Hopkins Statistical Genetics Working Group

• **Goals:**
  – Get to know each other
  – Foster collaboration
  – See new stuff
  – Get help

• **Scope:**
  statistical, computational, mathematical issues in genetics, genomics, molecular biology

• **Format:**
  – interesting papers, problems, works-in-progress
  – informal and interactive

• **Time:**
  – Mondays, 3:30pm
  – biweekly (next meeting 16 Oct)

• **Webpage:**
  biosun01.biostat.jhsph.edu/~kbroman/hsgwg
Identifying QTLs in experimental crosses

Karl W Broman

Humans vs model organisms

“Model”:  
• Genes  
• Genetic architecture  
• Analysis methods

Differences:  
• Complexity of pedigrees  
• Complexity of genet. architec.  
• Environmental variation
Backcross experiment

LL × HH → HL

Backcross progeny (LL or HL)
Data

Phenotypes (trait values)

\[ y_i = \text{phenotype for individual } i \]

Marker genotypes

\[ x_{ij} = 1/0 \text{ if } i \text{ is HL/LL at marker } j \]

Genetic map

Locations of markers

Models

Recombination:  No interference

Phenotype/genotype connection

\[ y = \mu + \sum \beta_j z_j + \epsilon \]

\[ \epsilon \sim \text{Normal}(0, \sigma^2) \]
Problem

100 to 1000 backcross progeny
100 to 400 markers
\[ y = \mu + \sum \beta_j x_j + \varepsilon \]

Find the \( x \)'s with \( \beta_j \neq 0 \)

Errors:

• Miss important loci
• Include extraneous loci
The usual method

At each location:

- Imagine a single QTL
- Infer genotypes
- Regression of phenotype on genotype
Major issues

- Missing genotype information
- Model selection
- Estimation of QTL effects
- Estimation of QTL location

Model Selection:

- Space of models
- Searching through models
- Comparing models
- Assessing performance


Ásaunak Sen and Gary Churchill

A statistical framework for quantitative trait mapping

http://www.jax.org/research/churchill

\[
y = \text{phenotypes} \\
m = \text{marker genotypes} \\
H = \text{genetic model} \\
g = \text{QTL genotypes} \\
\mu = \text{parameters of genetic model} \\
\theta = \text{locations of QTLs}
\]

\[
p_H(y, m, g, \mu, \theta) = \{ p_H(y | g, \mu) p_H(\mu) \} \\
\quad \{ p(g | m, \theta) p(m) p(\theta) \}
\]
Multiple imputation

Grid G of "pseudomarker" positions

(e.g., equally spaced at 2 cM over entire genome)

\( r_i(u) = \text{realization of matrix of genotypes simulated from } p(g \mid m, \theta = G) \)

- \( u = \text{locations on grid} \)
- \( i = 1 \ldots q \text{ realizations} \)

Weights:

\[ W_H[r_i(u)] = p_H[y \mid g = r_i(u)] p(\theta = u) \]

- integrate over \( \mu \)
- \( p(\theta = u) \text{ may be ignored} \)

e.g., Normal model (\( v = \# \text{ QTLs} \)):

\[-2 \log W_H[r_i(u)] = v \log n + n \log \text{RSS} \]
Estimating QTL locations

\[ p_H(\theta = u \mid y, m) \propto \sum_i W_H[r_i(u)] \]

Estimating QTL model parameters

\[ p_H(\mu \mid y, m) \propto \sum_i \sum_u p_H[\mu \mid y, g=r_i(u)] W_H[r_i(u)] \]

Model selection

Bayes factor \( B(H,K) = p_H(y,m) / p_K(y,m) \)

\[ p_H(y \mid m) \approx \sum_i \sum_u W_H[r_i(u)] / (q \cdot s) \]

- \( q = \) number of realizations
- \( s = \) number of \( p \)-tuples of qtl loc

General approach:

Main-scan
Pair-scan for interactions
Main scan
Pair scan
Pair scan: closeup view