Lecture 6 (part): Association Mapping

Bruce Walsh lecture notes
Tucson Winter Institute
9 - 11 Jan 2013

Association mapping

- Review of basic ideas
- General pedigree (random effects) models
- Mixed-models
  - SNP as fixed effects, background as random effects
  - QK method
  - Heterotic marker effects

Basic ideas behind association mapping

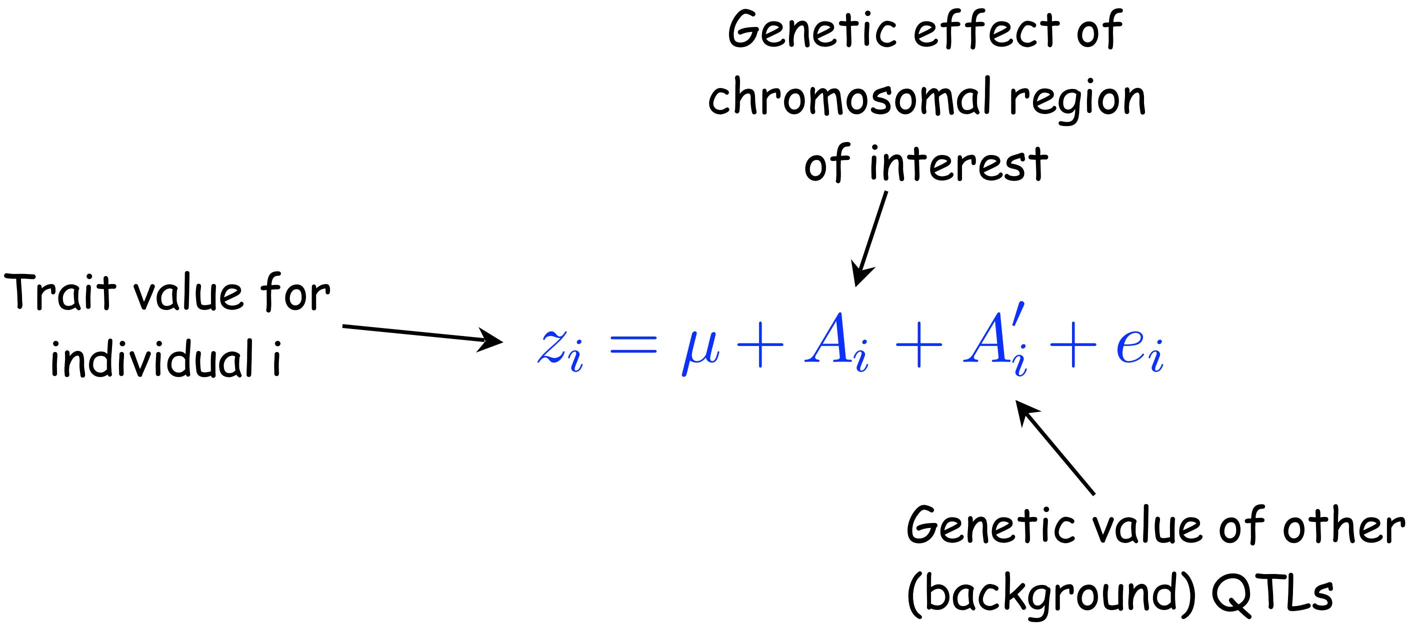
- A new mutation is initially in complete LD with tightly-linked markers
  - Since LD declines as $(1-c)^t$, if c is small, LD will persist for a very long time
  - Generates a marker-trait association that will appear in a random sample of individuals from a population.
- Problem: If population subdivision is present, and mean trait values differ over subpopulations, a particular marker providing information on group status will also generate a marker-trait association

Model 1: Random effects, pedigree models

- Suppose our association sample of individuals contains some relatives
- Want to separate shared pedigree effects on a trait from shared marker effects
- First model appeared in human genetics, the random-effects pedigree model
    - Region was treated as a random effect (a significant marker variance indicated a QTL in the region)
    - A background polygenic effect shared over relatives also included.

General Pedigree Methods

Random effects (hence, variance component) method for detecting QTLs in general pedigrees



Genetic effect of chromosomal region of interest

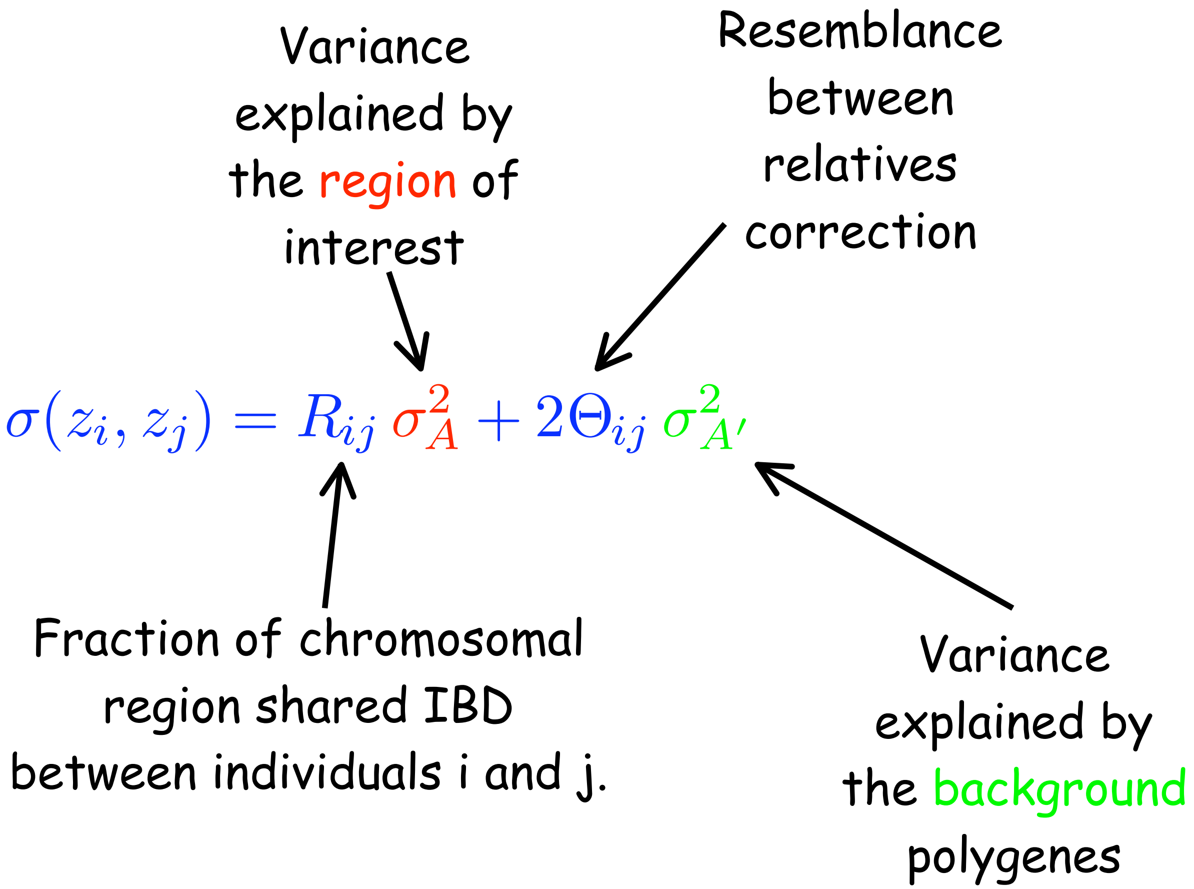
Trait value for individual i

$$z_i = \mu + A_i + A'_i + e_i$$

Genetic value of other (background) QTLs

$$z_i = \mu + A_i + A_i' + e_i$$

The covariance between individuals i and j is thus



$$\sigma(z_i, z_j) = R_{ij}\,\sigma^2_A + 2\Theta_{ij}\,\sigma^2_{A'}$$

Variance explained by the region of interest

Resemblance between relatives correction

Fraction of chromosomal region shared IBD between individuals i and j.

Variance explained by the background polygenes

Assume z is MVN, giving the covariance matrix as

$$\mathbf{V} = \mathbf{R}\,\sigma_A^2 + \mathbf{A}\,\sigma_{A'}^2 + \mathbf{I}\,\sigma_e^2$$

Here

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \widehat{R}_{ij} & \text{for } i \neq j \end{cases}, \qquad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases}$$

Estimated from marker data

Estimated from the pedigree

The resulting likelihood function is

$$\ell(\,\mathbf{z}\,|\,\mu, \sigma_A^2, \sigma_{A'}^2, \sigma_e^2\,) = \frac{1}{\sqrt{(2\pi)^n\,|\mathbf{V}|}}\,\exp\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\,\mathbf{V}^{-1}\,(\mathbf{z}-\boldsymbol{\mu})\right]$$

A significant $\sigma_A{}^2$ indicates a QTL in the focal region.

The next step, mixed models

- The random effects model (assuming marker effects are random) suffers from low power
  - Random effects estimate a variance, which has a much larger sampling variance than fixed effects which estimate a mean
- Mixed models treat SNP effects as fixed, background as random
  - As SNP maps because more dense, could treat each SNP as a fixed effects --- compare mean difference of 00, 01, and 11 genotypes

Basic mixed model

Assume a collection of unrelated individuals drawn from a single population. There are potentially a number of fixed effects in addition to the effects associated with the current SNP being considered

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{S}\mathbf{a} + \mathbf{e}$$

SNP effects

Fixed effects in blue, random effects in red

Example

Suppose individuals 1 and 2 have genotype 00, individual 3 genotype 11, and individual 2 genotype 10

A general model for effects is 00 = -a, 01 = d, 11 = a

Resulting **S** and **a** matrices for this model is

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} -a \\ d \\ a \end{pmatrix}$$

Expanding this mixed model

- One complication is the presence of related individuals within the sample
  - General pedigree methods can accommodate this
- A second complication is that the sample may come from several different populations, which may differ in their mean trait values
  - A SNP marker may be informative as to group membership, generating a marker-trait association even though it may be unlinked to any QTL

Structured Association Mapping

Pritchard and Rosenberg (1999) proposed Structured Association Mapping, wherein one assumes k subpopulations (each in Hardy-Weinberg).

Given a large number of markers, one then attempts to assign individuals to groups using an MCMC Bayesian classifier (their program STRUCTURE)

Once individuals assigned to groups, association mapping without any correction can occur in each group.

Structure plus Kinship Methods

Association mapping in plants often occurs by first taking a large collection of lines, some closely related, others more distantly related. Thus, in addition to this collection being a series of subpopulations (derivatives from a number of founding lines), there can also be additional structure within each subpopulation (groups of more closely related lines within any particular lineage).

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{S}\mathbf{a} + \mathbf{Q}\mathbf{v} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

Fixed effects in blue, random effects in red

Q-K method

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{S}\mathbf{a} + \mathbf{Q}\mathbf{v} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

β = vector of fixed effects

$\mathbf{a}$ = SNP effects

$\mathbf{v}$ = vector of subpopulation effects (STRUCTURE)
Q_{ij} = Prob(individual i in group j). Determined from STRUCTURE output

$\mathbf{u}$ = shared polygenic effects due to kinship.
Cov($\mathbf{u}$) = var(A)****A***, where the relationship matrix ***A*** estimated from marker data matrix K

Even more general models

- The mixed-model machinery introduced in the last lecture easily extends to incorporating marker information
- Bernardo (11.8.1) gives an example where sets of inbreds (P1 and P2) from two heterotic groups are crossed
  - Can estimate the GCAs and SCAs in additional to the effects for marker alleles on variation in P1, variation in P2, and the interaction effects in the cross

Fixed effects in blue, random effects in red

$$Y = X\beta + Z_1 g_1 + Z_2 g_2 + Z_3 s + W_1 m_1 + W_2 m_2 + W_3 m + e$$

g_i is the vector of GCAs for the lines in the set P_i, **s** the vector of SCA for these crosses

m_i is the vector of SNP marker effects for the Lines in the set P_i, **m** the vector SNP interaction effects in their cross