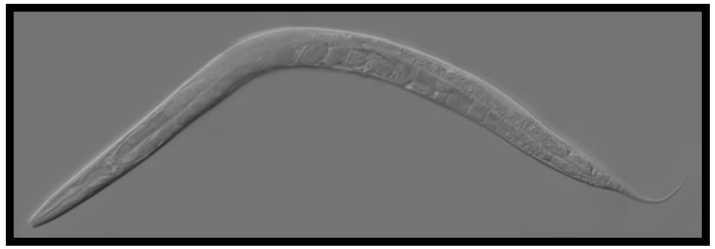
Estimation of Transcript Expression Levels

Tracy Holsclaw
James Ireland
June 6, 2007

What explains complexity?

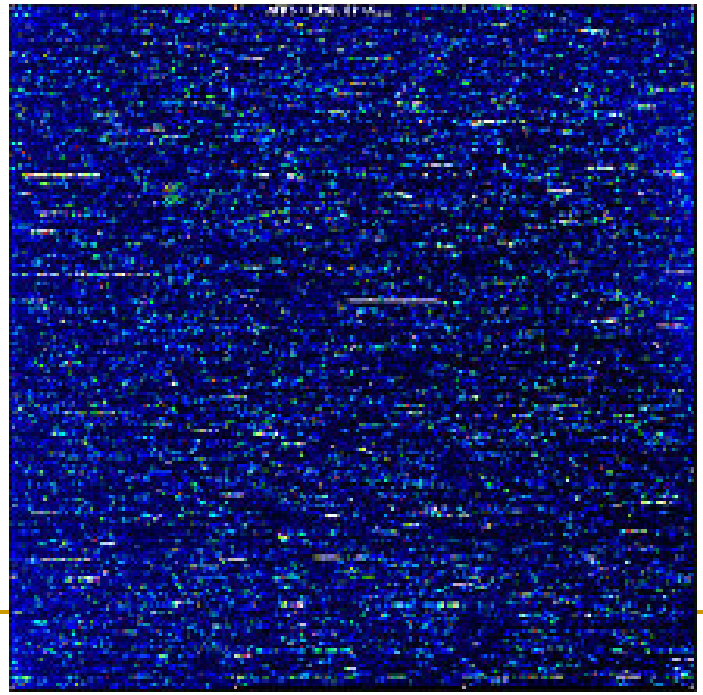


Worm: ~20K Genes

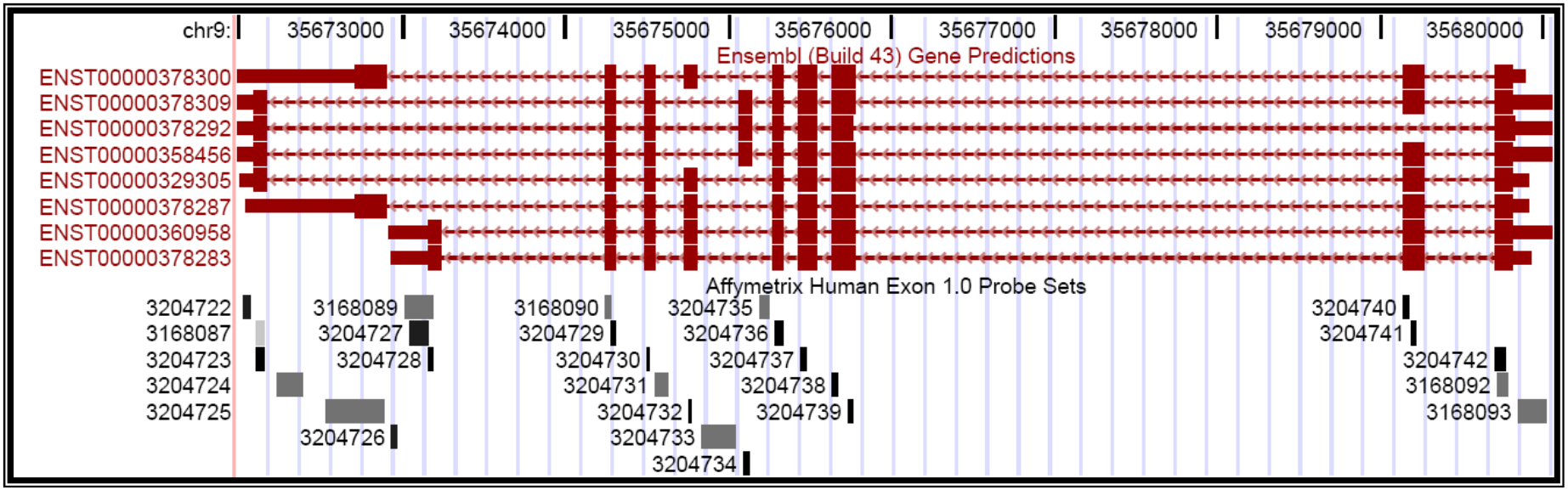
Human: ~24K Genes

Fly: ~14K Genes

Alternative Exon Splicing



Example Gene: TPM2



SampleId	Tissue Type
1	breast
2	cerebellum
3	muscle
4	liver

Matrix Format for the Equation

$$y_i = AGT + e_i$$

$$y_{17x4} = \begin{bmatrix} 25.5 & 36.8 & 17.3 & 32.3 \\ 32.7 & 43.5 & 22.6 & 34.7 \\ \vdots & \vdots & \vdots & \vdots \\ 37.0 & 39.8 & 43.9 & 45.2 \end{bmatrix}$$

$$A_{17x17} = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_{17} \end{bmatrix}$$

$$G_{17x3} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$T_{3x4} = \begin{bmatrix} t_1 & t_4 & t_7 & t_{10} \\ t_2 & t_5 & t_8 & t_{11} \\ t_3 & t_6 & t_9 & t_{12} \end{bmatrix}$$

$$\mathbf{y}_i \sim MVN(\mathbf{AGT}, \mathbf{I}\sigma^2)$$

Priors: assuming here that the RVs are independent:

$$p(a_1, a_2, ..., a_{17}) = p(a_1) * p(a_2) * ... * p(a_{17})$$

$$p(t_1, t_2, ..., t_{12}) = p(t_1) * p(t_2) * ... * p(t_{12})$$

Informative priors: $p(a_i) = N(3\ ,\ 2)$

Non-Informative priors: $p(t_i) \propto 1$

Non-informative prior: $p(\sigma^2) \propto \frac{1}{\sigma^2}$

Posterior:

$$\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^{17}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{AGT})'(\mathbf{y}-\mathbf{AGT})\right\} * p(a_1) * p(a_2) * ... * p(a_{17}) * p(t_1) * p(t_2) * ... * p(t_{12}) * p(\sigma^2)$$

MCMC algorithm:

Posterior:

$$= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^{17}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{AGT})'(\mathbf{y}-\mathbf{AGT})\right\} * p(a_1) * p(a_2) * \ldots * p(a_{17}) * p(t_1) * p(t_2) * \ldots * p(t_{12}) * p(\sigma^2)$$

A Gibb's step is only possible for σ^2 , where n =14*7 and k =17+12:

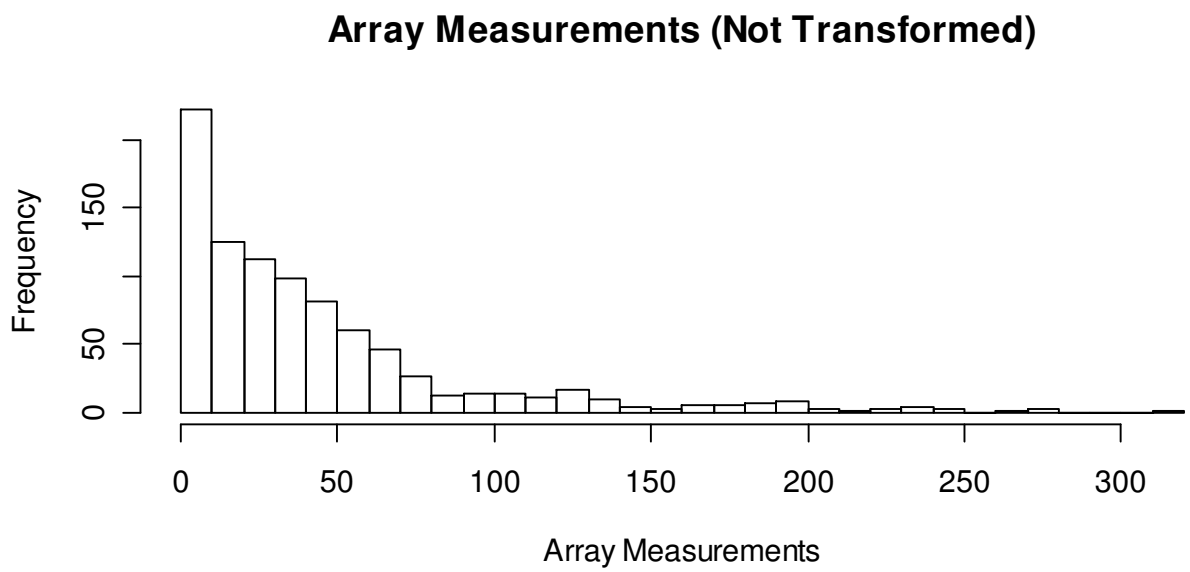
$$\sigma^2 \mid a_i, t_i, y \sim Inv\chi^2(n-k)$$

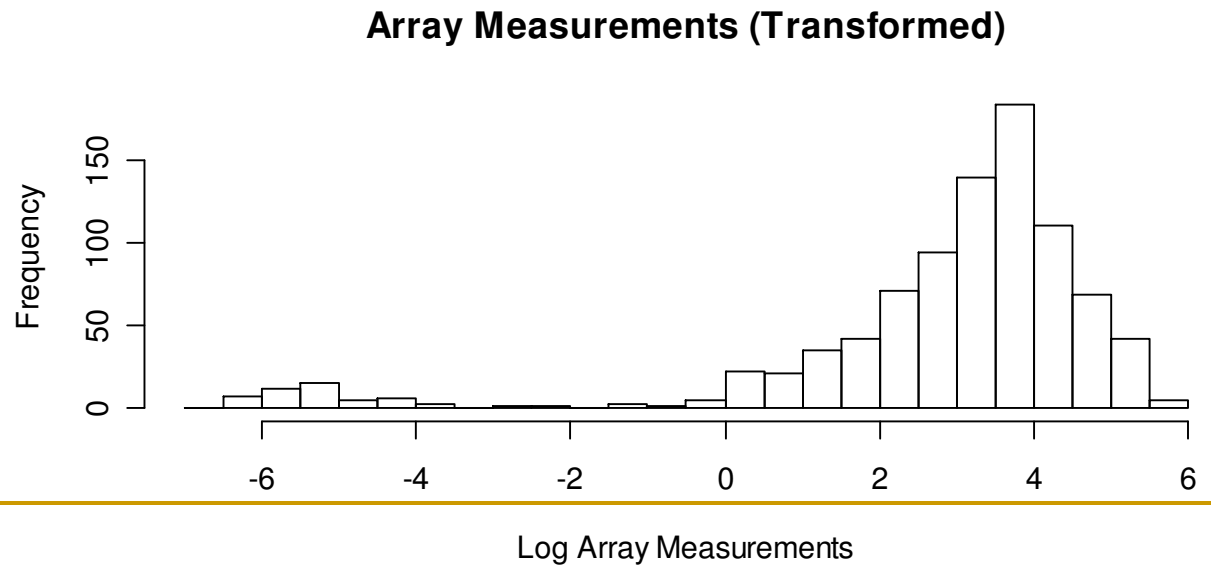
A Metropolis-Hasting step was used for all the 17 a-values as a group.
The proposal distribution was a Uniform so that we could ensure the a values would all be positive.

$$a_{i,t} \mid \sigma_t^2, t_{i,t-1}, y$$

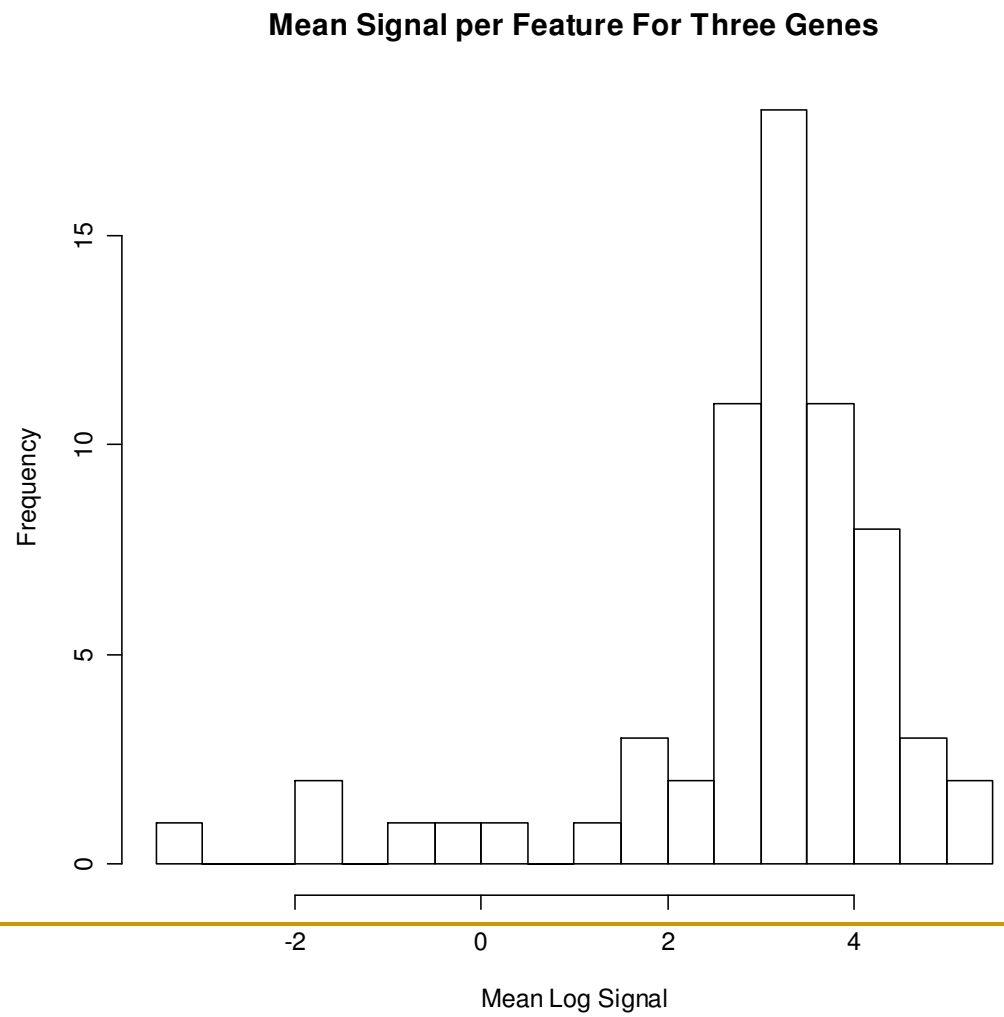
A Metropolis-Hasting step was used for all the 12 t-values as a group.
The proposal distribution was a Uniform so that we could ensure the a values would all be positive.

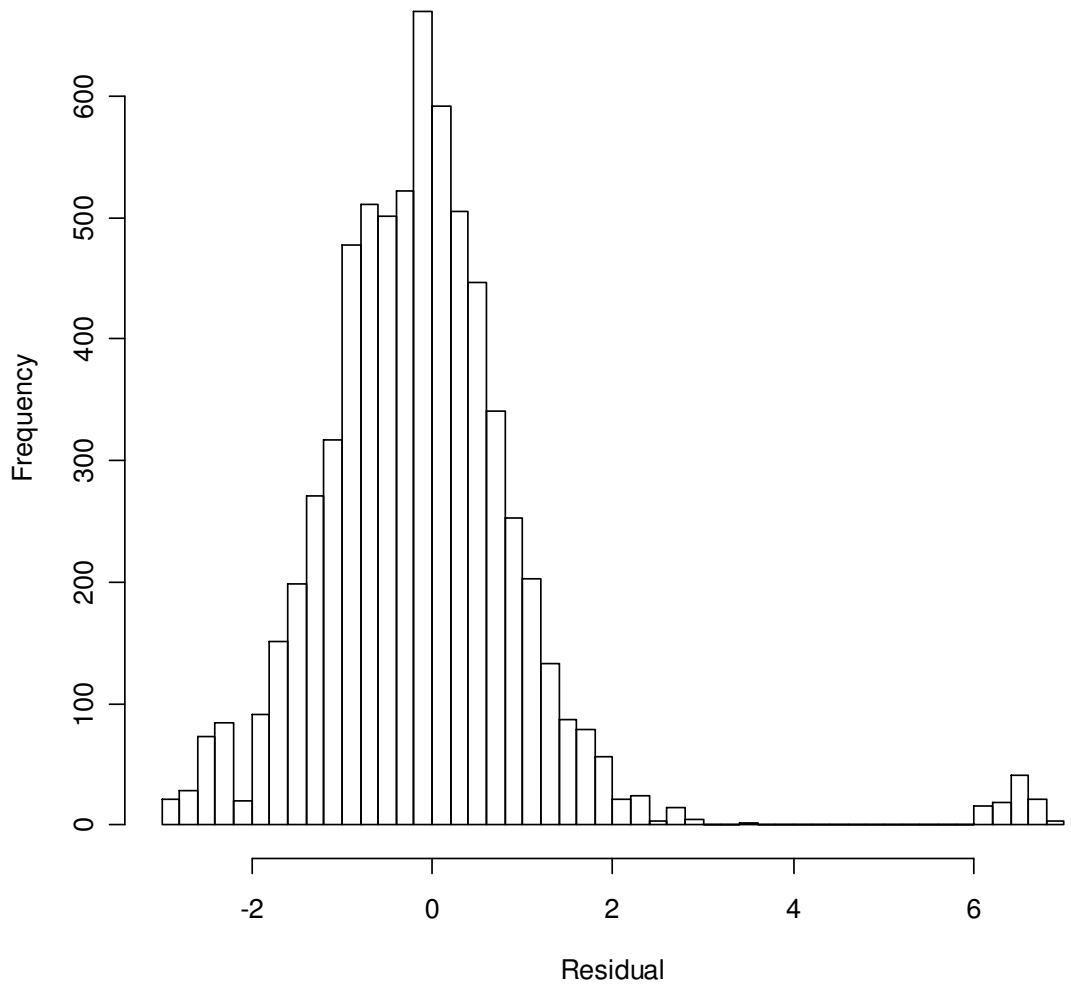
$$t_{i,t} \mid \sigma_t^2, a_{i,t}, y$$

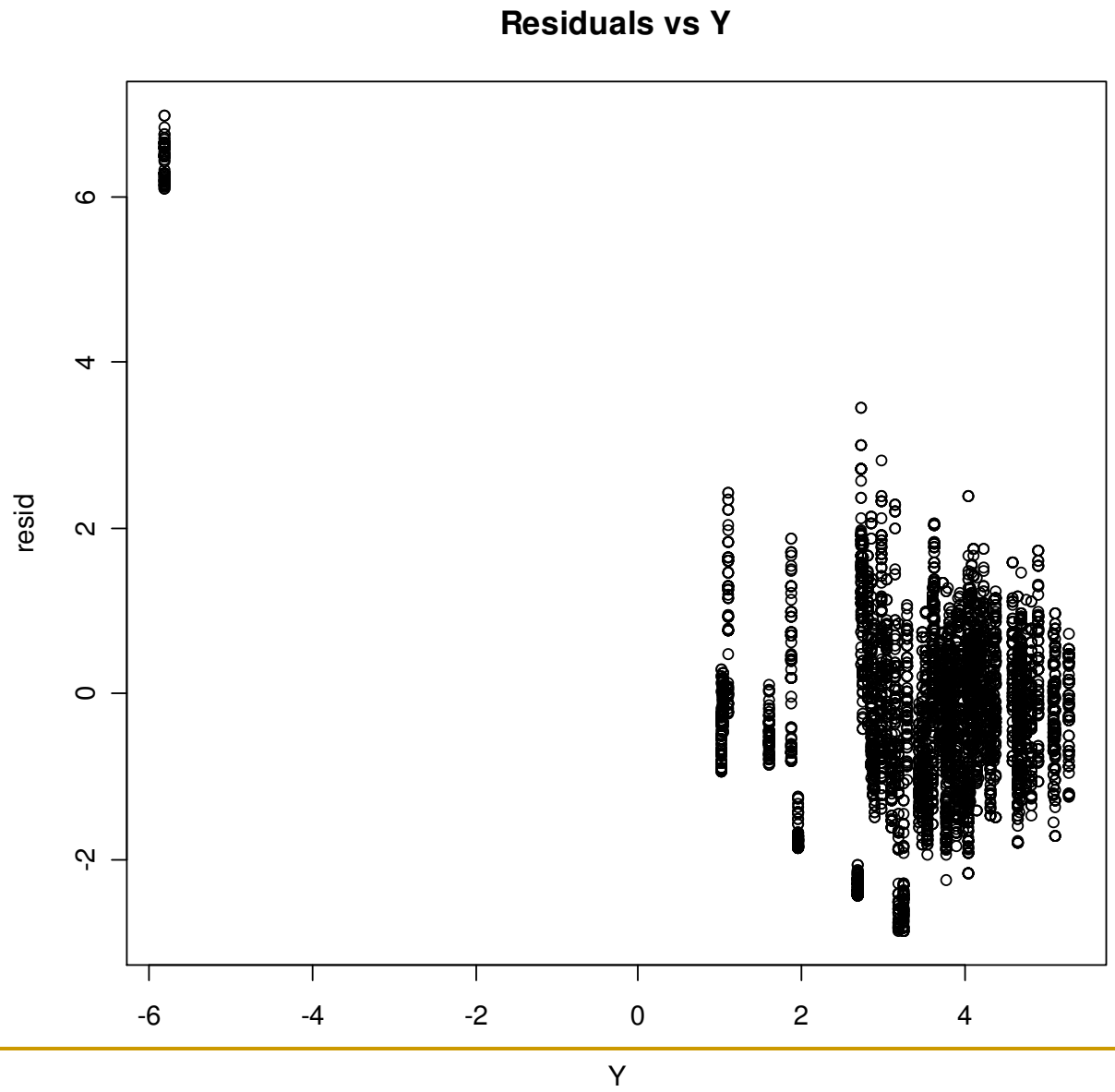

Array Measurements (Not Transformed)
Frequency
0
50
150
0
50
100
150
200
250
300
Array Measurements

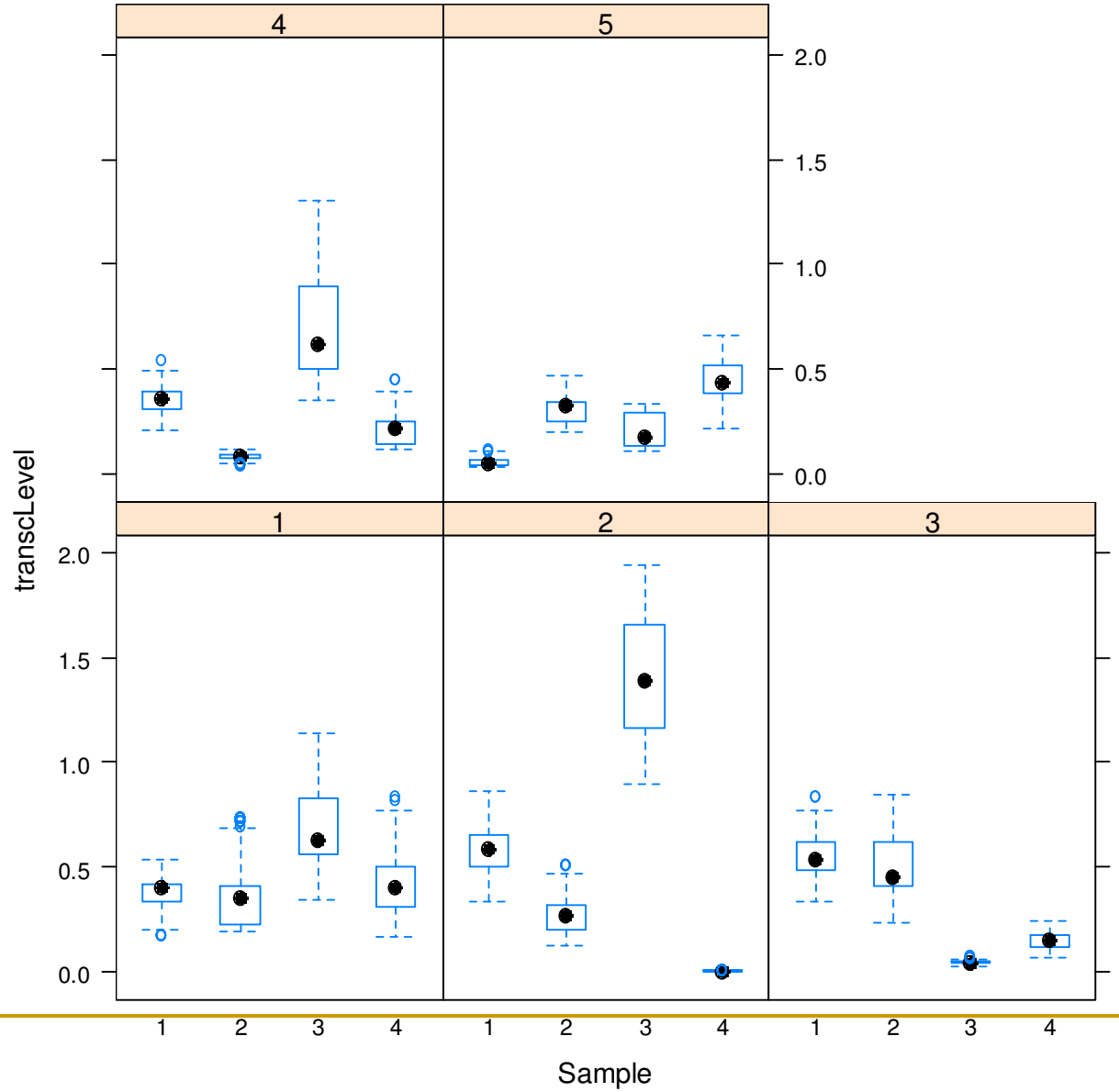

Array Measurements (Transformed)
Frequency
0
50
100
150
-6
-4
-2
0
2
4
6
Log Array Measurements

Informative Prior for A Matrix



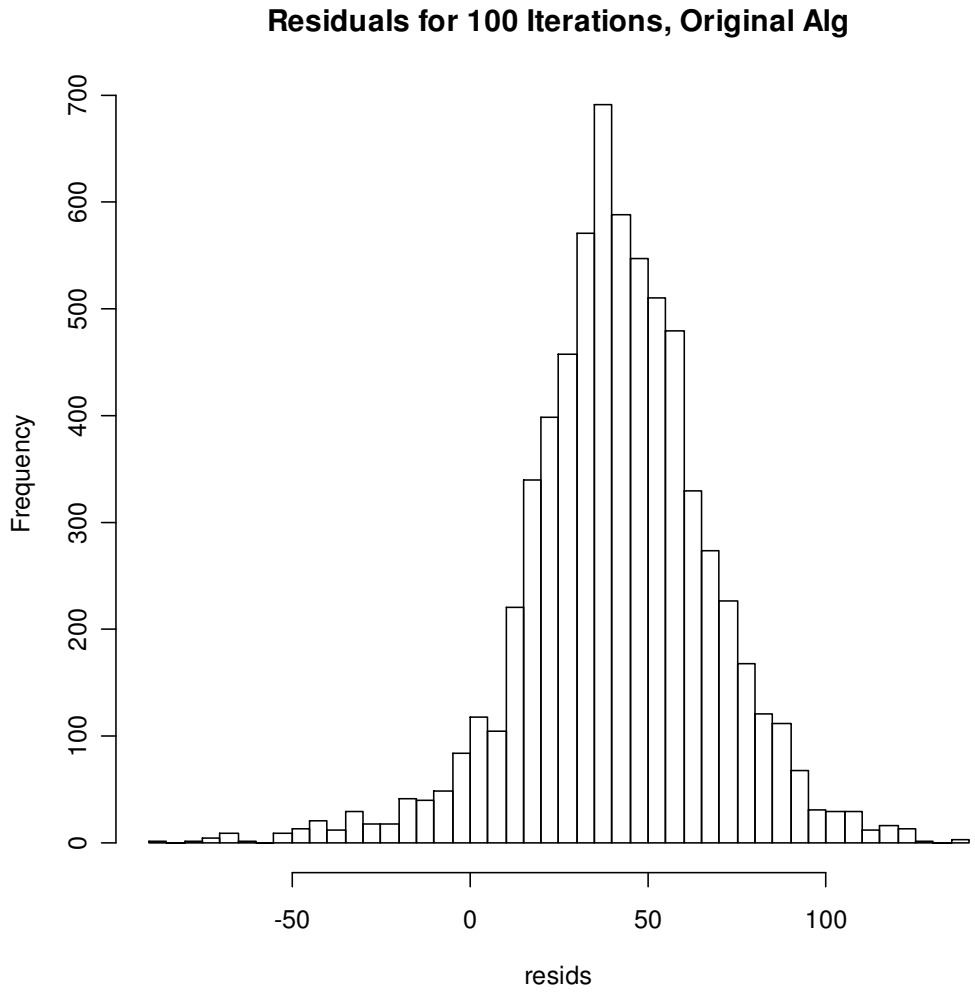

Distribution of Residuals
Frequency
0
100
200
300
400
500
600
-2
0
2
4
6
Residual

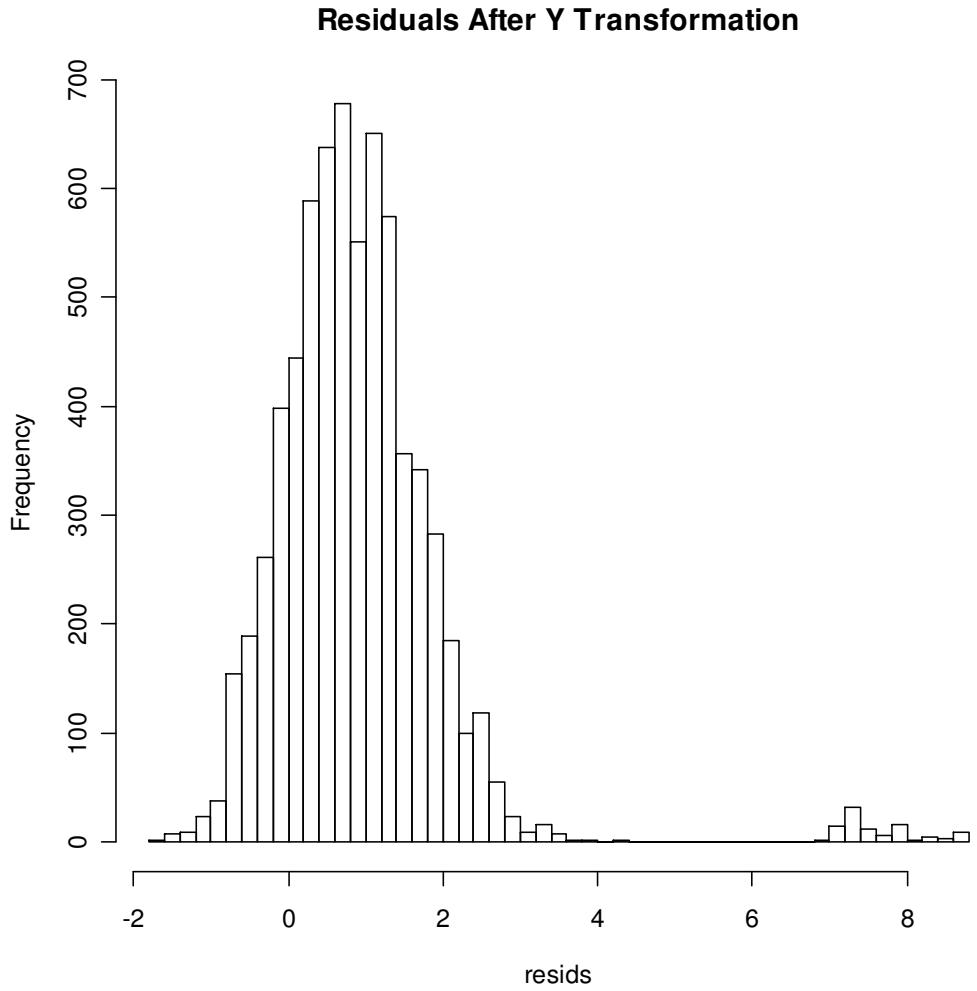

Residuals vs Y
resid
6
4
2
0
-2
-6
-4
-2
0
2
4
Y

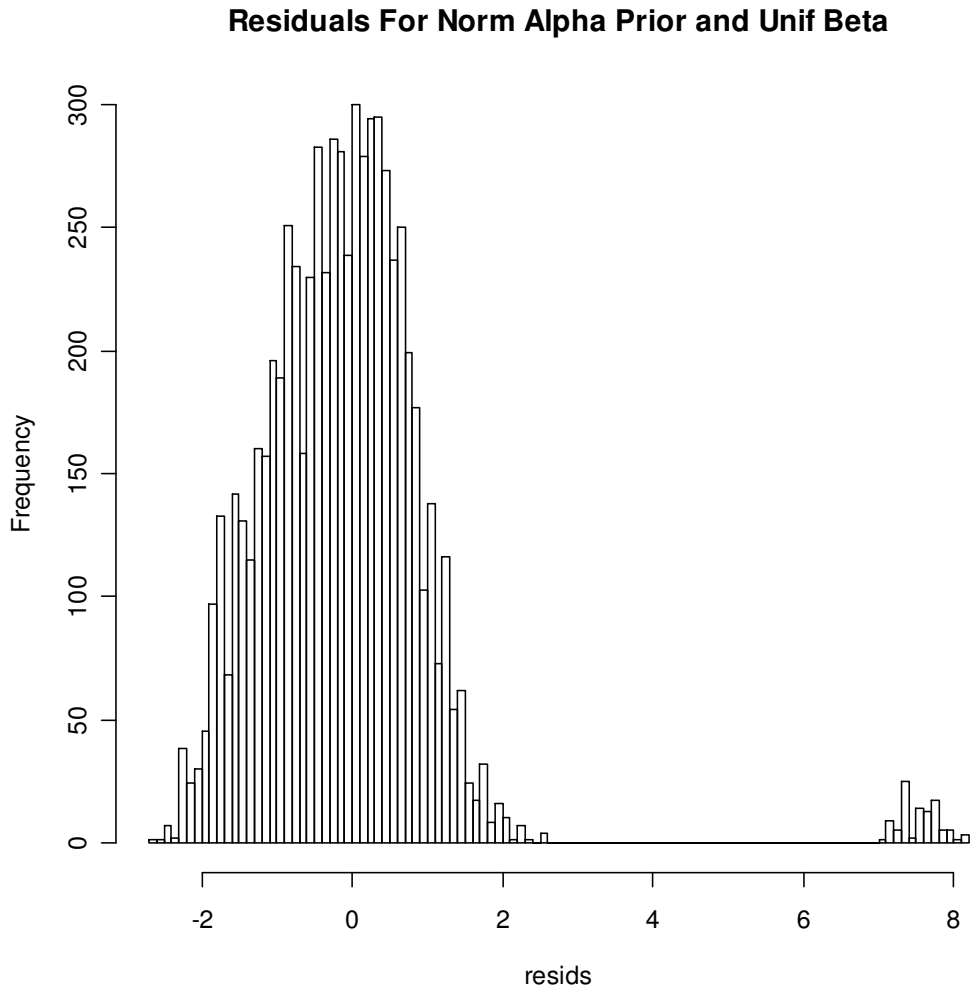

Estimated Transcript Levels per Sample
4
5
1
2
3
transcLevel
0.0
0.5
1.0
1.5
2.0
1
2
3
4
Sample

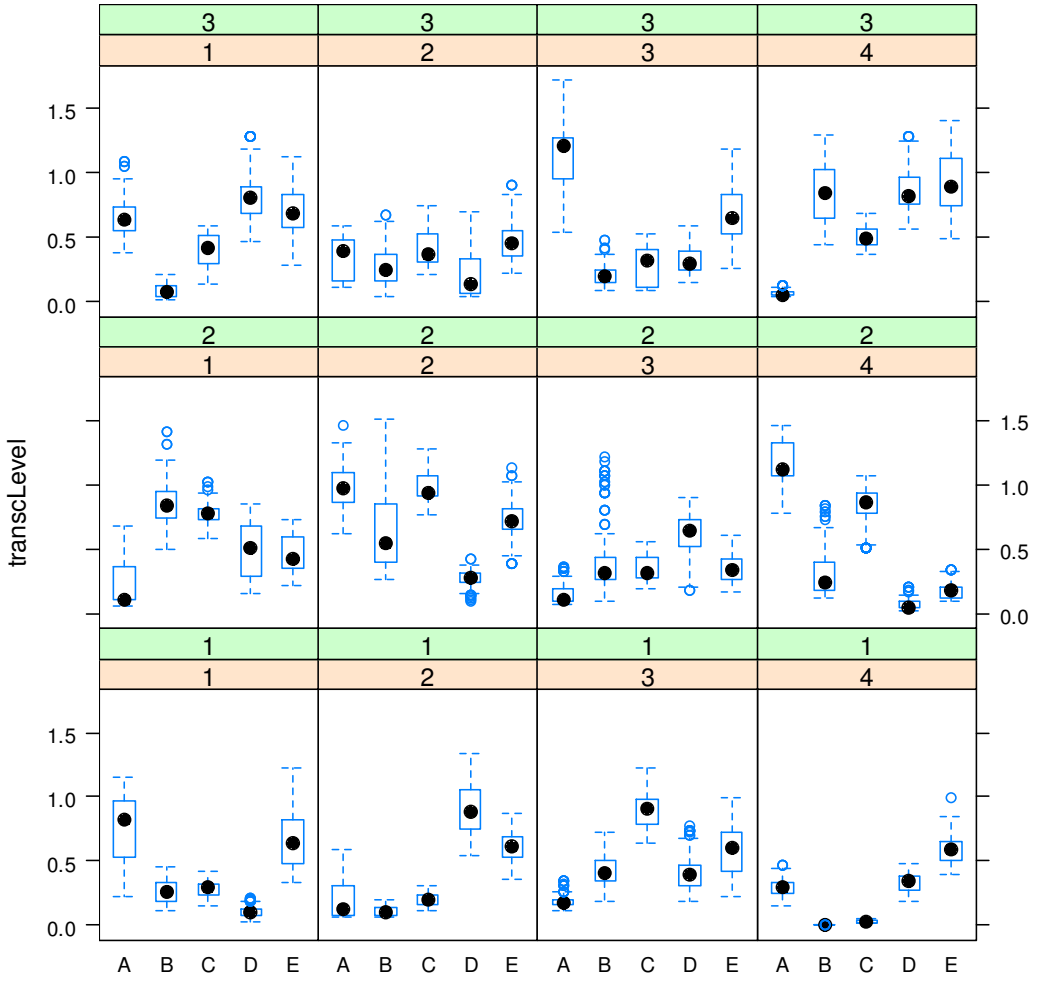
Further Work

- Mixture Model
- Better Estimation of sigma
- Work with more genes + sample types


Residuals for 100 Iterations, Original Alg
Frequency
0
100
200
300
400
500
600
700
-50
0
50
100
resids


Residuals After Y Transformation
Frequency
0
100
200
300
400
500
600
700
-2
0
2
4
6
8
resids


Residuals For Norm Alpha Prior and Unif Beta
Frequency
0
50
100
150
200
250
300
-2
0
2
4
6
8
resids


3
3
3
3
1
2
3
4
2
2
2
2
1
2
3
4
1
1
1
1
1
2
3
4
transcLevel
0.0
0.5
1.0
1.5
A
B
C
D
E

Test Gene: ACHE