

General Introduction

Error Correction and Clustering
Introduction
Methods
Results
Conclusions

Probe-level data
Introduction
Methods
Results
Conclusions

Future work
Motivation
Methods
Preliminary results

Summary

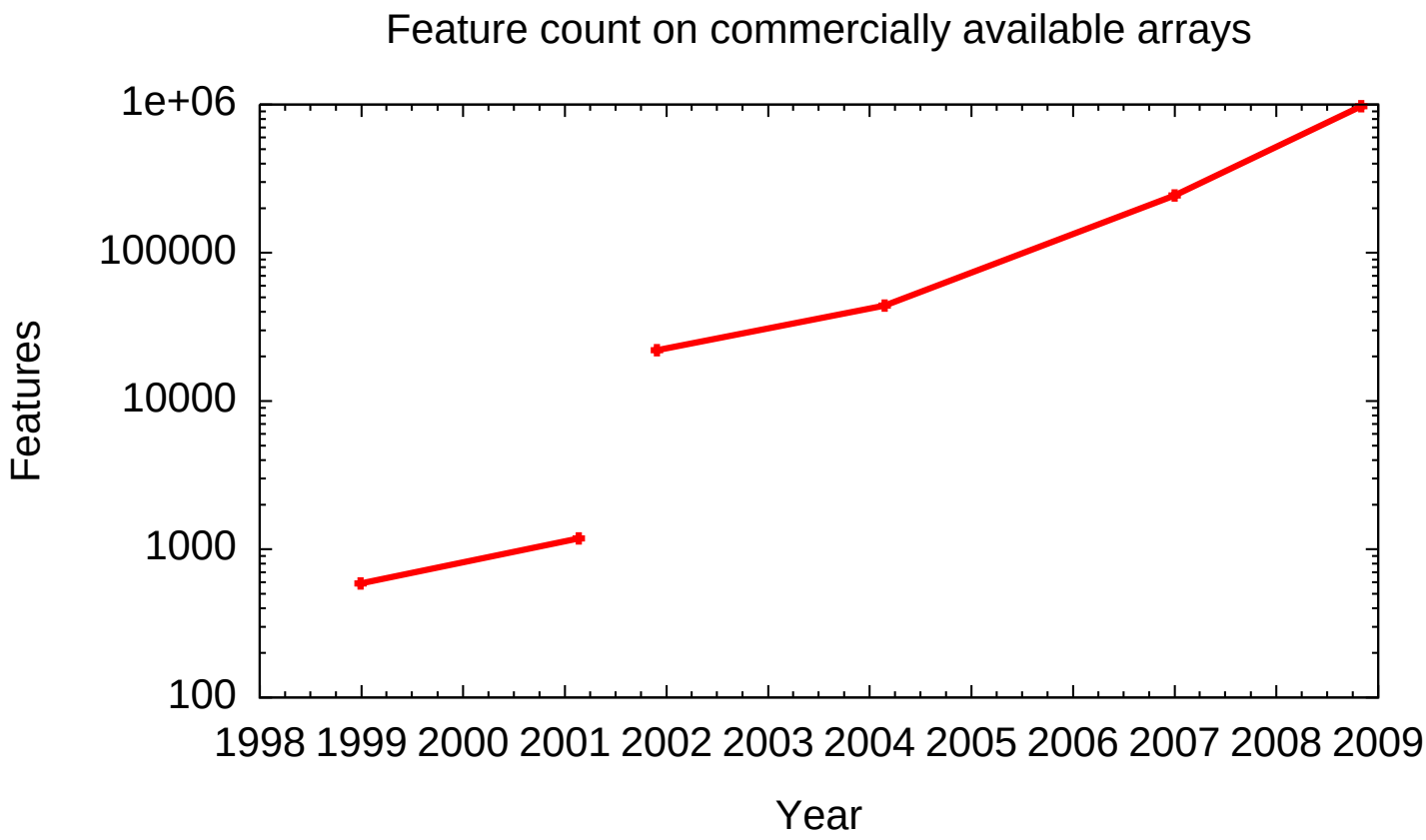
- We have designed a technique to validate microarray analysis methods
- The technique measures concordance between two data sets derived from different micorarray technologies on the same experimental conditions
- We have performed the experiment using data from an odor avoidance training time course on fruit flies
- The results demonstrate our prior error correction techniques produce the most concordance between the data sets in this experiment

Acknowledgments

This work was supported in part by a SCORE grant (S06GM08102) and an INBRE grant (P20RR016470) from the National Institutes of Health.
The data was kindly provided by Dart Neuroscience LLC.

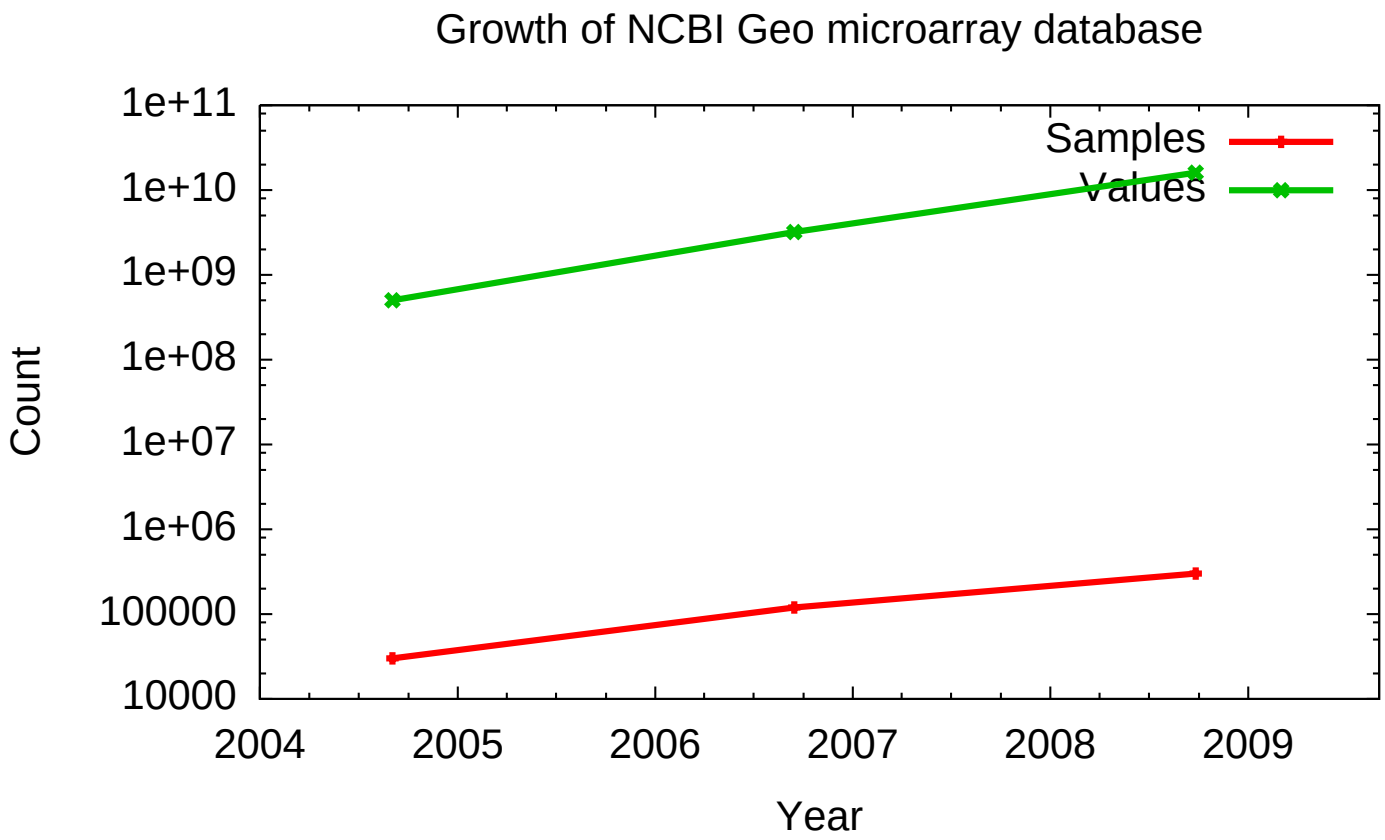
Outline

"Moore's law" for microarray features



Feature density increases exponentially.

"Moore's law" for microarray data



Extrachromosomal DNA

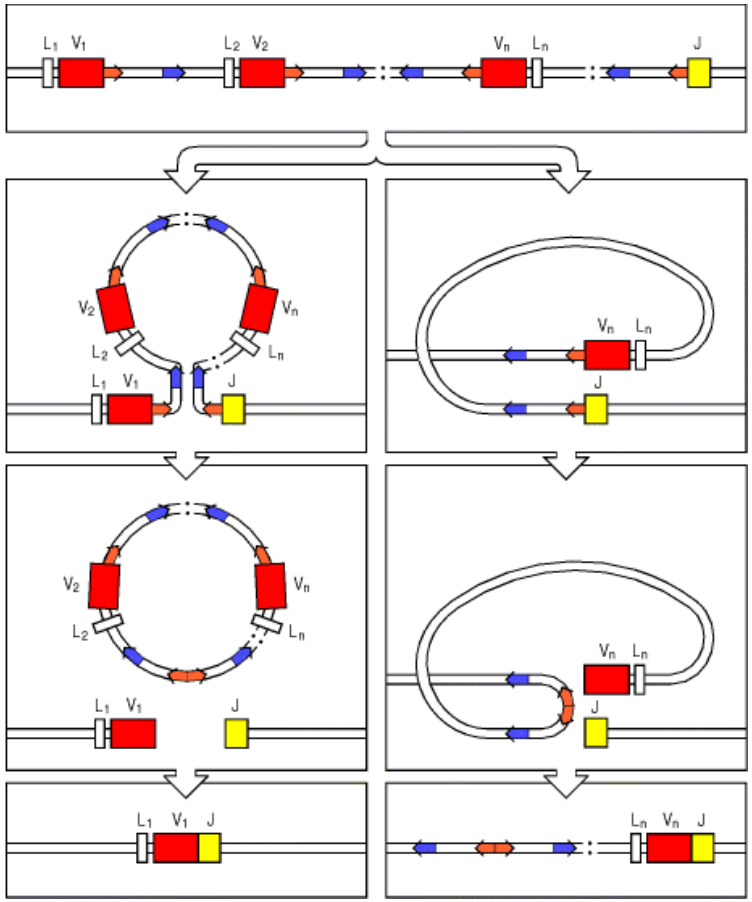
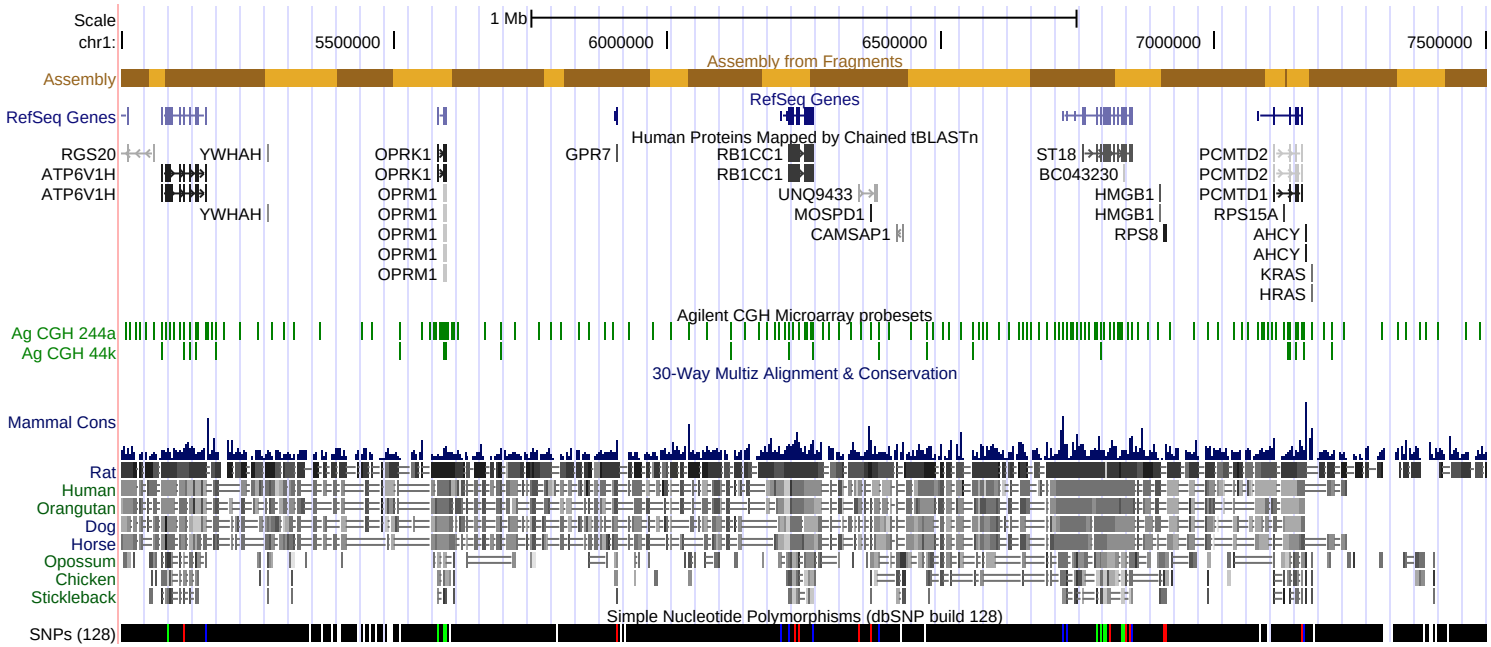


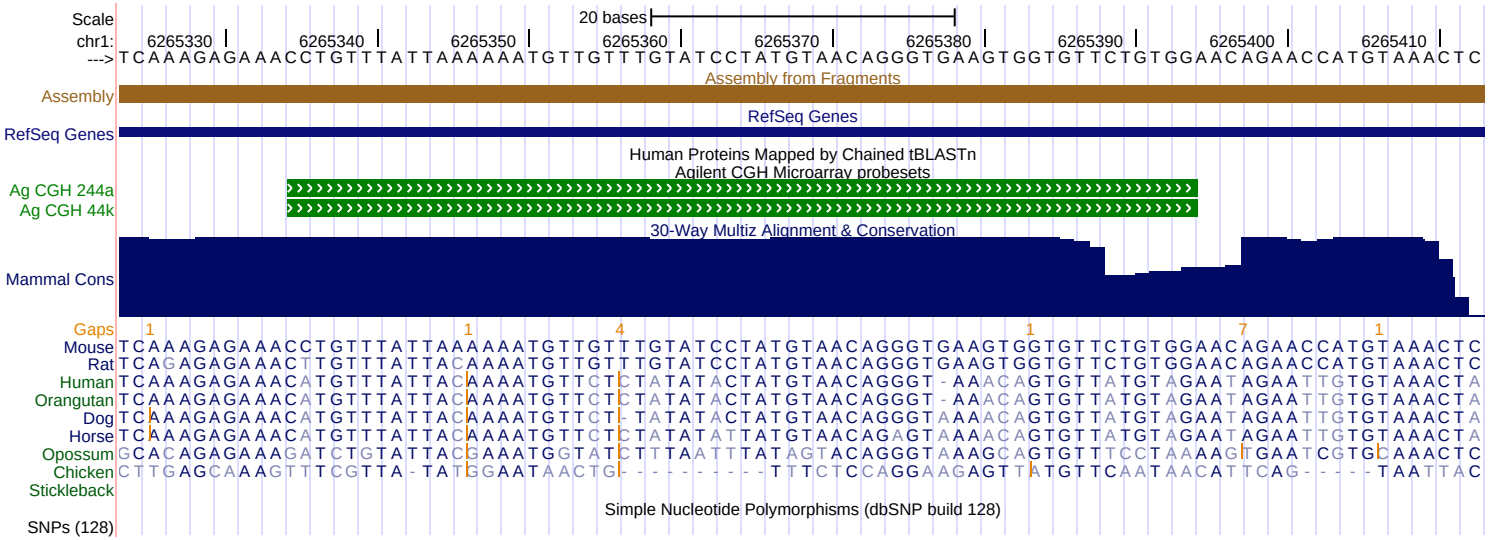
Figure from *Immunobiology.* Janeway, Charles A.; Travers, Paul; Walport, Mark; Shlomchik, Mark. New York and London: Garland Science; 2001.

Outline

Probes on the genome

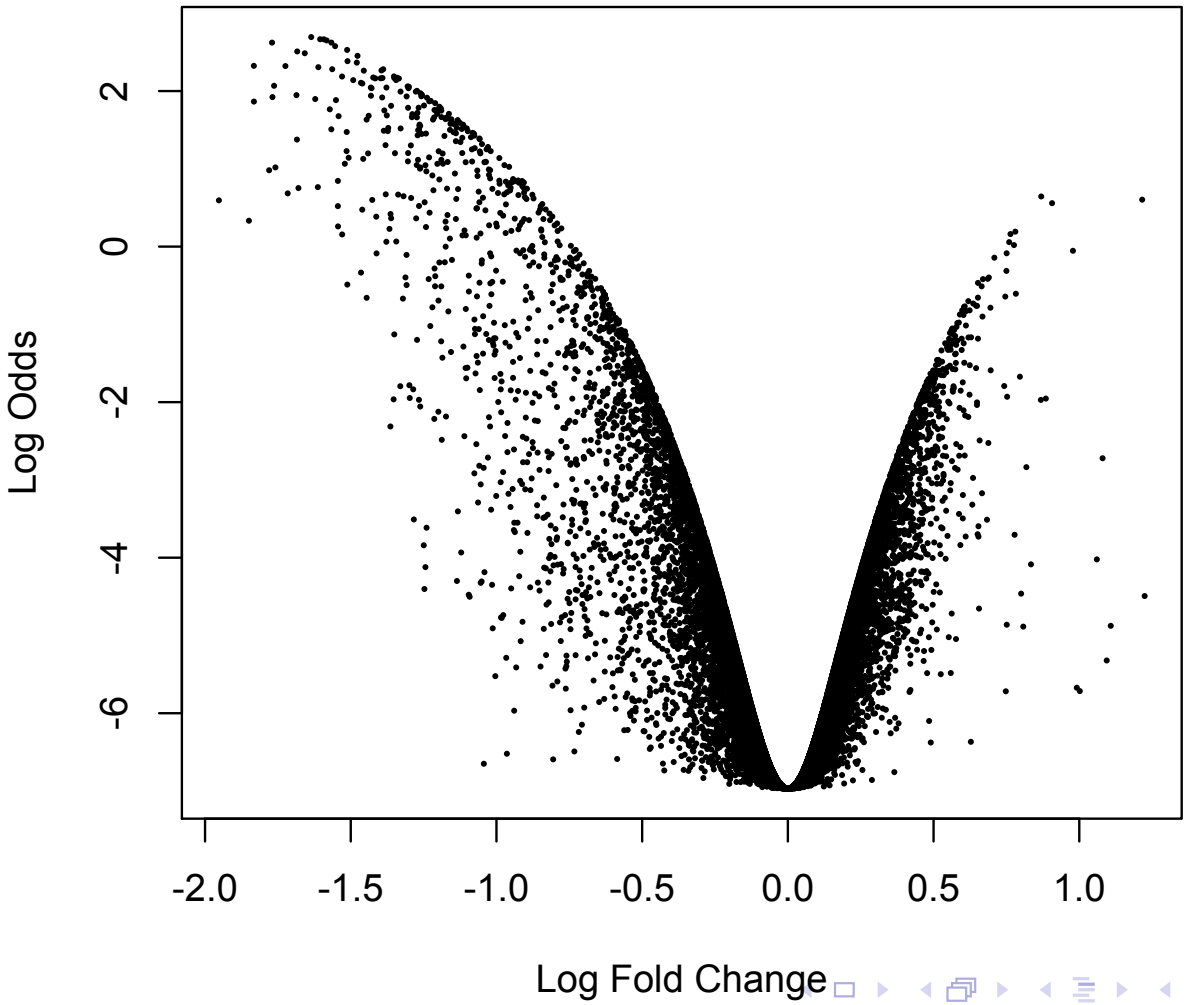


Probes on the sequence

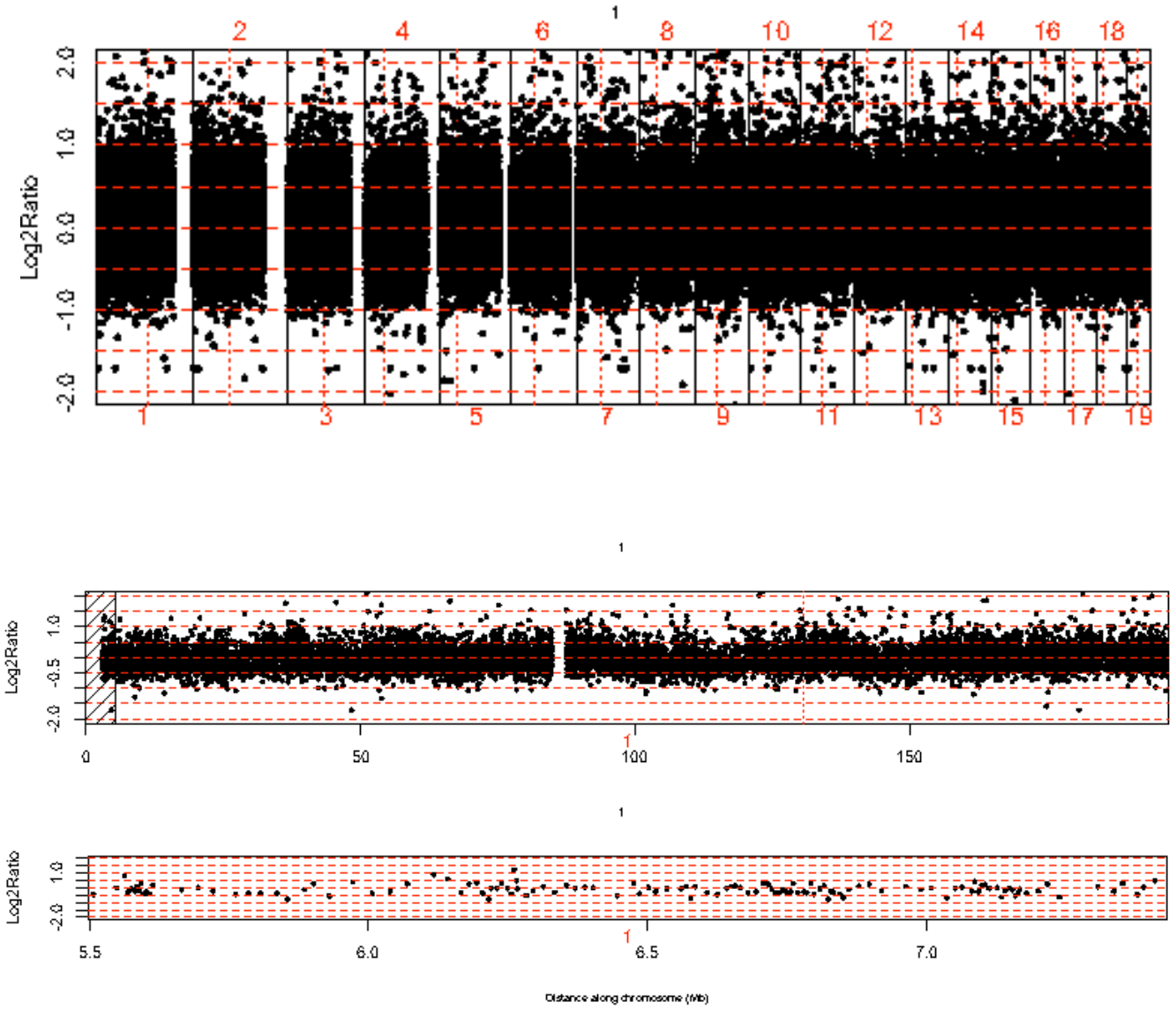


Outline

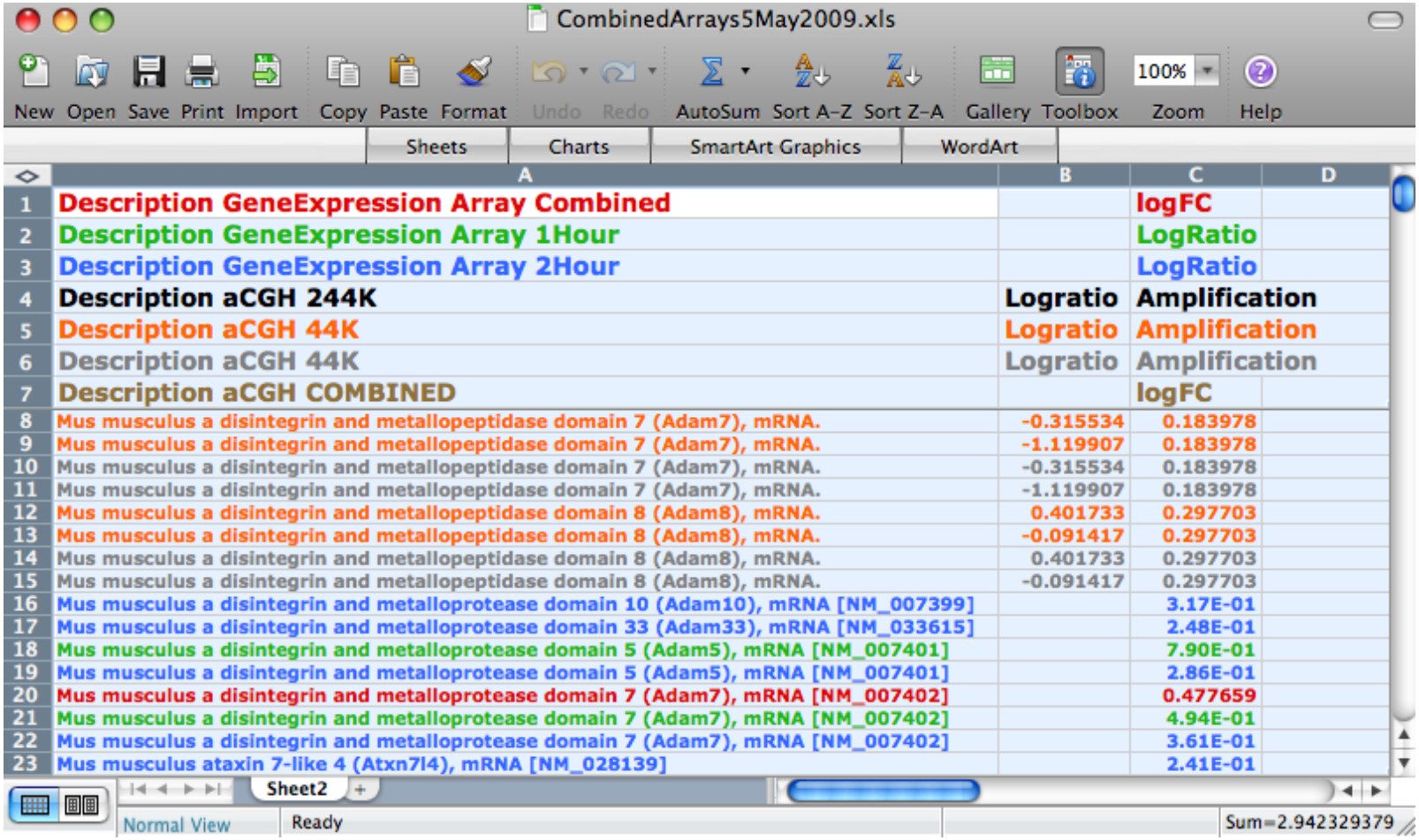
Differentially expressed genes



Genome wide probe abundance



Candidate genes



	A	B	C	D
1	Description GeneExpression Array Combined		logFC	
2	Description GeneExpression Array 1Hour		LogRatio	
3	Description GeneExpression Array 2Hour		LogRatio	
4	Description aCGH 244K	Logratio	Amplification	
5	Description aCGH 44K	Logratio	Amplification	
6	Description aCGH 44K	Logratio	Amplification	
7	Description aCGH COMBINED		logFC	
8	Mus musculus a disintegrin and metallopeptidase domain 7 (Adam7), mRNA.	-0.315534	0.183978	
9	Mus musculus a disintegrin and metallopeptidase domain 7 (Adam7), mRNA.	-1.119907	0.183978	
10	Mus musculus a disintegrin and metallopeptidase domain 7 (Adam7), mRNA.	-0.315534	0.183978	
11	Mus musculus a disintegrin and metallopeptidase domain 7 (Adam7), mRNA.	-1.119907	0.183978	
12	Mus musculus a disintegrin and metallopeptidase domain 8 (Adam8), mRNA.	0.401733	0.297703	
13	Mus musculus a disintegrin and metallopeptidase domain 8 (Adam8), mRNA.	-0.091417	0.297703	
14	Mus musculus a disintegrin and metallopeptidase domain 8 (Adam8), mRNA.	0.401733	0.297703	
15	Mus musculus a disintegrin and metallopeptidase domain 8 (Adam8), mRNA.	-0.091417	0.297703	
16	Mus musculus a disintegrin and metalloprotease domain 10 (Adam10), mRNA [NM_007399]		3.17E-01	
17	Mus musculus a disintegrin and metalloprotease domain 33 (Adam33), mRNA [NM_033615]		2.48E-01	
18	Mus musculus a disintegrin and metalloprotease domain 5 (Adam5), mRNA [NM_007401]		7.90E-01	
19	Mus musculus a disintegrin and metalloprotease domain 5 (Adam5), mRNA [NM_007401]		2.86E-01	
20	Mus musculus a disintegrin and metalloprotease domain 7 (Adam7), mRNA [NM_007402]		0.477659	
21	Mus musculus a disintegrin and metalloprotease domain 7 (Adam7), mRNA [NM_007402]		4.94E-01	
22	Mus musculus a disintegrin and metalloprotease domain 7 (Adam7), mRNA [NM_007402]		3.61E-01	
23	Mus musculus ataxin 7-like 4 (Atxn7l4), mRNA [NM_028139]		2.41E-01	

Outline

Summary

- Microarray probe counts, data, and experiments continue to grow at exponential rates.
- Sequence databases also continue to grow at exponential rates.
- This growth presents a challenge, and opportunity to researchers.
- Combining multiple sources of information is possible, but requires new analysis and visualization techniques to be effective.

Acknowledgments

This work was supported in part by a SCORE grant (SC1MH086072) and an INBRE grant (P20RR016470) from the National Institutes of Health.
The data was kindly provided by Dr. Sandra Peña de Ortiz and Edgardo Castro.

Fin