Genomics 104 (2014) 472–476

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

Cox regression model for dissecting genetic architecture of survival time

Dan Jiang [a,1], Hongwei Wang [b,1], Jiahan Li [c], Yang Wu [d], Ming Fang [a], Runqing Yang [e,*]

[a] Life Science College Heilongjiang Bayi Agricultural University, Daqing 163319, People's Republic of China
[b] Fishery Technical Extension Station, Beijing Daxing Animal Health Supervisory Commission, Beijing 102600, People's Republic of China
[c] Applied and Computational Mathematics and Statistics, University of Notre Dame, IN 46637, USA
[d] Institute of Animal Science, Chinese Academy of Agricultural Science, Beijing 100193, People's Republic of China
[e] Research Centre for Aquatic Biotechnology, Chinese Academy of Fishery Sciences, Beijing 100141, People's Republic of China

ARTICLE INFO

Article history:
Received 13 April 2014
Accepted 3 October 2014
Available online 12 October 2014

Keywords:
Survival time
QTL
LASSO
Cox regression model
Partial likelihood algorithm

ABSTRACT

Common quantitative trait locus (QTL) mapping methods fail to analyze survival traits of skewed normal distributions. As a result, some mapping methods for survival traits have been proposed based on survival analysis. Under a single QTL model, however, those methods perform poorly in detecting multiple QTLs and provide biased estimates of QTL parameters. For sparse oversaturated model used to map survival time loci, the least absolute shrinkage and selection operator (LASSO) for Cox regression model can be employed to efficiently shrink most of genetic effects to zero. Then, a few non-zero genetic effects are re-estimated and statistically tested using the standard maximum Cox partial likelihood method. Simulation shows that the proposed method has higher statistic power for QTL detection than that of the LASSO for logarithmic linear model or the interval mapping based on Cox model, although it somewhat underestimates QTL effects. Especially, computational speed of the method is very fast. An application of this method illustrates mapping main effect and interacting QTLs for heading time in the North American Barley Genome Mapping Project.

1. Introduction

Survival traits measured as time to event usually come with two features: skewed distribution with heavier right tail and censored observations [1]. By incorporating survival analysis theory into the traditional quantitative trait loci (QTL) mapping frameworks, the QTL mapping for survival traits has been introduced to efficiently locate survival time loci, allowing a better understanding of genetic architectures underlying survival traits. Broman [2] considered a cure-rate model that treated the mice alive at the end of the study as cured ones when mapping QTL of the time to death of bacterial infection. The survival time was modeled by log-normal distribution. A Cox proportional hazards (PH) model was proposed to characterize the effects of the QTL genotype on failure time [3], where model parameters and computed LOD scores were estimated by a variant of EM algorithm [4]. Diao et al. [5] also used a Cox PH model with a Weibull baseline hazard function to locate QTLs, and then developed efficient likelihood-based inference procedures. These two Cox PH models belong to parametric models for mapping survival time loci, due to the component of estimating baseline hazard functions. Along this line, parametric Cox PH model for mapping QTLs of heading time in rice was optimized by Luo et al. [6], where the best baseline hazard distribution was selected from six commonly used survival distributions. Other than parametric algorithms, Diao and Lin [6] developed semi-parametric statistical methods for Cox PH model to search for survival trait loci. Without need to estimate baseline hazard functions, Fang [7] proposed further a simple and efficient non-parametric approach to estimate QTL parameters through partial likelihood function. Using simulated data with different structures, Moreno et al. [8] systematically compared the parametric model based on Weibull distribution, semi-parametric model, and classical interval mapping based on the normal distribution. Additionally, accelerated failure time model was also used to model the genetic effects of QTL on survival traits [9–11].

All these methods of mapping survival traits are developed for interval mapping, where the whole genome is scanned, but only one QTL is analyzed at each time. For survival traits controlled by multiple QTLs, this mapping strategy is suboptimal due to the existence of linked QTLs. Over the past decade, QTL mapping methods have been developed to simultaneously analyze multiple QTLs for normal traits. Mapping multiple QTLs by either non-Bayesian or Bayesian methods is in fact a model selection problem about selecting QTLs from a large number of genetic loci over the entire genome. Although Bayesian shrinkage analysis greatly facilitates modeling multiple QTLs and the shrinkage estimation, full Bayesian shrinkage mapping is practically infeasible due to high computational cost [12–17]. As an equivalent strategy to Bayesian shrinkage estimation with double exponential priors for regression coefficients [18], the LASSO [19] is widely used to solve for the oversaturated linear model. Most recently, Liu etc. [20] employed

* Corresponding author. Fax: +86 10 68670701.
E-mail address: runqingyang@sjtu.edu.cn (R. Yang).
[1] These authors have contributed equally to this work.

http://dx.doi.org/10.1016/j.ygeno.2014.10.002
0888-7543/

Table 1
Mean estimates and standard deviations (in parentheses) of QTL positions obtained with three mapping methods for the simulated datasets.

Sample size	Method	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
	True position	23	56	49	94	69	35	93	80	27	79
150	Cox-LASSO	23.9(2.3)	56.6(1.8)	50.0(2.5)	94.9(2.6)	70.3(3.0)	34.5(2.3)	94.1(2.0)	80.7(2.6)	26.4(2.6)	77.9(2.3)
	Gau-LASSO	23.2(3.1)	56.8(1.3)	49.8(2.4)	94.0(3.0)	69.5(2.7)	34.6(2.5)	94.5(1.9)	81.0(1.3)	26.2 (2.2)	80.1(1.9)
	Cox-LS	28.5(5.5)	74.0(5.3)	54.3(8.2)	90.2(7.2)	–	32.17(6.7)	98.5(3.5)	81.3(6.6)	–	–
300	Cox-LASSO	24.2(1.6)	56.6(1.3)	49.9(2.3)	95.1(2.1)	70.0(2.8)	34.7(2.3)	94.4(1.6)	80.8(1.7)	26.0(2.2)	79.9(2.4)
	Gau-LASSO	23.9(2.7)	56.6(1.6)	49.6(2.5)	94.9(2.8)	70.5(2.9)	34.8(2.5)	94.0 (2.1)	80.9(2.3)	26.5(2.3)	79.7(2.6)
	Cox-LS	29.3(4.9)	74.1(4.0)	55.7(7.1)	90.3(6.3)	–	32.7(5.9)	95.7(2.8)	80.9(5.0)	22.8(7.9)	–

a fast LASSO with coordinate descent algorithm [21] to successfully search for the QTLs of normal traits. To date, however, no multiple-QTL mapping method is reported for survival traits. In this study, a little modified LASSO for Cox regression model [22], as a non-parametric approach, is used to efficiently analyze sparse oversaturated model for mapping survival loci. Then for the subset of selected loci, their genetic effects are unbiasedly estimated and statistical tests are carried out using the standard maximum Cox partial likelihood method [24]. Simulation is conducted to investigate statistical efficiency of the mapping method proposed. A dataset from the North American Barley Genome Mapping Project is analyzed to map the QTLs for heading times.

2. Method

2.1. Multiple-QTL proportional hazards model

Assume that n individuals from a backcross (BC) population are observed for a survival trait, and are genotyped for m co-dominant markers with known genetic linkage map. To map the QTLs of the observed survival trait, entire genome is evenly divided by k loci that are 1 or 2 cM away from each other, and the candidates of QTLs include the genotyped markers as well as ungenotyped loci between markers. Our genetic design implies that there are two genotypes at each locus on chromosome, denoted by QQ and Qq. With the proportional hazard model, the effects of the QTL candidates can be formulated as

$$\lambda(t_i) = \lambda_0(t_i)\exp\left(\sum_{j=1}^{p} b_j x_{ij}\right) \tag{1}$$

where t_i is survival time for the ith individual, $\lambda(t_i)$ is the hazard function evaluated at time t_i, $\lambda_0(t)$ is a common baseline hazard function, $p = m + k$ is the number of the QTL candidates, b_j is additive effect of the jth QTL candidate, and x_{ij} is the indicator variable of the jth QTL candidate for individual i determined by QTL genotypes. If the QTL candidate is the genotyped marker, x_{ij} is defined as $+1$ for QQ and -1 for Qq, respectively. Otherwise for any locus between two markers, the indicator variable can be estimated by its expectation conditional on flanking markers, according to least square method by Haley and Knott [23]. Specifically, the expectation is calculated as

$$E\left(x_{ij}\right) = (+1)p(QQ) + (-1)p(\mathrm{Qq}) = p(QQ) - p(\mathrm{Qq})$$

where $p(QQ)$ and $p(Qq)$ are probabilities for two QTL genotypes estimated by flanking markers.

2.2. Shrinkage estimation of genetic effects

Usually survival data are censored because of random loss to follow-up, failures from competing causes, or limited duration of the experiment. Besides, base hazard function in model (1) is unknown in general. To address these issues, Cox partial likelihood algorithm was developed to handle censoring problem without estimating baseline hazard function. Based on model (1), the Cox's partial likelihood [24] can be written as

$$PL = \prod_{i=1}^{n} \frac{\exp\left(\sum_{j=1}^{p} b_j x_{l(i)j}\right)}{\sum_{r=1}^{l(i)} \exp\left(\sum_{j=1}^{p} b_j x_{rj}\right)}, \tag{2}$$

where $l(i)$ for $l = 1, 2, \ldots, t$ denotes the index at lth survival time in the increasing list of unique survival times for ith individuals.

By defining $\boldsymbol{\mu} = \partial \log L/\partial \boldsymbol{\eta}$, $\mathbf{A} = -\partial^2 \log L/\partial \boldsymbol{\eta}\boldsymbol{\eta}^{\mathbf{T}}$ and $\mathbf{z} = \boldsymbol{\eta} + \mathbf{A}^{-1}\boldsymbol{\mu}$ with $\boldsymbol{\eta} = \mathbf{b}'\mathbf{x}$, the logarithm of log-partial likelihood is approximated by a two term Taylor series, as follows

$$\log(PL) = \left(\mathbf{z} - \mathbf{b}^T\mathbf{x}\right)^T\left(\mathbf{z} - \mathbf{b}^T\mathbf{x}\right). \tag{3}$$

where, $\mathbf{b} = [b_1, b_2, \cdots, b_p]$ and $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^T$ with $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{ip}]$ Decomposing matrix $\mathbf{A}$ into $\mathbf{V}^T\mathbf{V}$ by Cholesky decomposition, we transform objective function (3) into

$$\log(PL) = \left(\mathbf{y} - \mathbf{b}^T\mathbf{x}'\right)^T\left(\mathbf{y} - \mathbf{b}^T\mathbf{x}'\right). \tag{4}$$

where $\mathbf{y} = \mathbf{Vz}$ and $\mathbf{x}' = \mathbf{Vx}$. In QTL mapping with linkage analysis, the number of estimated parameters is far greater than the sample size. Moreover, the number of QTLs with non-zero genetic effects is very limited. The LASSO with coordinate descent algorithm [21,25] can be therefore employed to efficiently shrink most of genetic effects to be zero by minimizing

$$\left(\mathbf{y} - \mathbf{b}^T\mathbf{x}'\right)^T\left(\mathbf{y} - \mathbf{b}^T\mathbf{x}'\right) + \lambda|\mathbf{b}|. \tag{5}$$

where λ is a tuning parameter, which will be chosen through cross validation. Since both y and x' are the function of parameter b, iterations are required in solving the model parameters with the LASSO.

Table 2
Mean estimates and standard deviations (in parentheses) of QTL effects obtained with three mapping methods for the simulated datasets.

Sample size	Method	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
	True effect	1.80	2.30	1.02	1.40	−0.52	1.00	−0.95	1.55	0.65	−1.10
150	Cox-LASSO	1.17(0.25)	1.52(0.3)	0.68(0.15)	0.91(0.21)	−0.48(0.08)	0.68(0.17)	−0.6(0.13)	0.98(0.2)	0.5(0.09)	1.17(0.25)
	Gau-LASSO	0.55(0.21)	0.75(0.15)	0.41(0.21)	0.47(0.13)	−0.19(0.19)	0.39(0.11)	0.17(0.09)	0.22(0.07)	0.23(0.1)	0.29(0.07)
	Cox-LS	0.69(0.12)	0.8(0.11)	0.4(0.07)	0.42(0.09)	–	0.33(0.04)	0.31(0.05)	0.39(0.07)	–	–
300	Cox-LASSO	1.14(0.18)	1.42(0.2)	0.62(0.12)	0.86(0.14)	−0.36(0.07)	0.61(0.11)	−0.6(0.11)	0.97(0.14)	0.37(0.08)	−0.68(0.12)
	Gau-LASSO	0.61(0.13)	0.68(0.09)	0.46(0.18)	0.36(0.13)	−0.27(0.21)	0.38(0.08)	−0.29(0.25)	0.39(0.13)	0.27(0.17)	0.37(0.14)
	Cox-LS	0.68(0.08)	0.79(0.08)	0.37(0.06)	0.41(0.06)	–	0.22(0.03)	0.22(0.02)	0.35(0.06)	0.2(0.01)	–