

Mixture Cure Model with an Application to Interval Mapping of Quantitative Trait Loci

Mengling Liu

Division of Biostatistics, School of Medicine
New York University

Bioinformatics, Hangzhou, June 14, 2007

Joint work with Y. Shao (NYU) and W. Lu (NCSU)

Outline

Goal: explain why and how to use the mixture cure models in interval mapping of quantitative traits (time-to-event data) from a population of heterogenous susceptibilities

- Review
- Motivating examples
- Mixture cure models for mapping QTL
- Numerical results
- Remarks

Review of Concepts

Trait (Phenotype):

- Qualitative traits, e.g. diseased/nondiseased;
- Quantitative traits, e.g. blood pressure, BMI.
- Typical distributional assumption: Normal
- Time-to-event data as quantitative traits, e.g. age-of-onset, survival times, etc.
Examples: breast and ovarian cancers, Miki et al. (1994);
Prostate cancer, Carter et al. (1992).
- Incomplete time-to-event data due to censoring

Review of Concepts (Cont')

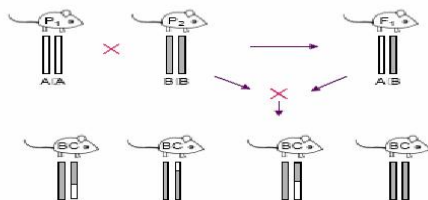
Heterogeneous susceptibilities:

For mapping *time-to-event traits*, challenges include

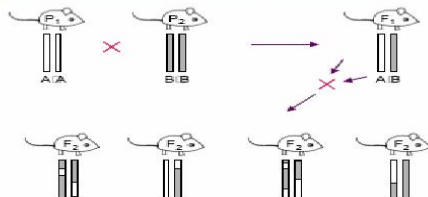
- a subset of the population are not at risk of developing a certain disease of interest.
E.g.: Immune to a virus infection (Boyartchuk et al., 2001), different susceptibilities to v-abl transgene-induced plasmacytoma (Symons et al., 2002).
- the susceptibility status is not completely observable, e.g. subject is censored before observing any event.

Common Experimental Cross Designs:

Backcross experiment



Intercross experiment



Listeria Data

Real data for analysis: (Listeria Data in R)

- $N=116$ intercross mice
- Genotyping 133 markers over 20 chromosomes
- Trait of interest: survival times after infection of *Listeria Monocytogenes*

After the infection, most mice die around 72 hours. However, a proportion of mice survived over 240 hours.

References:

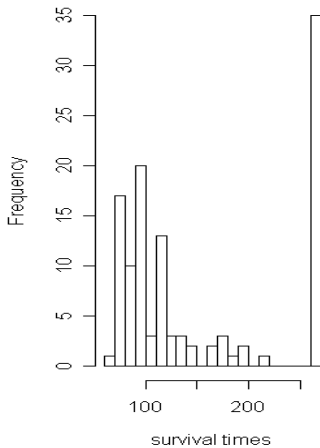
Boyartchuk et al 2001 Nat Genetics 27:259-260

Broman KW 2003 Genetics 163: 1169-1175 (two-part model)

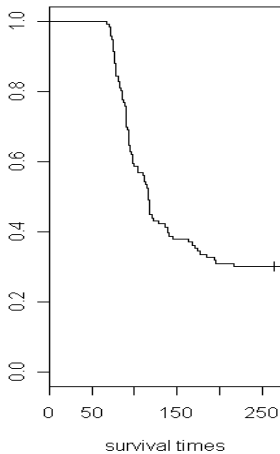
Diao G. et al 2004 Genetics 168: 1689-1698 (parametric PH)

Diao G and Lin D Y 2005 Biometrics 61:789-798 (semi-parametric PH)

Histogram of survival times



K-M curve



- **Observed:**

survival times $Y_i = \min(T_i, C_i)$, $\delta_i = I(Y_i \leq C_i)$,
Marker genotypes M_i ,
Covariates Z_i ,

- **Partially Observed:**

Susceptible indicator η_i
If $\delta_i = 1 \Rightarrow \eta_i = 1$.
If $\delta_i = 0$, it is unknown about η_i .

- **Unobserved:**

Genotypes of QTL, denoted by G_i ,

E.g.: in the intercross design, G_i takes the values of (1, 0), (0, 1) and (0, 0) corresponding to the possible genotypes QQ , Qq and qq .

Survival time of an individual from a population with heterogeneous susceptibilities

$$T_i = \eta_i T_i^* + (1 - \eta_i)\infty$$

Cure Models given G_i

Logistic model for the susceptible indicator:

$$\text{pr}(\eta_i = 1 | G_i, Z_i^*) = \frac{\exp(\gamma' Z_i^* + \gamma_g' G_i)}{1 + \exp(\gamma' Z_i^* + \gamma_g' G_i)},$$

where $Z_i^* = (1, Z_i)$ to include an intercept term.

Survival models for the time-to-event T_i^* of a susceptible subject:
 $f(t | G_i, Z_i; \beta, \beta_g, \theta)$.

log-normal model: $\log T_i^* \sim N(\beta' Z_i^* + \beta_g' G_i, \sigma^2)$,

proportional hazards model:

$$\lambda(t | G_i, Z_i) = \lambda_0(t; \theta) \exp(\beta' Z_i + \beta_g' G_i)$$

At any specific location d on the chromosome, the *complete-data* likelihood function is specified as

$$\begin{aligned}
 L_C(\mu; d) &= \prod_{i=1}^n \left\{ \frac{\exp(\gamma' Z_i^* + \gamma'_g G_i)}{1 + \exp(\gamma' Z_i^* + \gamma'_g G_i)} \right\}^{\eta_i} \left\{ \frac{1}{1 + \exp(\gamma' Z_i^* + \gamma'_g G_i)} \right\}^{1-\eta_i} \\
 &\times \left[\{f(Y_i|Z_i, G_i; \beta, \beta_g, \theta)\}^{\delta_i} \{1 - F(Y_i|Z_i, G_i; \beta, \beta_g, \theta)\}^{1-\delta_i} \right]^{\eta_i} \\
 &\times \text{pr}(G_i|M_i; d). \tag{1}
 \end{aligned}$$

$\text{pr}(G_i|M_i; d)$ is determined by the flanking markers of location d [e.g. Lynch and Walsh 1998].

Hypotheses

- No overall QTL effects,
 $H_0 : \gamma_g = 0$ and $\beta_g = 0$ vs. $H_1 : \gamma_g \neq 0$ or $\beta_g \neq 0$;
- No QTL effects on susceptibility,
 $H_{0\gamma} : \gamma_g = 0$ vs. $H_{1\gamma} : \gamma_g \neq 0$;
- No QTL effects on the survival of susceptible subjects,
 $H_{0\beta} : \beta_g = 0$ vs. $H_{1\beta} : \beta_g \neq 0$.

Testing procedures:

1. Construct LRT profile over chromosomes

- The value of LRT statistic at each location
 - Maximum likelihood estimates of the parameters and restricted maximum likelihood estimates under the Null hypothesis.

2. Determine a genome-wide threshold of certain significance level

- The distribution of the supremum of LRT profile under the Null hypothesis.

Testing procedures:

1. Construct LRT profile over chromosomes

- The value of LRT statistic at each location
 - Maximum likelihood estimates of the parameters and restricted maximum likelihood estimates under the Null hypothesis.

2. Determine a genome-wide threshold of certain significance level

- The distribution of the supremum of LRT profile under the Null hypothesis.

Testing procedures:

1. Construct LRT profile over chromosomes

- The value of LRT statistic at each location
 - Maximum likelihood estimates of the parameters and restricted maximum likelihood estimates under the Null hypothesis.

2. Determine a genome-wide threshold of certain significance level

- The distribution of the supremum of LRT profile under the Null hypothesis.

Testing procedures:

1. Construct LRT profile over chromosomes

- The value of LRT statistic at each location
 - Maximum likelihood estimates of the parameters and restricted maximum likelihood estimates under the Null hypothesis.

2. Determine a genome-wide threshold of certain significance level

- The distribution of the supremum of LRT profile under the Null hypothesis.

Testing procedures:

1. Construct LRT profile over chromosomes

- The value of LRT statistic at each location
 - Maximum likelihood estimates of the parameters and restricted maximum likelihood estimates under the Null hypothesis.

2. Determine a genome-wide threshold of certain significance level

- The distribution of the supremum of LRT profile under the Null hypothesis.

Maximum likelihood estimations: EM algorithm

Because of missing data, EM algorithm is natural choice.
Generally intensive computations, here with some advantages:

- Because the missing data η_i and G_i are discrete data, calculations of the expectations of η_i , G_i and $\eta_i G_i$ suffice for the E-step.
- The components of the *complete-data* likelihood function (1) involve distinct parameters, hence the maximization reduces to low dimensional calculation.

Genome-wide threshold: a resampling method

Multiple testing issue: LRT are conducted at each location over the chromosomes using the same trait data - - the test statistics are dependent.

- Bonferroni correction
- Theoretical asymptotic distribution of the supremum of testing statistics (Rebai, 1995)
- A resampling approach
(Diao et al. 2004; Lin et al. 2004; Zou et al., 2004; Lin 2005)

The resampling method

Using the asymptotic equivalence between the LRT and the score test under the null model, the resampling approach computes the empirical threshold for LRT by generating a large number of randomly perturbed score test statistics (e.g., Lin D Y 2005, Bioinformatics, 21: 781-787).

Resampling method

Score functions

$$\mathbf{U}(\boldsymbol{\mu}; d) = \sum_{i=1}^n \mathbf{U}_i(\boldsymbol{\mu}; d) \doteq \sum_{i=1}^n \partial l_i(\boldsymbol{\mu}; d) / \partial \boldsymbol{\mu}'$$

For example, to test $H_0 : \boldsymbol{\gamma}_g = 0$ and $\boldsymbol{\beta}_g = 0$, the corresponding scores are

$$\mathbf{U}_{\boldsymbol{\gamma}_g \boldsymbol{\beta}_g}(\boldsymbol{\mu}; d) \doteq \begin{pmatrix} \sum_{i=1}^n \partial l_i(\boldsymbol{\mu}; d) / \partial \boldsymbol{\gamma}_g' \\ \sum_{i=1}^n \partial l_i(\boldsymbol{\mu}; d) / \partial \boldsymbol{\beta}_g' \end{pmatrix}.$$

The efficient scores are

$$\hat{\mathbf{U}}_{\gamma_g \beta_g}(\tilde{\boldsymbol{\mu}}; d) = \begin{pmatrix} \mathbf{U}_{\gamma_g}(\tilde{\boldsymbol{\mu}}; d) \\ \mathbf{U}_{\beta_g}(\tilde{\boldsymbol{\mu}}; d) \end{pmatrix} - \mathbf{I}_{\gamma_g, \beta_g | \mu}^{-1} \begin{pmatrix} \mathbf{U}_{\gamma}(\tilde{\boldsymbol{\mu}}; d) \\ \mathbf{U}_{\beta}(\tilde{\boldsymbol{\mu}}; d) \\ \mathbf{U}_{\theta}(\tilde{\boldsymbol{\mu}}; d) \end{pmatrix}.$$

Randomly perturbed score statistics are:

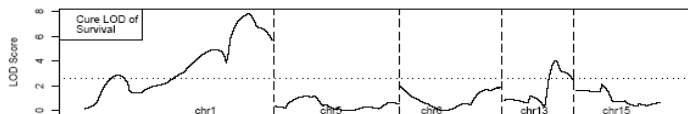
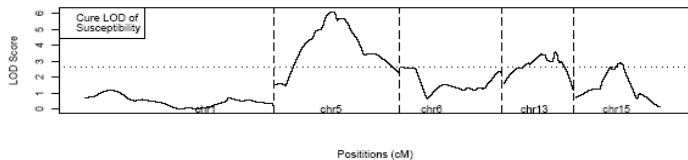
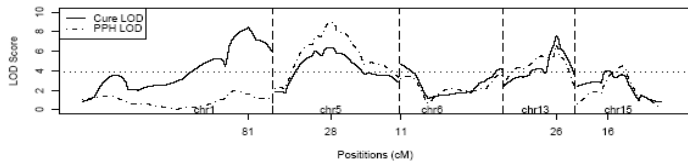
$$\hat{\mathbf{U}}_{\gamma_g \beta_g}^*(d) = \sum_{i=1}^n \hat{\mathbf{U}}_{\gamma_g \beta_g, i}(\tilde{\boldsymbol{\mu}}; d) X_i, \quad X_i \sim N(0, 1).$$

Resampled test statistics:

$$W^* = \sup_d \left\{ \frac{1}{n} \hat{\mathbf{U}}_{\gamma_g \beta_g}^{*T}(\tilde{\boldsymbol{\mu}}; d) \hat{\mathbf{V}}^{-1}(d) \hat{\mathbf{U}}_{\gamma_g \beta_g}^*(\tilde{\boldsymbol{\mu}}; d) \right\} \quad (2)$$

Analysis of Listeria data

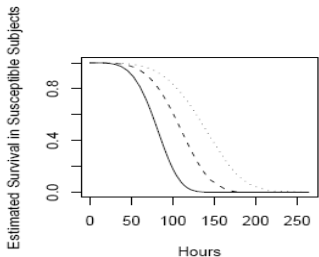
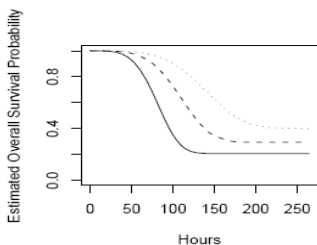
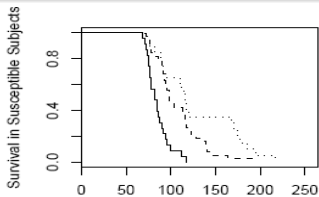
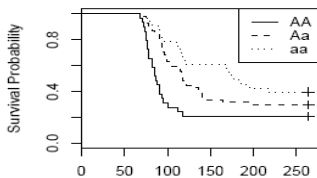
- Overall effects
 $H_0 : \gamma_g = 0$ and $\beta_g = 0$ vs. $H_1 : \gamma_g \neq 0$ or $\beta_g \neq 0$;
- Susceptibility effect
 $H_{0\gamma} : \gamma_g = 0$ vs. $H_{1\gamma} : \gamma_g \neq 0$;
- Survival effect
 $H_{0\beta} : \beta_g = 0$ vs. $H_{1\beta} : \beta_g \neq 0$.



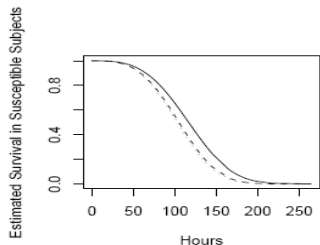
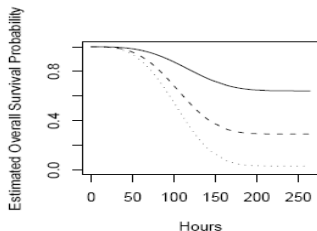
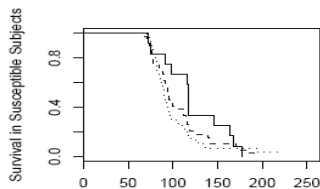
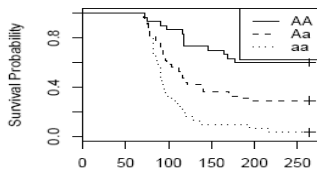
Next we examine chromosomes 1 and 5 more closely to interpret the differences between our results and findings using the PPH method.

QTL	Chrom 1	Chrom 5
Susceptibility	No	Yes
Survival	Yes	No
PPH	No	Yes

Chromosome 1: D1M355



Chromosome 5: Marker D5M357



- **H₀**: $\gamma_{g1} = 0, \gamma_{g2} = 0$ and $\beta_{g1} = 0, \beta_{g2} = 0$
- **H_{1a}**: $\gamma_{g1} = 0, \gamma_{g2} = 0$ and $\beta_{g1} = 0.5, \beta_{g2} = -0.5$
(Chromosome 1)
- **H_{2a}**: $\gamma_{g1} = 0.75, \gamma_{g2} = -0.75$ and $\beta_{g1} = 0.0, \beta_{g2} = 0.0$
(Chromosome 5)
- **H_{3a}**: $\gamma_{g1} = 0.25, \gamma_{g2} = -0.25$ and $\beta_{g1} = 0.5, \beta_{g2} = -0.5$
(Chromosome 13)

Simulated Type I errors and powers

	H_0		H_{1a}		H_{2a}		H_{3a}	
	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%
C	6.5	1.0	85.0	67.7	82.2	62.8	91.7	79.8
P	18.5	6.7	22.4	10.4	95.6	90.2	53.5	35.2
C	6.3	1.4	87.4	72.5	86.9	71.3	92.9	84.3
P	16.3	4.7	26.2	10.7	95.1	89.2	56.3	37.5

Remarks

- biological supports
- long enough follow-up time
- Model identifiability
- Further research worthwhile

Acknowledgements

- Yongzhao Shao, New York University
- Wenbin Lu, North Carolina State University
- NIH, NCI, and NSF