# who was Bayes?

- Reverend Thomas Bayes (1702-1761)
  - part-time mathematician
  - buried in Bunhill Cemetary, Moongate, London
  - famous paper in 1763 *Phil Trans Roy Soc London*
  - was Bayes the first with this idea? (Laplace?)
- basic idea (from Bayes' original example)
  - two billiard balls tossed at random (uniform) on table
  - where is first ball if the second is to its left (right)?

$\theta$

first

second

$Y=0$   $Y=1$

prior      $\text{pr}(\theta) = 1$

likelihood $\text{pr}(Y \mid \theta) = \theta^{1-Y}(1-\theta)^{Y}$

posterior  $\text{pr}(\theta \mid Y) = ?$

---
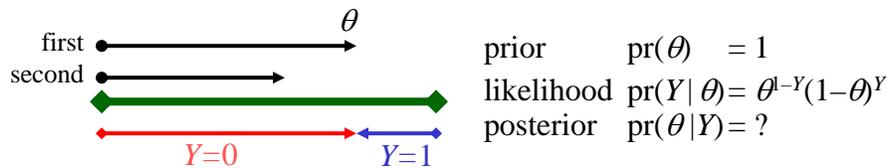
# what is Bayes theorem?

- posterior = likelihood * prior / C

  pr( parameter | data ) =
     pr( data | parameter ) * pr( parameter ) / pr( data )

$\theta$

$Y=0$   $Y=1$

$$\text{pr}(\theta \mid Y) = \frac{\text{pr}(\theta, Y)}{\text{pr}(Y)} = \frac{\text{pr}(Y \mid \theta) \times \text{pr}(\theta)}{\text{pr}(Y)}$$

- prior: probability of parameter before observing data
  - pr( $\theta$ ) = pr( parameter )
  - equal chance of first ball being anywhere on the table
- posterior: probability of parameter after observing data
  - pr( $\theta \mid Y$ ) = pr( parameter | data )
  - more likely second to left if first is near right end of table
- likelihood: probability of data given parameters
  - pr( $Y \mid \theta$ ) = pr( data | parameter )
  - basis for classical statistical inference about $\theta$ given $Y$
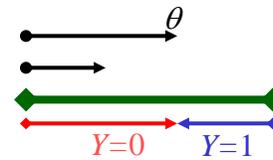
# Bayes posterior for normal data

n small prior

n small

n large

actual mean

prior mean

y = phenotype values

6   8   10   12   14   16

**small prior variance**

n small prior

n large

actual mean

prior mean

y = phenotype values

6   8   10   12   14   16

**large prior variance**

---
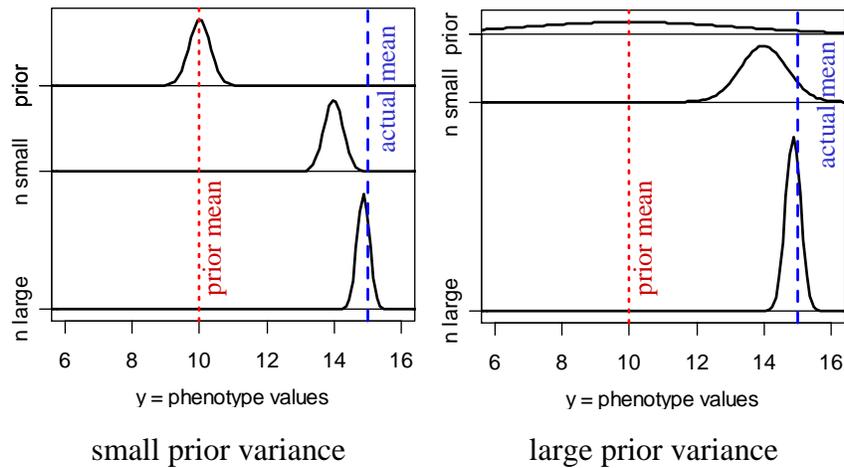
# Bayes posterior for normal data

| | |
|---|---|
| model | $Y_i = \mu + E_i$ |
| environment | $E \sim N(0, \sigma^2)$, $\sigma^2$ known |
| likelihood | $Y \sim N(\mu, \sigma^2)$ |
| prior | $\mu \sim N(\mu_0, \kappa\sigma^2)$, $\kappa$ known |

| | |
|---|---|
| posterior: | mean tends to sample mean |
| single individual | $\mu \sim N(\mu_0 + B_1(Y_1 - \mu_0), B_1\sigma^2)$ |

sample of $n$ individuals $\quad \mu \sim N\left(B_n\overline{Y}_\bullet + (1 - B_n)\mu_0, B_n\sigma^2/n\right)$

$$\text{with } \overline{Y}_\bullet = \sum_{\{i=1,\dots,n\}} Y_i / n$$

fudge factor
(shrinks to 1)

$$B_n = \frac{\kappa n}{\kappa n + 1} \to 1$$

# posterior genotypic means $G_q$



BS Yandell © 2005       Plant Microarray Course       5

---
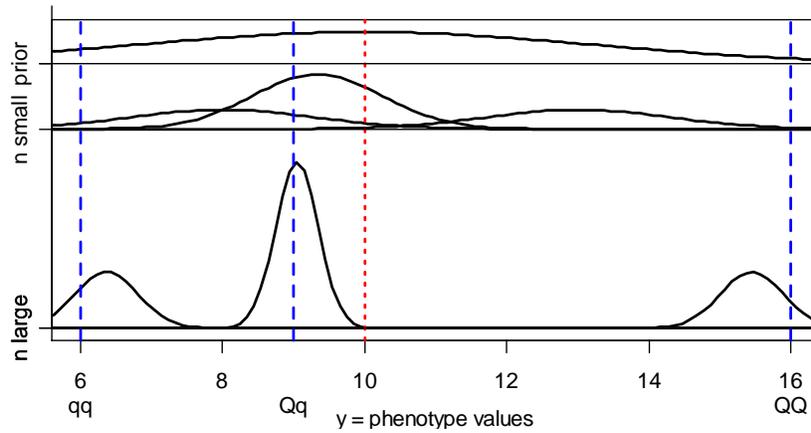
# posterior genotypic means $G_q$

posterior centered on sample genotypic mean
but shrunken slightly toward overall mean

prior:
$$G_q \sim N\left(\overline{Y}_\bullet, \kappa\sigma^2\right)$$

posterior:
$$G_q \sim N\left(B_q \overline{Y}_q + (1 - B_q)\overline{Y}_\bullet, B_q \sigma^2 / n_q\right)$$

$$n_q = \text{count}\{Q_i = q\}, \overline{Y}_q = \sum_{\{Q_i=q\}} Y_i / n_q$$

fudge factor:
$$B_q = \frac{\kappa n_q}{\kappa n_q + 1} \to 1$$

BS Yandell © 2005       Plant Microarray Course       6

# Are strain differences real?

strain differences?
  similar pattern
  parallel lines
  no interaction
noise negligible?
  few d.f. per gene
  Can we trust $SD_g$ ?

---
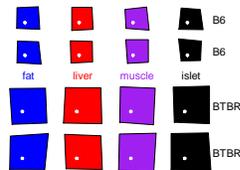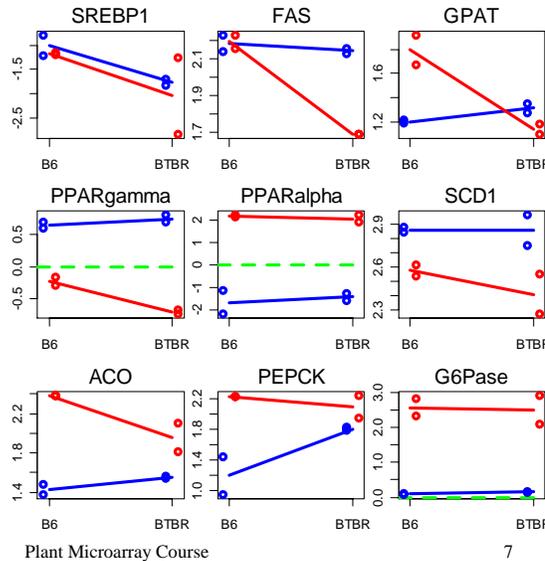
# Bayesian shrinkage of gene-specific SD

- gene-specific SD from replication
  - $SD_g$ = gene-specific standard deviation (df = $\nu_1$)
- robust abundance-based estimate
  - $\sigma(A_g)$ = smoothed over mRNA
  - depends only on abundance level $A_g$ (or constant)
- combine ideas into gene-specific hybrid
  - "prior" $\sigma_g^2 \sim$ inv-$\chi^2(\nu_0, \sigma(A_g)^2)$
  - "posterior" shrinkage estimate
  $$\frac{\nu_1 SD_g^2 + \nu_0 \sigma(A_g)^2}{\nu_1 + \nu_0}$$
  - combines two "statistically independent" estimates

## SD for strain differences

gene-specific $\sigma_g$

smooth of $\sigma_g$

main effects

fat $\sigma(A_g)$

liver $\sigma(A_g)$

interaction

fat-liver $\sigma(A_g)$

B6
B6
fat    liver    muscle    islet
BTBR
BTBR
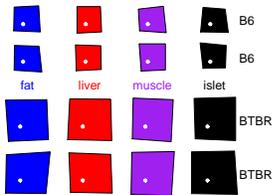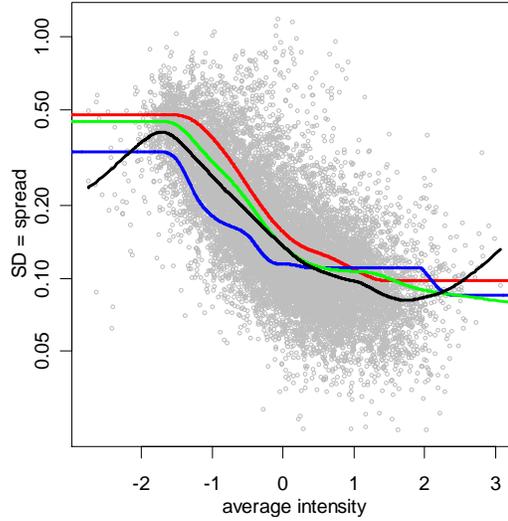
## Shrinkage Estimates of SD

gene-specific $\sigma_g$

abundance $\sigma(A_g)$

95% $\chi_8^2$ limits

new (shrunk) $\sigma_g$

size of shrinkage

$$\frac{\nu_1\sigma_g^2 + \nu_0\sigma(A_g)^2}{\nu_1 + \nu_0}$$

B6
B6
fat    liver    muscle    islet
BTBR
BTBR
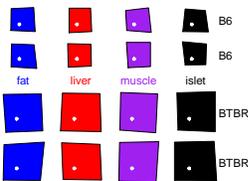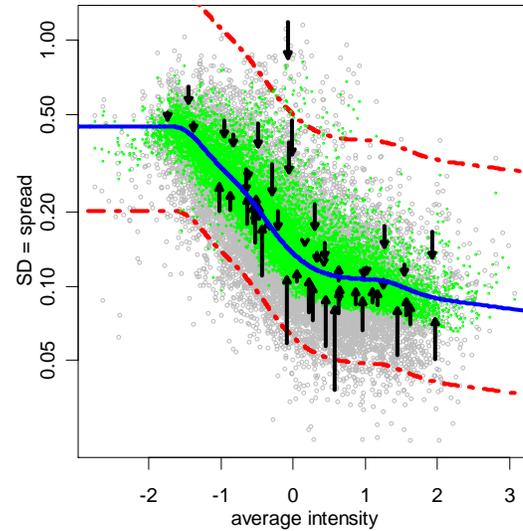
# How good is shrinkage model?

prior for gene-specific

$$\sigma_g^2 \sim \text{inv-}\chi^2(\nu_0, \beta\sigma(A_g)^2)$$

fudge $\beta$ to adjust mean

$$\frac{\nu_1\sigma_g^2 + \nu_0\beta\sigma(A_g)^2}{\nu_1 + \nu_0}$$

histogram of ratio

$$\sigma_g^2 / \sigma(A_g)^2$$

empirical Bayes estimates

$\chi^2$ approximation

$\nu_0 = 5.45, \beta = 1$

$\chi^2$ approximation with $\beta$

$\nu_0 = 3.56, \beta = .809$

---
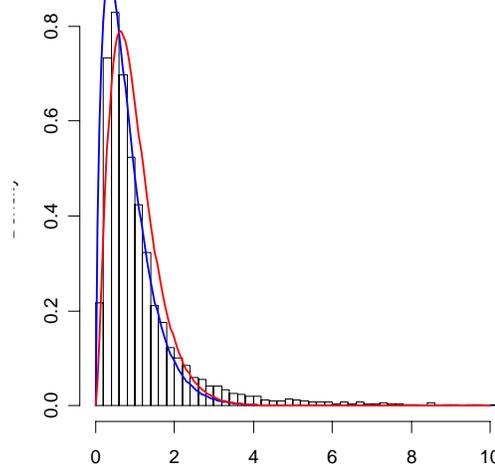
# Effect of SD Shrinkage on Detection

fat-liver interaction

shrinkage-based

abundance-based

9 genes identified



S = (D-center)/spread

A = average intensity

B6
B6
fat   liver   muscle   islet
BTBR
BTBR
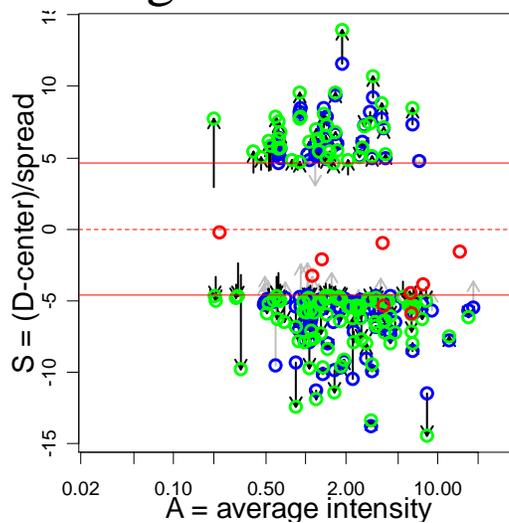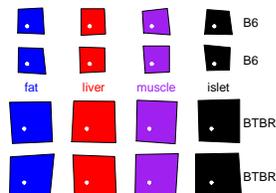
# QTL Mapping (Gary Churchill)

Key Idea:  Crossing two inbred lines creates
linkage disequilibrium which in turn creates
associations and linked segregating QTL



QTL

Marker  - - - - - - - ▸  Trait

---
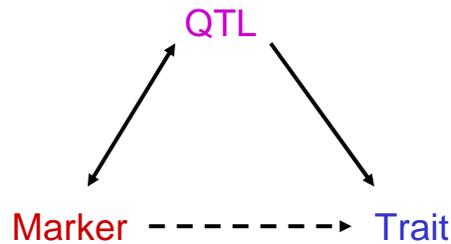
# Bayes factors for comparing models

- goal of BF: balance model fit with model "complexity"
    - want "best model" that captures key features (model bias)
    - want to avoid "overfitting" the data in hand (poor prediction)
- what is a Bayes factor (*BF*)?
    - ratio of posterior odds to prior odds
    - ratio of model likelihoods
- *BF* is same as Bayes Information Criteria (*BIC*)
    - penalty on likelihood ratio (*LR*)
- want Bayes factor to be much larger than 1 (ideally > 10)

$$BF_{12} = \frac{\text{pr( model}_1 \,|\, \text{data)} / \text{pr( model}_2 \,|\, \text{data)}}{\text{pr( model}_1) / \text{pr( model}_2)} = \frac{\text{pr(data} \,|\, \text{model}_1)}{\text{pr(data} \,|\, \text{model}_2)}$$

$$-2\log(BF_{12}) = -2\log(LR) - (p_2 - p_1)\log(n)$$