Journal of Statistical Planning and Inference 139 (2009) 978 – 989

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A robust QTL mapping procedure☆

Fei Zou[a,*], Lei Nie[b], Fred. A. Wright[a], Pranab K. Sen[a]

[a]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA
[b]Georgetown University Medical Center, Washington, DC 20057, USA

ARTICLE INFO

Article history:
Received 1 September 2006
Received in revised form
9 July 2007
Accepted 8 June 2008
Available online 25 June 2008

Keywords:
GEE
Generalized least squares estimate
Quantitative trait
Weighted least squares estimate
Wilcoxon–Mann–Whitney statistic

ABSTRACT

In quantitative trait linkage studies using experimental crosses, the conventional normal location-shift model or other parameterizations may be unnecessarily restrictive. We generalize the mapping problem to a genuine nonparametric setup and provide a robust estimation procedure for the situation where the underlying phenotype distributions are completely unspecified. Classical Wilcoxon–Mann–Whitney statistics are employed for point and interval estimation of QTL positions and effects.

Published by Elsevier B.V

1. Introduction

Genetic mapping of quantitative trait loci (QTL) has fundamental importance in revealing the genetic basis of phenotypic differences (Belknap et al., 1997; Haston et al., 2002; Wang et al., 2003). In plants and laboratory animals, backcross or F2 intercross populations are widely used for mapping quantitative traits (see Lynch and Walsh, 1998, for details). In QTL mapping, the basic problems are to test the existence of one or more QTLs, and to estimate the QTL map position and effect if there is evidence of linkage to a chromosomal region. QTL mapping methodologies, including the single marker t-tests (Sax, 1923) and likelihood interval mapping (Lander and Botstein, 1989; Haley and Knott, 1992; Kruglyak and Lander, 1995), have traditionally relied on parametric assumptions. In Kruglyak and Lander (1995), a nonparametric approach has been explored for testing linkage, but cannot produce QTL confidence intervals or specify effect sizes. Zou et al. (2002) proposed a semiparametric model that specifies an exponential tilt relationship between phenotype densities for different genotypes at the QTL.

In standard parametric linkage scans, the (profile) likelihood ratio test statistic is calculated for each position, and the maximum likelihood estimate (MLE) used as a point estimate for the QTL position. A difficulty in the use of the MLE in this setting is that it may exhibit nonstandard asymptotic behavior, depending on the asymptotic regime used (Kong and Wright, 1994). For realistic sample sizes and marker densities, the consequences are that the MLE of the QTL position might not be efficient and accurate confidence intervals are not readily available from the profile likelihood in the vicinity of the MLE. However, the reporting of plausible intervals is important (Flaherty et al., 2003). A number of approximate methods have been described, including LOD-drop intervals (Lander and Botstein, 1989), which may have unreliable coverage (Dupuis and Siegmund, 1999), and formulae in Darvasi and Soller (1997) for 95% confidence intervals based on their extensive simulations. Other computation-intensive

☆ This work was supported by National Institutes of Health (NIH) Grant MH070504 to F.Z.
* Corresponding author. Fax: +1 919 966 3804.
E-mail address: fzou@bios.unc.edu (F. Zou).

0378-3758/$ - see front matter Published by Elsevier B.V
doi:10.1016/j.jspi.2008.06.009

approaches include bootstrapping (Visscher et al., 1996) and the method of Mangin et al. (1994), which requires simulation to obtain an asymptotic distribution of a test statistic.

For backcross population, Kearsey and Hyne (1994), Wu and Li (1994, 1996) proposed a multipoint mapping by modeling the mean phenotype difference between two genotype groups at a marker as a function of the recombination frequency between that locus and a putative QTL. Their approach jointly uses the information of every marker on a chromosome. Instead of working on the profile likelihood across genomic positions, they proposed several least squares methods to estimate the QTL position and its effect simultaneously. Therefore, both the detection of the QTL and its position (with correct confidence intervals) are done simultaneously. Liang et al. (2001a, b) proposed a similar multipoint mapping of complex diseases for affected sib pair studies. The method carries out a parametric inference procedure to locate a susceptibility gene, using generalized estimating equations (GEE) to model the expected identical by descent (IBD) allele sharing on all genotyped markers at once with the ultimate goal of locating the susceptible gene more robustly.

The objectives of the current study are to extend the procedure of Kearsey and Hyne (1994) and Wu and Li (1994, 1996) to relax stringent model assumptions on the underlying phenotype distributions. Our proposed method differs from the approach of Kearsey and Hyne (1994) and Wu and Li (1994, 1996) in several ways. First, they considered mean phenotype differences at each marker while we calculate the rank difference of phenotype at each marker, which as shown later, increases mapping efficiency dramatically. Second, we directly express the covariance matrix analytically in terms of several meaningful parameters, while Kearsey and Hyne (1994) and Wu and Li (1994, 1996) did not. To simplify the illustration, we describe the method for backcross populations as done in Kearsey and Hyne (1994) and Wu and Li (1994, 1996).

The paper is organized as described below. Section 2 formulates the estimation procedures. Simulation studies in Section 3 demonstrate the properties of the proposed method and its utility. The discussion section describes extensions and suggestions for future work.

2. Methodology

Consider a backcross experiment with n genotyped individuals. For the inbred parental lines P1 and P2, we label an allele from P1 as m and that from P2 as M. The hybrid F1 individuals are completely heterozygous, with genotype Mm at each locus. Crossing F1 with one of the parental lines (say P2) generates a backcross population in which a subject's genotype has an equal probability $\frac{1}{2}$ of being either MM or Mm at every locus. For each individual i, $i = 1, \ldots, n$ where n is the total number of observations, the observed data consist of a quantitative trait value y_i and genotypes at K molecular markers $\{M_{ik}\}_{k=1}^{K}$. Details of the QTL experiments can be found in Lynch and Walsh (1998).

Suppose there exists a putative QTL at position μ on the genome. Further assume that the quantitative traits for individuals with QTL genotypes Qq and QQ follow distribution functions F and G, respectively. F and G will differ, for otherwise locus μ would not be considered a QTL. The quantity $\int F\,\mathrm{d}G$ is often used to measure the difference between F and G, and is interpretable as the probability that a random value from G exceeds a random value from F. It is also the area under the receiver–operator characteristic curve (AUC) comparing the two distributions, and is invariant to increasing monotone transformations. It is conceptually helpful to use the rescaled parameter $\delta = 2\int F\,\mathrm{d}G - 1$. Note that $|\delta|$ ranges from 0 (when $F = G$) to 1 (where F and G are completely nonoverlapping with each other).

For the QTL mapping problem, we note that the QTL position μ is unknown and the only genetic information consists of the marker genotypes, from which the genetic distances of the markers are estimated. If the recombination frequency between a particular marker locus $k \in \{1, \ldots, K\}$ and the QTL is θ_k, then given its kth marker genotype M_{ik}, the conditional phenotype distributions of individual i, will be $y_i|(M_{ik}=Mm) \sim \tilde{F}_k(y)=(1-\theta_k)F(y)+\theta_k G(y)$ and $y_i|(M_{ik}=MM) \sim \tilde{G}_k(y)=\theta_k F(y)+(1-\theta_k)G(y)$. Here θ_k is a function of μ, and by definition of the conditional distributions we have

$$\tilde{F}_k - \tilde{G}_k = (1-2\theta_k)(F-G).$$

This equation drives our ability to detect linkage nonparametrically, as $\tilde{F}_k$ and $\tilde{G}_k$ will exhibit their greatest difference for the marker closest to the QTL, and will show no difference at markers unlinked to the QTL (where $\theta_k = 0.5$). That is, the phenotypic differences between the two marker genotype groups will decrease as the marker and QTL distance increases. Specifically, when marker k is the QTL itself, $\theta_k = 0$ and $\tilde{F}_k = F$, $\tilde{G}_k = G$ (although the QTL need not be at a marker location). At the other extreme of no linkage, $\theta_k = \frac{1}{2}$ and $\tilde{F}_k = \tilde{G}_k = \frac{1}{2}(F+G)$.

For testing the existence of a QTL, we have the following two hypotheses:

H_0: There exist no QTLs, that is, $F = G$ for all positions on the chromosome vs.

H_A: There exists a QTL, that is, $F \neq G$ for μ somewhere on the chromosome.

At marker k, we divide the n individuals into $n_{1,k}$ individuals with genotype MM and $n_{2,k} = n - n_{1,k}$ individuals with genotype Mm. Let $y_{(1,1)}, \ldots, y_{(1,n_{1,k})}$ and $y_{(2,1)}, \ldots, y_{(2,n_{2,k})}$ be the corresponding trait values of those $n_{1,k}$ and $n_{2,k}$ individuals. We propose the following approach for estimation and testing. Define the Wilcoxon–Mann–Whitney (WMW) statistic at the kth marker as

$$U_{k,n} = \frac{1}{n_{1,k}n_{2,k}} \sum_{i=1}^{n_{1,k}} \sum_{j=1}^{n_{2,k}} \phi(y_{(1,i)}; y_{(2,j)}), \quad k = 1, 2, \ldots, K,$$