Using survival data in gene mapping : using survival data in genetic linkage and family-based association analysis

Callegaro, A.

Appendix A

Arthur - Weighted Allele Sharing Methods for Genetic Linkage Analysis

Abstract

Motivation: *Recently, a number of new score statistics have been proposed for genetic linkage mapping (Callegaro et al., 2009, 2010; Lebrec et al., 2004; Lebrec and van Houwelingen, 2007). These score tests are a computationally faster, locally more powerful, and more robust alternative to likelihood ratio tests. We have developed Arthur, a package to compute these score statistics, which are classical allele sharing statistics with particular weights.*

Availability: *The Arthur package is a collection of compiled exe files (*`ibd.variance`*,* `ARP.weight`*,* `QTL.weight` *,* `GLM.weight`*,* `AAO.weight`*,* `score.test`*) for the use in Windows. Package and documentation are freely available at http://www.msbi.nl/Genetics.*

A.1 Introduction

Although many traits are heritable, identification of responsible genes appears to be a challenge. Recently new loci have been discovered by genome wide association studies, but they explain only a part of the genetic variation and a lot remains to be recovered. Follow up of chromosomal areas with linkage signals in families by using extensive sequencing is a way to find the responsible genetic variants.

Several score statistics have been proposed for linkage analysis which are weighted allele sharing statistics with particular weight functions (Callegaro et al., 2009, 2010; Lebrec et al., 2004; Lebrec and van Houwelingen, 2007). These statistics are derived from statistical models, i.e. generalized linear mixed models and frailty models for survival data. To compute the weights, the user has to specify certain population parameters. For many traits these parameters are

known from twin studies. For N pedigrees, the weighted statistic is given by

$$Z_w = \frac{\sum_{i=1}^{N} vec(W_i)' vec(\hat{\Pi}_i - 2\Phi_i)}{\sqrt{\sum_{i=1}^{N} vec(W_i)' \text{var}_0(\hat{\Pi}_i) vec(W_i)}}, \tag{A.1}$$

where W_i is the weight matrix; $\hat{\Pi}_i$ is the matrix of pairwise estimated proportions of alleles shared identical by descent (IBD) and Φ_i is the matrix of kinship coefficients. The operator $vec(A)$ places the n columns of the $m \times n$ matrix A into a vector of $mn \times 1$. In the case of uncertain IBD status, the variance of the proportion of alleles shared IBD $\text{var}_0(\hat{\Pi}_i)$ can be estimated by simulations (Lebrec et al., 2004).

A.2 Methods

In order to compute the score statistic in equation (1) Arthur package uses three steps. In the first step the variance of the IBD $\text{var}_0(\hat{\Pi}_i)$ (`ibd.variance`) is computed by using Merlin (Abecasis et al., 2002). In the second step the weight matrix W_i is computed. For various types of outcome variables programs are available to compute the weight matrices (`ARP.weight`, `QTL.weight`, `GLM.weight`, `AAO.weight`). Finally all the available information is combined to compute the score statistics (`score.test`). In the next section we will describe these steps in more detail.

Step 1: IBD variance computation

`ibd.variance`: The program uses Merlin (Abecasis et al., 2002) to estimate the proportion of alleles IBD $\hat{\Pi}_i$ and its variance. Input files are in Merlin format. For the estimation of the variance $\text{var}_0(\hat{\Pi}_i)$ Arthur uses multipoint simulations. Specifically, B data-sets are simulated using the Merlin option `--simulate`. Let $\hat{\Pi}_{i0}^b$ denote the proportion of IBD estimated on the b-th, $(b = 1, ..., B)$ simulated data-set. The variance is $\text{var}_0(\hat{\Pi}_i) = \sum_{b=1}^{B} (\hat{\Pi}_{i0}^b - \sum_{b=1}^{B} \hat{\Pi}_{i0}/B)^2/B$.

Estimation of the variance can be time consuming in the case of moderate size pedigrees or large numbers of markers. However computation is only once and the variances can be used for testing of linkage for various traits.

Step 2: Weight computation

Affected relative pairs

`ARP.weight`: Arthur assigns weights equal to one to affected relative pairs and zero otherwise.

Quantitative Traits

`QTL.weight`: For quantitative traits, Arthur computes the weight function proposed by several authors, (e.g., Tang and Siegmund (2001), Lebrec et al. (2004)). Let y_i, μ_i, and Σ_i be the vector of phenotypes, its expectation and the variance-covariance matrix of the phenotype for the ith family. The weight matrix is given by

$$W_i = \Sigma_i^{-1}(y_i - \mu_i)(y_i - \mu_i)'\Sigma_i^{-1} - \Sigma_i^{-1}. \tag{A.2}$$

To compute W_i, the user has to specify the population mean (μ), its variance (σ^2) and the correlation (ρ) between sibling pairs.

Categorical and count data

`GLM.weight`: For the generalized linear mixed model, a score statistic was derived by using a quasi-likelihood approach Lebrec and van Houwelingen (2007). The weight function is similar to equation (A.2), with a slightly different parametrization of the variance-covariance matrix. The weight can also be adjusted for covariates with known effect sizes at the population level. For survival data, this program can be used when a log-normal frailty model is assumed. For affected relative pairs and various family sizes, this function can be used to weight different family sizes according to the correlation in the population, i.e. when the correlation is high affected pairs from a large family will be assigned less weight than affected pairs from small families while for small correlation there is not much difference between these weights.

Survival data

`AAO.weight`: When age at onset for affected and age at censoring for unaffected subjects are available, Arthur can be used to perform a linkage analysis for survival outcomes. Several weight functions are available namely assuming no residual correlation, assuming a correlated frailty model and including phenotypic information of the parents (Callegaro et al., 2009, 2010). Note that for large pedigrees either the composite likelihood or a quasi-likelihood approach (`GLM.weight`) should be used to relieve the computational burden (Callegaro et al., 2009).

Step 3: score test computation

`score.test`: At the final step, Arthur combines the quantities computed (and saved) in the previous steps ($\text{var}_0(\hat{\Pi}_i)$ and W_i) and the weighted score statistic of equation (A.1) and corresponding LOD-score are computed.

We separated step 2 and step 3 in order to provide the user complete flexibility. By using a separate weight file, Arthur can also use weight files specified by the user and it computes any kind of weighted NPL statistic.

Example

As an example, we present the results of an analysis on breast cancer data (Callegaro et al., 2009). Arthur was applied to 55 affected sibling pairs with known age at onset and without any mutations in BRCA1 and BRCA2 described (see Oldenburg et al. (2008)). Figure A.1 shows the LOD-scores derived by using three different weight functions: constant weights, age at onset weights assuming null variance of the random effect and using age specific incidence of breast cancer for the Dutch population, and age at onset weights using population parameters from twin studies (correlation of $\rho = 0.125$ and variance of $\sigma^2 = 25$ for the gamma distributed frailties) (Callegaro et al., 2009). Adjusting for age at onset increased the evidence of linkage at chromosome 9 around 82 cM.

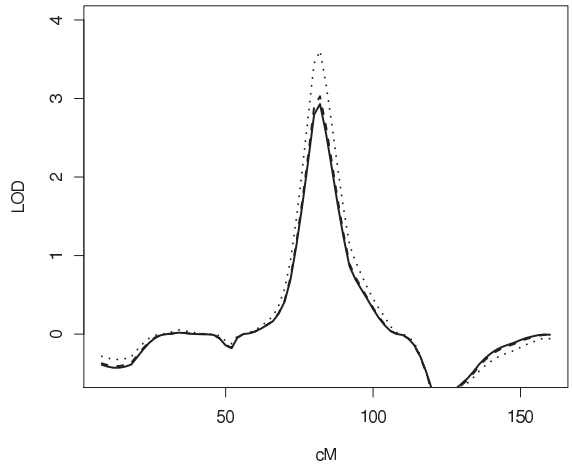

FIGURE A.1: *Results of genetic linkage analysis of breast cancer data for chromosome 9 using Arthur. Solid, dashed and dotted line represent the unweighted NPL method, the weighted NPL method based on a survival model without residual correlation, and the NPL method based on a correlated frailty model for age at onset respectively.*

A.3 Conclusion

Arthur is a package which computes weighted allele sharing statistics for genetic linkage analysis. For IBD computations the program uses MERLIN (Abecasis et al., 2002) - input files are in the MERLIN format. Various weights are currently implemented, namely for quantitative traits (Lebrec et al., 2004), for GLM traits with or without covariate adjustment, (Lebrec and van Houwelingen, 2007) and for age at onset traits with or without parental age at onsets

adjustment, (Callegaro et al., 2009, 2010). Arthur can further use different kind of weights specified in a weight file by the user.