Inferring Genetic Architecture of Complex Biological Processes
Brian S. Yandell\(^1\), Christina Kendziorski\(^2\), Hong Lan\(^3\), Elias Chaibub\(^4\), Alan D. Attie\(^5\)
1 Department of Statistics
2 Department of Horticulture
3 Department of Biostatistics & Medical Informatics
4 Department of Biochemistry
University of Wisconsin-Madison
http://www.stat.wisc.edu/~yandell/statgen

---

studying diabetes in an F2

- mouse model: segregating panel from inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (Chr 6)
  - sacrificed at 14 weeks, tissues preserved
- physiological study (Stoehr et al. 2000 *Diabetes*)
  - mapped body weight, insulin, glucose at various ages
- gene expression studies
  - RT-PCR for a few mRNA on 108 F2 mice liver tissues
  - Affymetrix microarrays on 60 F2 mice liver tissues
  - U47 A & B chips, RMA normalization
  - design: selective phenotyping (Jin et al. 2004 *Genetics*)

---

The intercross (from K Broman)

mRNA expression as phenotype:
interval mapping for SCD1 is complicated

taking a multiple QTL approach

- improve statistical power, precision
  - increase number of QTL detected
  - better estimates of loci: less bias, smaller intervals
- improve inference of complex genetic architecture
  - patterns and individual elements of epistasis
  - appropriate estimates of means, variances, covariances
    - asymptotically unbiased, efficient
  - assess relative contributions of different QTL
- improve estimates of genotypic values
  - less bias (more accurate) and smaller variance (more precise)
  - mean squared error = MSE = (bias)\(^2\) + variance

---

[Image of glucose and insulin levels, courtesy AD Attie]
Pareto diagram of QTL effects

Bayesian model assessment: number of QTL for SCD1 with R/bim

Bayesian model assessment genetic architecture: chromosome pattern

trans-acting QTL for SCD1

Bayesian LOD and $h^2$ for SCD1 (summaries from R/bim)

SCD mRNA expression phenotype 2-D scan for QTL (R/qtl)
sub-peaks can be easily overlooked

interval mapping basics
- observed measurements
  - \( Y \): phenotype trait
  - \( X \): markers & linkage map
- missing data
  - missing marker data
  - \( Q \): QTL genotypes
    - \( Q_1, Q_2, Q_3 \)
- unknown quantities
  - \( M \): genetic architecture
    - \( \lambda \): QT locus (or loci)
  - \( \mu \): phenotype model parameters
  - \( f(Y|\mu) \): phenotype model
    - grounded by linkage map, experimental errors
      - recombination yields multinomial for \( Q \) given \( X \)
- \( f(Y|X,M) \): phenotype model
- \( \beta_0 + \sum_{q \in M} \beta_q \gamma_q \): genotypic mean for \( Q \) given \( X \)

multiple QTL interval mapping
- genotypic mean depends on model \( M \)
  - \( \mu_q = \beta_0 + \sum_{q \in M} \beta_q \gamma_q \)
- interval mapping between flanking markers
  - \( f(Y|X,M) = \sum_q f(Y|\mu_q) f(Q = q|X,\lambda) \)
- model selection
  - choice of distribution: \( f \) is normal
  - sample many possible architectures
  - compare based on Bayes factors (BIC)

heterogeneity: many genes affect each trait
- major QTL on linkage map
  - polygenes

3 November 2004  UAB: Yandell © 2004 15

3 November 2004  UAB: Yandell © 2004 14

3 November 2004  UAB: Yandell © 2004 18

2M observations
30,000 traits
60 mice
modern high throughput biology

- measuring the molecular dogma of biology
  - DNA → RNA → protein → metabolites
  - measured one at a time only a few years ago
- massive array of measurements on whole systems ("omics")
  - thousands measured per individual (experimental unit)
  - all (or most) components of system measured simultaneously
  - whole genome of DNA, genes, promoters, etc.
  - all expressed RNA in a tissue or cell
  - all proteins
  - all metabolites

- systems biology: focus on network interconnections
  - chains of behavior in ecological community
  - underlying biochemical pathways
- genetics as one experimental tool
  - perturb system by creating new experimental cross
  - each individual is a unique mosaic

finding heritable traits
(from Christina Kendziorski)

- reduce 30,000 traits to 300-3,000 heritable traits
- probability a trait is heritable
  \[ p(H|Y,Q) = p(Y|Q,H) p(H|Q) / p(Y|Q) \] Bayes rule
\[ p(Y|Q,H) = p(Y|Q,H) p(H|Q) + p(Y|Q,not H) p(not H|Q) \]
- phenotype averaged over genotypic mean \( \mu \)
  \[ p(Y|Q,not H) = f_0(Y) = \int f(Y|\mu) p(\mu) d\mu \]
  if not heritable
  \[ p(Y|Q,H) = f_1(Y|Q) = \prod_{q} f_0(Y_q) \]
  if heritable
  \[ Y_q = \{ Y_i | Q_i = q \} \] trait values with genotype \( Q=q \)

expression meta-traits: pleiotropy

- reduce 3,000 heritable traits to 3 meta-traits(!)
- what are expression meta-traits?
  - pleiotropy: a few genes can affect many traits
    - transcription factors, regulators
    - weighted averages; \( Z = \sum W \)
  - principle components, discriminant analysis
- infer genetic architecture of meta-traits
  - model selection issues are subtle
    - missing data, non-linear search
    - what is the best criterion for model selection?
  - time consuming process
    - heavy computation load for many traits
    - subjective judgement on what is best

PC across microarray functional groups

Affy chips on 60 mice
-40,000 mRNA
2500+ mRNA show DE
(via EB arrays with marker regression)
1500+ organized in 85 functional groups
2-35 mRNA / group
which are interesting?
examine PC1, PC2
circle size = # unique mRNA
factor loadings for PC1&2

how well does PC1 do?
lod peaks for 2 QTL at best pair of chr
data (red) vs. 500 permutations (boxplots)

blue bars at 1%, 5%; width proportional to group size

84 PC meta-traits by functional group
focus on 2 interesting groups

red lines: peak for PC meta-trait
black/blue: peaks for mRNA traits
arrows: cis-action?

DA meta-traits: separate pleiotropy
from environmental correlation

pleiotropy only
environmental correlation only
both
Korol et al. (2001)
interaction plots for DA meta-trait
DA for all pairs of markers:
separate 9 genotypes based on markers
(a) same locus pair found with PC meta-trait
(b) Chr 2 region interesting from biochemistry (Jessica Byers)
(c) Chr 5 & Chr 9 identified as important for insulin, SCD

comparison of PC and DA meta-trait on 1500+ mRNA traits

relating meta-trait to mRNA traits
• genotype mean prior assumed independent across traits
• multivariate phenotype averaged over genotypic mean
• posterior for graph given multivariate trait & architecture

building graphical models
• infer genetic architecture of meta-trait
  - \( E(Z \mid Q, M) = \mu_q = \beta_0 + \sum_{q \in M} \beta q_k \)
• find mRNA traits correlated with meta-trait
  - \( Z \approx YW \) for modest number of traits \( Y \)
• extend meta-trait genetic architecture
  - \( M = \) genetic architecture for \( Y \)
  - expect subset of QTL to affect each mRNA
  - may be additional QTL for some mRNA

posterior for graphical models
• posterior for graph given multivariate trait & architecture
  \( \text{pr}(G \mid Y, Q, M) = \text{pr}(Y \mid Q, G) \text{ pr}(G \mid M) / \text{pr}(Y \mid Q) \)

  - \( \text{pr}(G \mid M) = \) prior on valid graphs given architecture
• multivariate phenotype averaged over genotypic mean \( \mu \)
  \( \text{pr}(Y \mid Q, G) = f_q(Y \mid Q, G) = \Pi_q f_q(Y_q \mid G) \)
  \( f_q(Y_q \mid G) = [L_{Y_q} \mid G, G \text{ pr}(G) \text{ d}G] \)
• graphical model \( G \) implies correlation structure on \( Y \)
• genotype mean prior assumed independent across traits
  \( \text{pr}(G) = \Pi, \text{pr}(\mu) \)
from graphical models to pathways

- build graphical models
  QTL → RNA1 → RNA2
  - class of possible models
  - best model = putative biochemical pathway
- parallel biochemical investigation
  - candidate genes in QTL regions
  - laboratory experiments on pathway components

\[
f_2(Y | Q, G=g) = f_1(Y_1 | Q) \cdot f_1(Y_2 | Q, Y_1)
\]