Gene Mapping for High Throughput Expression Profiles: Lessons from Diabetes

Brian S. Yandell
University of Wisconsin-Madison
www.stat.wisc.edu/~yandell/statgen

Outline

• why study diabetes in a mouse model?
• why map gene expression?
• what are QTL?
  – why multiple QTL?
  – how to select genetic architecture?
• how to map massive gene expression?
• preliminary results
Type 2 Diabetes Mellitus

from Unger & Orci FASEB J. (2001) 15,312
studying diabetes in an F2

- segregating cross of inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (chr 6)
  - measured and mapped body weight, insulin, glucose at various ages
    (Stoehr et al. 2000 Diabetes)
  - sacrificed at 14 weeks, tissues preserved
- gene expression data
  - Affymetrix microarrays on parental strains, F1
    - key tissues: adipose, liver, muscle, β-cells
    - novel discoveries of differential expression (Nadler et al. 2000 PNAS; Lan et al. 2002 in review; Ntambi et al. 2002 PNAS)
  - RT-PCR on 108 F2 mice liver tissues
    - 15 genes, selected as important in diabetes pathways
    - SCD1, PEPCK, ACO, FAS, GPAT, PPARgamma, PPARalpha, G6Pase, PDI,…
why map gene expression as a quantitative trait?

- *cis*- or *trans*-action?
  - does gene control its own expression?
  - evidence for both modes (Brem et al. 2002 Science)
- mechanics of gene expression mapping
  - measure gene expression in intercross (F2) population
  - map expression as quantitative trait (QTL technology)
  - adjust for multiple testing via false discovery rate
- research groups working on expression QTLs
  - review by Cheung and Spielman (2002 *Nat Gen Suppl*)
  - Kruglyak (Brem et al. 2002 *Science*)
  - Doerge et al. (Purdue); Jansen et al. (Wageningen)
  - Williams et al. (U KY); Lusis et al. (UCLA)
  - Dumas et al. (2000 *J Hypertension*)

What is a QTL?

- QTL = quantitative trait locus (or loci)
  - trait = phenotype = characteristic of interest
  - quantitative = measured somehow
    - qualitative traits can often be directly mapped
    - quantitative traits not readily mapped
  - locus = location in genome affecting trait
    - gene or collection of tightly linked genes
    - some physical feature of genome
LOD map for PDI: *cis*-regulation

interval mapping basics

- observed measurements
  - \(Y\) = phenotypic trait
  - \(X\) = markers & linkage map
    - \(i\) = individual index 1,…,\(n\)
- missing data
  - missing marker data
  - \(Q\) = QT genotypes
    - alleles \(QQ\), \(Qq\), or \(qq\) at locus
- unknown genetic architecture
  - \(\lambda\) = QT locus (or loci)
  - \(\theta\) = genetic action
  - \(m\) = number of QTL
- \(\text{pr}(Q|X,\lambda,m)\) recombination model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for \(Q\) given \(X\)
- \(\text{pr}(Y|Q,\theta,m)\) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters \(\theta\) (could be non-parametric)

after Sen Churchill (2001)
interval mapping
details and interpretation

• likelihood models relation of data to unknown architecture
  – \( L(\lambda, \theta | m) = pr(Y|X, \lambda, \theta, m) \)
    = product, \( \sum_Q pr(Q|X, \lambda, m) pr(Y|Q, \theta, m) \)
  – complicated to evaluate: product of sum of products

• classical interval mapping: maximize LOD
  – LOD(\( \lambda \)) = \( \max_\theta \log_{10} \frac{L(\lambda, \theta | Y, m)}{L(\mu | Y)} \)
    • scan loci systematically across genome
  – threshold for testing presence vs. no QTL
    • theory for single QTL
      (Lander Botstein 1989; Dupuis Siegmund 1999 Genetics)
    • permutation tests for more general setting
      (Churchill Doerge 1994; Doerge Churchill 1996 Genetics)

• study genetic architecture
  – assess with Bayesian Information Criteria (BIC)

high throughput:
which genes are the key players?

• one approach: clustering of expression
  seed by insulin, glucose
• advantage:
  subset relevant to trait
• disadvantage:
  still many genes to study
SCD1, FAS, GPAT, PEPCK: 
*trans*-regulation by multiple QTL?

Multiple Interval Mapping
SCD1: multiple QTL plus epistasis!
multiple QTL & gene expression

- does one locus affect expression of many genes?
  - is this a controlling locus?
  - is there coordinated expression across many genes?
- multiple QTL affecting gene expression?
  - multiple controlling loci for key pathways?
  - single QTL approach would be inadequate
- multiple QTL literature
  - multiple interval mapping (Kao, et al. 1999 Genetics; Zeng et al. 2000 Genetics; Broman Speed 2002 JRSSB)
  - Bayesian interval mapping (Satagopan et al. 1996 Genetics; Satagopan Yandell 1996; Stevens Fisch 1998 Biometrics; Silanpää Arjas 1998, 1999 Genetics; Sen Churchill 2001 Genetics; Gaffney 2001; Yi Xu 2002 Genetics)
how many (detectable) QTL?

- many, many QTL may affect most any trait
  - how many QTL are detectable with these data?
    - limits to useful detection (Bernardo 2000)
    - depends on sample size, heritability, environmental variation
  - consider probability that a QTL is in the model
    - avoid sharp in/out dichotomy
    - major QTL usually selected, minor QTL sampled infrequently
- build $m =$ number of QTL detected into QTL model
  - directly allow uncertainty in genetic architecture
  - model selection over number of QTL, architecture
  - use Bayes factors and model averaging
    - to identify “better” models
Bayesian interpretation

• consider likelihood of data augmented by QTL genotypes
  \[ \text{pr}(Y,Q|X,\lambda,\theta,m) = \text{product, pr}(Q|X,\lambda,m) \text{ pr}(Y|Q,\theta,m) \]

• reinterpret likelihood as posterior for architecture
  \[ \text{pr}(\lambda,Q,\theta,m|Y,X) = \text{[product, pr}(Q|X,\lambda,m) \text{ pr}(Y|Q,\theta,m)] \text{ [pr}(\lambda,\theta|X,m)\text{pr}(m)] \]
  = \text{[augmented likelihood] x [prior]}

• examine posterior of architecture given data
  – controlling loci \( \lambda \) and gene action \( \theta \)
  \[ \text{pr}(\lambda,\theta|Y,X,m) = \sum_Q \text{pr}(\lambda,Q,\theta|Y,X,m) \text{ pr}(m) \]
  with \( m \) fixed

• assess using Bayes factors
  – extends Bayes Information Criterion to compare any 2 models

Bayes factors to assess models

• Bayes factor: which model best supports the data?
  – ratio of posterior odds to prior odds
  – ratio of model likelihoods

• equivalent to LR statistic when
  – comparing two nested models
  – simple hypotheses (e.g. 1 vs 2 QTL)

• related to Bayes Information Criteria (BIC)
  – Schwartz introduced for model selection in general settings
  – penalty to balance model size \((p = \text{number of parameters})\)

\[ BF = \frac{\text{pr}(m|Y,X) / \text{pr}(m+1|Y,X)}{\text{pr}(m) / \text{pr}(m+1)} = \frac{\text{pr}(Y|m,X)}{\text{pr}(Y|m+1,X)} \]

\[ -2 \log(BF) = -2 \log(LR) - 2 \log(n) = BIC \]
computing QTL Bayes factors

- easy to compute Bayes factors from samples
  - sample posterior using MCMC
  - posterior \( \text{pr}(m|Y,X) \) is marginal histogram
  - posterior affected by prior \( \text{pr}(m) \)

- BF insensitive to shape of prior
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior

- BF sensitivity to prior variance on effects \( \theta \)
  - prior variance should reflect data variability
  - resolved by using hyper-priors
    - automatic algorithm; no need for user tuning

multiple QTL phenotype model

- \( Y = \mu + G_Q + \text{environment} \)
- partition genotypic effect into separate QTL effects
  - \( G_Q = \text{main QTL effects} + \text{epistatic interactions} \)
  - \( G_Q = \theta_{1Q} + \ldots + \theta_{mQ} + \theta_{12Q} + \ldots \)

- priors on mean and effects
  - \( G_Q \sim N(0, h^2 s^2) \) model independent genotypic prior
  - \( \theta_{1Q} \sim N(0, \kappa_1 s^2/m.) \) additive effects (down-weighted)
  - \( \theta_{12Q} \sim N(0, \kappa_2 s^2/m.) \) epistatic interactions (down-weighted)

- hyper-parameters (to reduce sensitivity of Bayes factors to prior)
  - \( s^2 = \text{total sample variance} \)
  - \( m = m + m_2 = \text{number of QTL effects and interactions} \)
  - \( h^2 = (m_1 \kappa_1 + m_2 \kappa_2)/m. = \text{unknown heritability, } h^2/2 - \text{Beta}(a,b) \)
Bayesian model assessment: number of QTL for SCD1

Bayesian LOD and $h^2$ for SCD1
Bayesian model assessment: chromosome QTL pattern for SCD1

trans-acting QTL for SCD1
(no epistasis yet: see Yi Xu 2002)
2-D scan: assumes only 2 QTL!

1-D and 2-D marginals

\[ \text{pr(QTL at } \lambda \mid Y, X, m) \]

unlinked loci

linked loci
false detection rates and thresholds

- multiple comparisons: test QTL across genome
  - size = \( \Pr(\text{LOD}(\lambda) > \text{threshold} \mid \text{no QTL at } \lambda) \)
  - threshold guards against a single false detection
    - very conservative on genome-wide basis
  - difficult to extend to multiple QTL
- positive false discovery rate (Storey 2001)
  - pFDR = \( \Pr(\text{no QTL at } \lambda \mid \text{LOD}(\lambda) > \text{threshold} ) \)
  - Bayesian posterior HPD region based on threshold
    - \( \Lambda = \{\lambda \mid \text{LOD}(\lambda) > \text{threshold} \} \approx \{\lambda \mid \Pr(\lambda \mid Y, X, m) \text{ large} \} \)
    - extends naturally to multiple QTL

pFDR and QTL posterior

- positive false detection rate
  - pFDR = \( \Pr(\text{no QTL at } \lambda \mid Y, X, \lambda \text{ in } \Lambda) \)
  - pFDR = \( \frac{\Pr(H=0) \times \text{size}}{\Pr(m=0) \times \text{size} + \Pr(m>0) \times \text{power}} \)
  - power = posterior = \( \Pr(\text{QTL in } \Lambda \mid Y, X, m>0) \)
  - size = (length of \( \Lambda \)) / (length of genome)
- extends to other model comparisons
  - \( m = 1 \) vs. \( m = 2 \) or more QTL
  - pattern = ch1,ch2,ch3 vs. pattern > 2*ch1,ch2,ch3
pFDR for SCD1 analysis

prior probability
fraction of posterior
found in tails

pr( H=0 | p>size )

BH pFDR(-) and size(.)

Storey pFDR(-)

trans-acting QTL for SCD1

hong7ecd blm summaries with pattern > ch2, ch5, ch9

dominance?
high throughput dilemma

- want to focus on gene expression network
  - ideally capture pathway in a few dimensions
  - allow for complicated genetic architecture
- may have multiple controlling loci
  - could affect many genes in coordinated fashion
  - could show evidence of epistasis
  - quick assessment via interval mapping may be misleading
- mapping principle component as quantitative trait
  - multiple interval mapping with epistatic interactions
  - Liu et al. (1996 *Genetics*); Zeng et al. (2000 *Genetics*) Mahler et al. (2002 *Genomics*)
pFDR for PC1 analysis

prior probability
fraction of posterior
found in tails

mapping controlling loci via PC

- $Y =$ expression data from chips for F2 population
  - principle components (singular value decomposition)
    - $Y = UDV^T$
    - $V$ has eigen-genes as rows, individuals as columns
      - Hilsenbeck et al. (1999); Alter et al. (2000); West et al. (2000)
- $V =$ combined expression of coordinated genes
  - map $V_1$, $V_2$ as quantitative traits
  - identify mRNA with strong correlation: coordinated expression?
\[ Y = UDVT \]

Alter et al. (2000 PNAS) yeast cell cycle

PC simply rotates & rescales to find major axes of variation
multivariate screen for gene expressing mapping

principal components

PC1 (42%) PC1(red) and SCD(black)

PC2 (22%)

Relation of Composite Phenotypes to Individual mRNA Expressions (after West et al. 2000)
SVD Pros and Cons

• advantages
  – superphenotypes $V_1$, $V_2$, ... are orthogonal
  – may only need a few
    • how fast do eigen-values $D$ drop?
  – can dramatically increase power to detect QTL

• disadvantages
  – less efficient if many large eigen-values
  – may be difficult to interpret some superphenotypes
  – PCs may not reflect genetic differential expression
    • could iterate on putative QTL to improve discrimination

Ongoing & Future Work

• fine mapping via congenic lines
  – ongoing for physiological traits
  – candidate genes emerging

• new F2 population focusing on islets
  – expression mapping on a large scale (100-200 mice)
  – development of new methodology (Jin, Yang, Lan)

• model selection for genetic architecture
  – fast computation for multiple QTL (Yi, Gaffney)
  – high throughput model assessment
Summary

• mouse model for diabetes
  – studying pathways via gene expression
  – massive number of phenotypes: expression arrays

• model selection for multiple QTL
  – Bayes factors for model assessment
  – posteriors can reveal subtle hints of QTL
  – multiple trait mapping…

• dimension reduction to elicit pathways
  – study genetic architecture of "supergenes"
  – unravel correlation with individual mRNA

• connection to false discovery rate
  – whole genome evaluation
  – calibrate posterior region with pFDR

Collaborators

www.stat.wisc.edu/~yandell/statgen

Alan D. Attie\textsuperscript{3}
Hong Lan\textsuperscript{3}
Samuel T. Nadler\textsuperscript{3}

\textsuperscript{3}UW-Madison Biochemistry

Yi Lin\textsuperscript{1}
Christina Kendziorski\textsuperscript{4}
Yang Song\textsuperscript{1}
Fei Zou\textsuperscript{5}
Pat Gaffney\textsuperscript{6}
Jaya Satagopan\textsuperscript{7}

\textsuperscript{1}UW-Madison Statistics
\textsuperscript{4}UW-Madison Biostatistics
\textsuperscript{5}UNC Biostatistics
\textsuperscript{6}Lubrizol
\textsuperscript{7}Memorial Sloan Kettering
software

- www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
  - module using QtlCart format
  - compiled in C for Windows/NT
  - extensions in progress
  - R post-processing graphics
    - library(bim) is cross-compatible with library(qtl)

- Bayes factor and reversible jump MCMC computation

- enhances MCMCQTL and revjump software
  - initially designed by JM Satagopan (1996)
  - major revision and extension by PJ Gaffney (2001)
    - whole genome
    - multivariate update of effects; long range position updates
    - substantial improvements in speed, efficiency
    - pre-burnin: initial prior number of QTL very large