Mapping Quantitative Trait Loci

by

J. Peng, H. K. Tang, D. Siegmund

(A) Goals

To give a systematic large sample theory for QTL mapping, which (i) clarifies the similarities and differences between QTL mapping in experimental genetics and in humans, (ii) treats issues of study design of recent interest, e.g., (a) the comparative value of large pedigrees versus sib pairs, (b) genotyping only selected pedigrees, and (iii) provides a framework to study gene × gene and gene × covariate interaction.

(B) Methods

Starting from the standard components of variance model and a parameterization of the genetic effects that makes "linkage parameters" orthogonal to "segregation parameters" use the framework of local alternatives employed in large sample statistical theory, in order to obtain explicit expressions for robust score statistics and for asymptotic noncentrality parameters, which can be used to compare the power of different strategies.

Tang and Siegmund (2001) *Biostatistics* **2**, 147-162.

Tang and Siegmund (2002) *Genetic Epidemiology*, **22**, 313-327.

Models

Model A: One locus modeled.

$$Y = \mu + \alpha_x + \alpha_y + \delta_{x,y} + e,$$

where α_a denotes the additive effect of allele a at locus τ, $\delta_{a,b}$ the dominance effect of alleles a and b, and e incorporates both environmental and residual genetic effects.

Notation: $\sigma_A^2 = 2E\alpha_x^2$, $\sigma_D^2 = E\delta_{x,y}^2$; then

$$\sigma_Y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_e^2.$$

Model B: Two additively interacting loci.

$$Y = \mu + \alpha_x + \alpha_y + \delta_{x,y} + \tilde{\alpha}_{\tilde{x}} + \tilde{\alpha}_{\tilde{y}} + \tilde{\delta}_{\tilde{x},\tilde{y}}$$

$$+\gamma_{x,\tilde{x}} + \gamma_{x,\tilde{y}} + \gamma_{y,\tilde{x}} + \gamma_{y,\tilde{y}} + e.$$

Although e is allowed to contain residual genetic effects, we always assume it is uncorrelated with the genetic effects that are explicitly modeled. Both Hardy-Weinberg and linkage equilibrium are assumed.

Additional Notation: $\sigma_{A\tilde{A}}^2 = 4E\gamma_{x,\tilde{x}}^2$.

Parameters

Consider two siblings with phenotypic values Y_1 and Y_2. Segregation parameters are $\mu = E(Y_i)$, $\sigma_Y^2 = \mathrm{Var}(Y_i)$, and $\rho = \mathrm{Corr}(Y_1, Y_2)$.

Linkage parameters:

Model A: Let $\nu = \nu(\tau)$ be the number of alleles (0,1,or 2) inherited identical by descent by the siblings at the modeled QTL. Then

$$\rho_\nu = \mathrm{Corr}(Y_1, Y_2|\nu) = \rho + \sigma_Y^{-2}[(\nu - 1)\alpha_0 + (1/2 - 1_{\{\nu=1\}})\delta_0],$$

where $\alpha_0 = \sigma_A^2/2 + \sigma_D^2/2$ and $\delta_0 = \sigma_D^2/2$. Note that $\delta_0 \leq \alpha_0$.

Model B: Let ν and $\tilde{\nu}$ denote the number of alleles inherited identical by descent at the two modeled QTLs. Then

$$\rho_{\nu,\tilde{\nu}} = \rho + \{(\nu - 1)\alpha_0 + (\tilde{\nu} - 1)\tilde{\alpha}_0$$
$$+(\nu - 1)(\tilde{\nu} - 1)\gamma_0 + \cdots\}/\sigma_Y^2,$$

where

$$\alpha_0 = (\sigma_A^2 + \sigma_D^2)/2 + (\sigma_{A\tilde{A}}^2 + \sigma_{D\tilde{A}}^2)/4 + (\sigma_{A\tilde{D}}^2 + \sigma_{D\tilde{D}}^2)/8,$$

$$\gamma_0 = (\sigma_{A\tilde{A}}^2 + \sigma_{D\tilde{A}}^2 + \sigma_{A\tilde{D}}^2 + \sigma_{D\tilde{D}}^2)/4,$$

and the $\cdots$ indicate five terms associated with dominance deviations, which are frequently small enough to be neglected. Note that $\gamma_0 \leq 2\alpha_0$.

Log Likelihoods and Efficient Score

We consider N pedigrees containing s siblings and let Y denote the vector of phenotypes within the nth pedigree. For simplicity assume that $\mu = 0$. Let $\nu_{ij}(t)$ denote the number of alleles shared identical by descent at the marker locus t by the ith and jth sibs in the nth sibship. Let A_ν denote the $s \times s$ matrix with entries $\nu_{ij} - 1$ for $i \neq j$ and zeroes along the diagonal. Let $\mathbf{\Sigma}_\nu = E(YY'|A_\nu)$, so

$$\mathbf{\Sigma}_\nu = \mathbf{\Sigma} + \alpha_0 A_\nu + \delta_0 D_\nu.$$

The log likelihood for a single QTL at τ is $\ell(\tau, \alpha_0, \delta_0, \rho, \sigma_Y)$ given by

$$\ell = -2^{-1} \Sigma_{n=1}^{N} \left\{ \log |\mathbf{\Sigma}_\nu| + \mathrm{tr} \mathbf{\Sigma}_\nu^{-1} YY' \right\},$$

where $\nu = \nu(\tau)$. The efficient score for the parameter α_0 is by partial differentiation

$$\ell_\alpha = 2^{-1} \Sigma_n \left\{ -\mathrm{tr}(\mathbf{\Sigma}_\nu^{-1} A_\nu) + \mathrm{tr}(\mathbf{\Sigma}_\nu^{-1} A_\nu \mathbf{\Sigma}_\nu^{-1} YY') \right\}.$$

Under the hypothesis $\alpha_0 = 0$ the efficient scores for the segregation parameters ρ, σ_Y and μ are uncorrelated with the efficient scores for the linkage parameters α_0 and δ_0. Should be standardized by

$$\hat{\sigma} = \{ E_0[\ell_\alpha^2 | Y_1, \cdots, Y_N] \}^{1/2}.$$

The robust score statistic at a marker locus t is

$$Z_t = \ell_\alpha(t)/\hat{\sigma}.$$

At a marker locus linked to the trait locus τ, it has asymptotic expected value proportional to

$$(\alpha_0/\sigma_Y^2)\exp(-\beta|t-\tau|).$$

Since we do not know τ, we use

$$Z_{\max} = \max_t Z_t$$

to detect QTL at unknown positions in the genome.

For markers equally spaced at intermarker distance 1cM, for a 22 autosome 3300 cM human genome, a threshold of approximately $z_{\max} = 3.91$ is required for a 0.05 genome-wide false positive probability, and a total noncentrality of $EZ_\tau \approx 5$ gives power of roughly 0.9.

Gene × Covariate Interactions

Assume that

$$Y = \mu + b'w + \sum [\alpha_x + \alpha_y + w'(\gamma_x + \gamma_y)] + e, \quad (1)$$

where the summation extends over all QTLs. Here w is a (vector of) covariates, which we regard as having a fixed value for each individual. Without loss of generality we assume that w has mean zero and covariance matrix the identity. The residual term e is assumed not to depend on w and to be uncorrelated with the genetic effects modeled in (1).

Example. Assume that each individual has a covariate w^* that is either 1, with probability p, or 0, with probability $q = 1 - p$, and that the model is given by

$$Y = \mu + \sum w^*(\gamma_x^* + \gamma_y^*) + e.$$

In this model of interaction, genetic effects are present in individuals with the "right" covariate, but not otherwise.

Observe that $E(Y|w) = \mu + b'w$, and

$$\mathrm{Var(Y|w) = V_{\alpha\alpha} + 2w'V_{\alpha\gamma} + w'V_{\gamma\gamma}w + \sigma_e^2},$$

$$\mathrm{Cov(Y_1, Y_2|w_1, w_2) = V_{\alpha\alpha}/2 + (w_1 + w_2)'V_{\alpha\gamma}/2}$$

$$+ w_1' V_{\gamma\gamma} w_2/2 + r\sigma_e^2,$$

$$\mathrm{Cov(Y_1, Y_2|w_1, w_2}, \nu) = \mathrm{Cov(Y_1, Y_2|w_1, w_2)}$$

$$+[\alpha_0 + (w_1 + w_2)'\beta_0 + w_1'\gamma_0 w_2](\nu - 1),$$

where $\alpha_0 = \sigma_\alpha^2/2$, $\beta_0 = \sigma_{\alpha\gamma}/2$, and $\gamma_0 = \sigma_\gamma^2/2$ are defined in terms of the locus specific variance components, and r is the residual correlation between sibs. The null hypothesis is α_0, γ_0, and β_0 all equal 0.

The log likelihood function for $(\mathbf{Y_1}, \cdots, \mathbf{Y_N})$ is

$$\ell = -2^{-1} \sum_n [\log|\Sigma_{w_n,\nu_n}| + \mathbf{Y_n}'\Sigma^{-1}_{w_n,\nu_n}\mathbf{Y_n}].$$

The full score statistic is three dimensional, with constraints $\alpha_0 \geq 0$, $\gamma_0 \geq 0$. The threshold for a genome scan at 1 cM resolution is $b \approx 4.7$, and a total noncentrality of about 5.63 is required to have power of

0.9. If there is in fact no gene-environment interaction, the higher threshold will lead to a loss of efficiency of about 27%.

Qualitative Traits: Affected Sib Pairs

Let $X_{i,j} = X_{i,j}(\tau)$ denote the indicator that in the jth sib pair i alleles are shared IBD at the locus τ. We suppress the subscript j and the genetic location τ. The log likelihood function is

$$\sum_j \sum_{i=0}^{2} [X_i \log\{1+(i-1)\frac{[\alpha_0 + (w_1 + w_2)\beta_0 + w_1 w_2 \gamma_0]}{E(Y_1 Y_2 | w_1, w_2)}\},$$

so components of the score statistic depend on the nuisance parameters $V_{\alpha\alpha}, V_{\alpha\gamma}$, etc., which cannot generally be estimated from data on affected sib pairs alone.

Table 1: Number of sib pairs for naive and score statistics.

The general model is given by equation (1). We assume that $H^2 = 1/2$ and $h^2 = 1/4$ at the primary locus. The column headed "naïve" is the sample size for the statistic (6). The column headed "two-point" is the sample size for the 3 degrees of freedom score statistic when the covariate is two-valued. The column headed "normal" is the sample size for the same statistic, but obtained under the assumptions that the standardized covariates are bivariate normal within sib pairs with correlation R_w and unit variances, and the noncentrality parameters satisfy relation (5). For this case the noncentrality parameter is determined by simulations with 10^6 samples.

			Naïve		Score Statistic	
p	R_w	ρ	Two-point	Normal	Two-point	Normal
0.75	1.0	0.25	2791	3065	3232	3501
	0.9	0.244	2957	3223	3305	3590
	0.5	0.219	3737	3842	3660	4146
	0.1	0.194	4812	4803	4160	4789
	0	0.188	5128	5109	4302	4892
0.5	1.0	0.25	3090	3298	3267	3708
	0.9	0.238	3497	3659	3383	3883
	0.5	0.188	5720	5647	4315	4682
	0.1	0.138	10364	10494	5995	5712
	0	0.125	12666	12683	6302	6096
0.25	1.0	0.25	3985	3419	3925	3847
	0.9	0.231	4832	4005	4100	4188
	0.5	0.156	10297	8791	6346	6359
	0.1	0.081	33475	32232	11975	6407
	0	0.063	52580	55264	16564	6510