Multiple Trait Analysis

observations on multiple traits
one or more traits in multiple environments

Does QTL have pleiotropic effects on multiple traits?
Does QTL show genotype-environment interaction?
What is genetic correlation between different traits?
Is correlation due to pleiotropy or linkage? Where?

view multiple traits as multivariate vector
Falconer (1952); Jiang Zeng (1995)

statistical models and likelihood analyses
hypothesis tests of QTL effects
pleiotropy vs. close linkage
QTL by environment interaction

Statistical Models and Likelihood Analyses
CIM model for multiple traits

sample of n individuals from a F_2 population
additive effects only (for now)
observe m quantitative traits
CIM scan for QTL on a marker interval (M_i, M_{i+1})

$$\begin{aligned} y_{j1} &= b_{01} + a_1^* x_j^* + \sum_l b_{l1} x_{jl} + e_{j1} \\ y_{j2} &= b_{02} + a_2^* x_j^* + \sum_l b_{l2} x_{jl} + e_{j2} \\ &\vdots \quad\quad \vdots \\ y_{jm} &= b_{0m} + a_m^* x_j^* + \sum_l b_{lm} x_{jl} + e_{jm} \end{aligned}$$

y_{jk} : phenotype of kth trait on individual j
b_{0k} : mean effect (reference) for trait k
a_k^* : additive effect of putative QTL on trait k
x_j^* : number of alleles of P_1 at putative QTL
x_{jl} : genotype at marker l
b_{lk} : marker regression coefficients
e_{jk} : residual effect on trait k for individual j

assumptions on residual "error" effects e_{jk}

errors correlated among traits within individuals
covariance: $Cov(e_{jk}, e_{jl}) = \sigma_{kl} = \sigma_{lk} = \rho_{kl}\sigma_k\sigma_l$
independent individuals: $Cov(e_{jk}, e_{ik}) = 0$

variance-covariance matrix

errors multivariate normal among individuals
mean zero
general covariance matrix ($\sigma_{kl} = \sigma_{lk}$)

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{pmatrix}$$

vector notation

$$\mathbf{y}_j = x_j^* \mathbf{a}^* + \mathbf{x}_j^{\mathsf{T}} \mathbf{B} + \mathbf{e}_j$$

$\mathbf{y}_j$ = m vector of phenotypes y_{jk}
$\mathbf{a}^*$ = m vector of QTL effects a_k^*
$\mathbf{x}_j$ = $n_p + 1$ vector of 1 and marker data x_{jk}
$\mathbf{B}$ = $(n_p + 1) \times m$ matrix of cofactor effects
= (reference b_{0k} and cofactors b_{lk})
$\mathbf{e}_j$ = m vector of errors e_{jk}
$Cov(\mathbf{e}_j)$ = $\mathbf{V}$ covariance matrix

matrix notation

$$\mathbf{Y} = \mathbf{x}^{*\mathsf{T}} \mathbf{a}^* + \mathbf{X}\mathbf{B} + \mathbf{E}$$

$\mathbf{Y}$ = $n \times m$ matrix of y_{jk} (row j = $\mathbf{y}_j$)
$\mathbf{x}^*$ = n vector of x_j^*
$\mathbf{X}$ = $n \times (n_p + 1)$ marker matrix (column j = $\mathbf{x}_j$)
$\mathbf{E}$ = $n \times m$ error matrix of e_{jk} (row j = $\mathbf{e}_j$)

choice of background markers?
additive and dominance effects?
same issues for selecting cofactors as ordinary CIM

likelihood analysis

$$\begin{aligned}\mathbf{Y} &= \mathbf{x}^{*\mathsf{T}}\mathbf{a}^* + \mathbf{XB} + \mathbf{E} \\ \mathbf{y}_j &= x_j^*\mathbf{a}^* + \mathbf{x}_j^{\mathsf{T}}\mathbf{B} + \mathbf{e}_j \\ Cov(\mathbf{e}_j) &= \mathbf{V}\end{aligned}$$

however we do not know $\mathbf{x}^* = \{x_j^*\}$
mixture model with multivariate normal

$$L_1 = \prod_{j=1}^{n}\left[\sum_k p_{kj} f_k(\mathbf{y}_j)\right]$$

p_{kj} = $Prob\{x_j^* = k|\text{markers}\}$ for putative QTL
$f_k(\mathbf{y}_j)$ = $\phi(k\mathbf{a}^* + \mathbf{x}_j^{\mathsf{T}}\mathbf{B}, \mathbf{V})$ multivariate normal

maximum likelihood estimates

Expectation/Conditional Maximization (ECM)
special version of general EM algorithms
Meng Rubin 1993

Expectation E-step

individual posterior QTL genotype probabilities

$$P_{kj}^{(t+1)} = \frac{p_{kj} f_k^{(t)}(\mathbf{y}_j)}{\sum_l p_{lj} f_l^{(t)}(\mathbf{y}_j)}$$

$f_k^{(t)}(\mathbf{y}_j)$ = normal density functions with parameters replaced by estimates in iteration t

Conditional Maximization CM-step

model parameters divided into three groups:
QTL $(\mathbf{a}^*, \mathbf{d}^*)$, cofactors $(\mathbf{B})$, covariance $(\mathbf{V})$
estimated consecutively between groups
but simultaneously within each group

Conditional Maximization CM-step

model parameters divided into three groups
estimated consecutively between groups
but simultaneously within each group

QTL

$$\mathbf{a}^{*(t+1)} = \mathbf{P}_2^{(t+1)\mathsf{T}}(\mathbf{Y} - \mathbf{XB}^{(t)})/(2\mathbf{P}_2^{(t+1)\mathsf{T}}\mathbf{1})$$

$$\mathbf{d}^{*(t+1)} = \left[\mathbf{P}_1^{(t+1)\prime}/(\mathbf{P}_1^{(t+1)\prime}\mathbf{1}) - \mathbf{P}_2^{(t+1)\prime}/(2\mathbf{q}_2^{(t+1)\prime}\mathbf{1})\right](\mathbf{Y} - \mathbf{XB}^{(t)})$$

cofactors

$$\mathbf{B}^{(t+1)} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\left[\mathbf{Y} - (2\mathbf{P}_2^{(t+1)} + \mathbf{P}_1^{(t+1)})\mathbf{a}^{*(t+1)} - \mathbf{P}_1^{(t+1)}\mathbf{d}^{*(t+1)}\right]$$

covariance

$$\begin{aligned}\mathbf{V}^{(t+1)} = &\Big[(\mathbf{Y} - \mathbf{XB}^{(t+1)})^{\mathsf{T}}(\mathbf{Y} - \mathbf{XB}^{(t+1)}) - 4(\mathbf{P}_2^{(t+1)\mathsf{T}}\mathbf{1})\mathbf{a}^{*(t+1)\mathsf{T}}\mathbf{a}^{*(t+1)} \\ &-(\mathbf{P}_1^{(t+1)\mathsf{T}}\mathbf{1})(\mathbf{a}^{*(t+1)} + \mathbf{d}^{*(t+1)})^{\mathsf{T}}(\mathbf{a}^{*(t+1)} + \mathbf{d}^{*(t+1)})\Big]/n\end{aligned}$$

$\mathbf{P}_k^{(t+1)}$ = n vector of $P_{kj}^{(t+1)}$ genotype probabilities
$\mathbf{1}$ = column vector of ones
$P_{kj}^{(0)}$ = p_{kj} initial values
$\mathbf{a}^{*(0)}$ = 0 (or some other initial value)

iterations terminated with a predetermined criterion
changes of estimates or log-likelihood value is negligible ($< 10^{-8}$)
final estimates $\hat{\mathbf{a}}^*$, $\hat{\mathbf{B}}$, $\hat{\mathbf{V}}$ used for LR (or LOD) test

log-likelihood with parameter estimates

$$\ln(L_1(\lambda)) = -\frac{nm\ln(2\pi)}{2} - \frac{n}{2}\ln(|\hat{\mathbf{V}}|) +$$

$$\sum_{j=1}^{n}\ln\left\{\sum_k p_{kj}\exp\left[-\frac{1}{2}(\mathbf{y}_j - k\hat{\mathbf{a}}^* - \mathbf{x}_j^{\mathsf{T}}\hat{\mathbf{B}})^{\mathsf{T}}\hat{\mathbf{V}}^{-1}(\mathbf{y}_j - k\hat{\mathbf{a}}^* - \mathbf{x}_j^{\mathsf{T}}\hat{\mathbf{B}})\right]\right\}$$

$$\ln(L_1(\lambda)) = -\frac{nm\ln(2\pi)}{2} - \frac{n}{2}\ln(|\hat{\mathbf{V}}|) - \frac{1}{2}\sum_{j=1}^{n}(\mathbf{y}_j - \mathbf{x}_j^{\mathsf{T}}\hat{\mathbf{B}})^{\mathsf{T}}\hat{\mathbf{V}}^{-1}(\mathbf{y}_j - \mathbf{x}_j^{\mathsf{T}}\hat{\mathbf{B}})$$

$$+\sum_{j=1}^{n}\ln\left\{\sum_k p_{kj}\exp\left[k\hat{\mathbf{a}}^{*\mathsf{T}}\hat{\mathbf{V}}^{-1}(2\mathbf{y}_j - k\hat{\mathbf{a}}^* - 2\mathbf{x}_j^{\mathsf{T}}\hat{\mathbf{B}})\right]\right\}$$

$|\hat{\mathbf{V}}|$ = determinant of covariance matrix

log-likelihood under null model of no QTL

QTL dropped, but cofactors remain
note that covariance matrix estimate changes

$$\begin{aligned}\ln(L_0) &= -\frac{nm\ln(2\pi)}{2} - \frac{n}{2}\ln(|\hat{\mathbf{V}}_0|) - \frac{nm}{2} \\ \hat{\mathbf{V}}_0 &= (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_0)^{\mathsf{T}}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_0)/n \\ \hat{\mathbf{B}}_0 &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\end{aligned}$$

Hypothesis Tests of QTL Effects

model 1 = full model (QTL for all m traits)
model 0 = null model (no QTL)
intermediate models: QTL for only some traits
additive and/or dominance
case of $m = 2$ traits has key features

joint mapping for QTL on two traits

map QTL for each trait individually or jointly on both?
joint mapping hypotheses

H_0 : $a_1^* = 0,\ d_1^* = 0,\ a_2^* = 0,\ d_2^* = 0$
H_1 : At least one of them is not zero

likelihood ratio test statistic

$$LR_1 = -2\ln(L_0/L_1(\lambda))$$

approximately chi-square distributed under H_0

threshold value for significance

hard to determine critical value for whole genome
same problem as CIM (Zeng 1994)
Bonferroni approximate test

extend permutation test of Churchill Doerge
available in theory–not implemented

why perform joint mapping?

formal procedures to test biologically interesting hypotheses
– pleiotropic effects of QTL
– QTL by environment interaction
– pleiotropy vs. close linkage

may perform better than separate CIM
– putative QTL has pleiotropic effects on both traits
– genotypic & environmental correlation opposite

Testing pleiotropic effects

does one QTL affect more than one trait?
pick a genome position λ and jointly test traits
pleiotropic effects on both traits

H_{10} : $a_1^* = 0, d_1^* = 0, a_2^* \neq 0, d_2^* \neq 0$ at λ
only trait 2 is affected
H_{11} : $a_1^* \neq 0, d_1^* \neq 0, a_2^* \neq 0, d_2^* \neq 0$ at λ
both traits affected by QTL

and

H_{20} : $b_1^* \neq 0, d_1^* \neq 0, b_2^* = 0, d_2^* = 0$ at λ
only trait 2 is affected
H_{21} : $b_1^* \neq 0, d_1^* \neq 0, b_2^* \neq 0, d_2^* \neq 0$ at λ
both traits affected by QTL

$H_{11} = H_{21}$ is alternative of pleiotropy
need to test H_{10} and H_{20} together

estimates and tests under restrictions

test of H_{10} vs. H_{11} differs from test of trait 1 alone since traits are correlated

test has more power than separate analyses

estimates of model parameters under H_{10} and H_{20} use ECM with some parameters set to 0

likelihood ratio test statistics use these estimates

testing pleiotropic effects against close linkage

rejecting both H_{10} and H_{20} supports hypothesis of pleiotropic effects of a single QTL

what if there were two closely linked QTL?
want to separate genetic correlation from linkage

two closely linked QTL may behave like one pleiotropic QTL
one pleiotropic QTL may be estimated as two QTL with separate trait analysis

implications for genetics and breeding
power to detect the difference?
linkage vs. fine mapping: what is a QTL?

need to focus on small region for test of 2 QTL
only genome regions significant under joint mapping
linkage at distance may be obvious
computation costs

likelihood analysis: pleiotropy vs. close linkage

two QTL at positions λ_1, λ_2
$|\lambda_1 - \lambda_2| < 5cM$ for convenience

$$\begin{aligned} H_0 &: \lambda_1 = \lambda_2 \\ H_1 &: \lambda_1 \neq \lambda_2 \end{aligned}$$

allow both QTL to have effects ($a_k^* \neq 0$)
H_1 is special case of many possible alternatives
more general alternative: both QTL have pleiotropic effects (more complicated)

statistical model for closely linked QTL

$$\begin{aligned} y_{j1} &= b_{01} + a_1^* x_{1j}^* + \sum_l b_{l1} x_{jl} + e_{j1} \\ y_{j2} &= b_{02} + a_2^* x_{2j}^* + \sum_l b_{l2} x_{jl} + e_{j2} \end{aligned}$$

looks like multiple trait model defined earlier
but QTL genotypes $x_{kj}^* = x_{kj}^*(\lambda_k)$
defined for separate QTL at different positions

caution on choice of cofactors

avoid using markers inside search region
models under hypotheses depend on cofactors

mixture model over two loci

nine components: recombination for two loci in F_2
$p_{kij} = \text{Prob}\{x_{1j}^* = k, x_{2j}^* = i | \lambda_1, \lambda_2\}$
probability p_{kij} inferred from flanking markers

– different marker intervals: independence $p_{kij} = p_{kj} p_{ij}$

– same marker interval: Table 11.1 for 4 positions

linked QTL likelihood function

$$L_2(\lambda_1, \lambda_2) = \prod_{j=1}^{n} \sum_{k,i} p_{kij} f_{ki}(\mathbf{y}_j)$$

bivariate normal density $f_{ki}(\mathbf{y}_j)$:

$$E\begin{pmatrix} y_{j1} \\ y_{j2} \end{pmatrix} = \begin{pmatrix} ka_1^* + \mathbf{x}_j^{\mathsf{T}} \mathbf{b}_1 \\ ia_2^* + \mathbf{x}_j^{\mathsf{T}} \mathbf{b}_2 \end{pmatrix}, \quad Var(\mathbf{y}_j) = \mathbf{V}$$

ECM iteration to maximize likelihood

E-step: posterior probabilities of QTL genotypes

$$P_{kij}^{(t+1)} = \frac{p_{kij} \hat{f}_{ki}^{(t)}(\mathbf{y}_j)}{\sum_{k,i} p_{kij} \hat{f}_{ki}^{(t)}(\mathbf{y}_j)}$$

CM-step: maximize likehood estimates

QTL effects

$$a_1^{*(t+1)} = \text{blah}$$

cofactors

$$\mathbf{B}^{(t+1)} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}^{(t+1)}$$

variance

$$\mathbf{V}^{(t+1)} = \frac{(\mathbf{W} - \mathbf{XB})^{\mathsf{T}}(\mathbf{W} - \mathbf{XB})}{n}$$

where $\mathbf{B} = (\mathbf{B}_1\ \mathbf{B}_2)$, $\mathbf{W} = (\mathbf{W}_1\ \mathbf{W}_2)$,

$$\begin{aligned} \mathbf{W}_1 &= \mathbf{Y}_1 - (\textstyle\sum k\mathbf{P}_{k\cdot})^{\mathsf{T}} a_1^* \\ \mathbf{W}_2 &= \mathbf{Y}_2 - (\textstyle\sum i\mathbf{P}_{\cdot i})^{\mathsf{T}} a_2^* \end{aligned}$$

with $\mathbf{Y}_l = \{y_{jl}\}$ and $\mathbf{P}_{k\cdot} = \sum_i \mathbf{P}_{ki}$, $\mathbf{P}_{\cdot i} = \sum_k \mathbf{P}_{ki}$

joint log likelihood for linked QTL

depends on putative QTL λ_1 and λ_2

$$\ln(L_2(\lambda_1,\lambda_2)) = -\frac{nm\ln(2\pi)}{2} - \frac{n}{2}\ln(|\hat{\mathbf{V}}|) +$$

$$\sum_{j=1}^{n} \ln\left\{\sum_{k,i} p_{kij} \exp\left[-\frac{1}{2}(\mathbf{y}_j - \hat{\mathbf{u}}_{kij})\hat{\mathbf{V}}^{-1}(\mathbf{y}_j - \hat{\mathbf{u}}_{kij})^{\mathsf{T}}\right]\right\}.$$

with $\hat{\mathbf{u}}_{kij} = \hat{E}(\mathbf{y}_j)$
p_{kij}, $\hat{\mathbf{u}}_{kij}$ and $\hat{\mathbf{V}}$ depend on λ_1, λ_2

search for maximum likelihood

search possible λ_1, λ_2 in region
test statistic

$$LR_2 = -2\ln\left(\frac{\max_{\lambda} L_2(\lambda,\lambda)}{\max_{\lambda_1,\lambda_2} L_2(\lambda_1,\lambda_2)}\right)$$

nested hypotheses: asymptotically χ^2_1 under H_0

scan LODs for joint and separate QTL
approximate test: do peaks match?
grid search in neighborhood

QTL by environment interaction

different environments → different gene effects
Paterson *et al.* (1991); Stuber *et al.* (1992)

Design I: same genotypes evaluated in different environments (paired comparison)
Design II: different genotypes (individuals) from common population evaluated in different environments (group comparison)

QTL × environment interaction hypotheses

$$H_0:\ a_1^* = a_2^* = a^*,\ \ d_1^* = d_2^* = d^*$$
$$H_1:\ a_1^* \neq a_2^*,\ \ d_1^* \neq d_2^*$$

only test in regions suggested by joint mapping (why?)

recombination probabilities
$p_{kij} = \text{Prob}\{x^*_{kj} = i\},\ k = 1,2,\ j = 1,\cdots,n_k$

Design I paired comparison

same $\mathbf{X}$ (marker data) matrix
multiple phenotypic vectors $\mathbf{Y}$ across environments
same statistical model as multiple traits
$\mathbf{V}$ reflects within and between environment variation

log likelihood under H_0: not G× E

construct likelihood L_3 under restriction of H_0
maximize likelihood using ECM again
E-step: substitute a^* for a_1^*, a_2^* (d^* for d_l^*)
CM-step: $a^{*(t+1)}$ is $\mathbf{V}^{(t)}$ weighted average

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

likelihood ratio test for G× E

$$LR_3(\lambda) = -2\ln(L_3(\lambda)/L_1(\lambda))$$

asymptotically chi-square under H_0

Design II group comparison

statistical model

$$y_{1j} = x^*_{1j}a_1^* + \mathbf{x}_{1j}^{\mathsf{T}}\mathbf{b}_1 + e_{1j} \quad j = 1,2,\cdots,n_1$$
$$y_{2j} = x^*_{2j}a_2^* + \mathbf{x}_{2j}^{\mathsf{T}}\mathbf{b}_2 + e_{2j} \quad j = 1,2,\cdots,n_2$$

matrix notation

$$\mathbf{y}_1 = \mathbf{x}_1^* a_1^* + \mathbf{X}_1\mathbf{b}_1 + \mathbf{e}_1$$
$$\mathbf{y}_2 = \mathbf{x}_2^* a_2^* + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}_2$$

assume environmental errors e_{1j}, e_{2j} independent normal with means zero and variances σ_1^2, σ_2^2

estimate separately by environment under H_1

sum of separate $\ln(L_1)$s by environment

$$\begin{aligned}\ln(L_4(\lambda)) &= \sum_{j=1}^{n_1}\ln\left(\sum_i p_{1ij}f_i(y_{1j})\right) + \sum_{j=1}^{n_2}\ln\left(\sum_i p_{2ij}f_i(y_{2j})\right) \\ &= \ln(L_{11}(\lambda)) + \ln(L_{12}(\lambda))\end{aligned}$$

$L_{11}(\lambda)$, $L_{12}(\lambda)$ are $L_1(\lambda)$ for groups 1,2

Design II group comparison
estimate jointly under H_0

one QTL effect parameter a^*
same λ, different individuals $1j$ and $2j$
p_{1ij} and p_{2ij} independent but posterior probabilities depend on a^* in E-step through normal density

$$P_{kij}^{(t+1)} = \frac{p_{kij} f_i^{(t)}(y_{kj})}{\sum_{i=0}^{2} p_{kij} f_i^{(t)}(y_{kj})},\ k = 1, 2$$

CM-step involves block update
– QTL effect a^*
– cofactors $\mathbf{B}_1, \mathbf{B}_2$
– variances σ_1^2, σ_2^2

$\ln(L_5(\lambda))$ looks like $\ln(L_4(\lambda))$ with $\hat{a}_1^* = \hat{a}_2^* = \hat{a}^*$
likelihood ratio test statistic

$$LR_{4(\lambda)} = -2 \ln(L_5(\lambda)/L_4(\lambda))$$

asymptotically chi-square under H_0: no G×E
degrees of freedom depend on model (BC, F_2)

relative efficiency of Designs I and II

mapping QTL, testing QTL × environment interaction
assume $n_1 = n_2 = n$ and n large

$LR_4(\lambda)$ is special case of $LR_3(\lambda)$ with $\rho = 0$
Design II: has more power for mapping QTL
Design I: more power to detect QTL × env interaction

QTL × environment as fixed effects here
random effects → mixed models