Computer Methods and Programs in Biomedicine 151 (2017) 171–177

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

The R-package GenomicTools for multifactor dimensionality reduction and the analysis of (exploratory) Quantitative Trait Loci

Daniel Fischer[a,b]

[a] Natural Resources Institute Finland (Luke), Myllytie 1, Jokioinen, Finland
[b] University of Tampere, School of Health Sciences, Tampere, Finland

ARTICLE INFO

Article history:
Received 18 August 2016
Revised 11 July 2017
Accepted 21 August 2017

Keywords:
eQTL
QTL
MDR
R-package

ABSTRACT

Background and objectives: We introduce the R-package GenomicTools to perform, among others, a Multifactor Dimensionality Reduction (MDR) for the identification of SNP-SNP interactions. The package further provides a new class of tests for an (exploratory) Quantitative Trait Loci analysis that overcomes some of the limitations of other popular (e)QTL approaches. Popular (e)QTL approaches that use linear models or ANOVA are often based on over-simplified models that have weak statistical properties and which are not robust against outlying observations.

Method: The algorithm to calculate the MDR is well established. To speed up its calculation in R, we implemented it in C++. Further, our implementation also supports the combination of several MDR results to an MDR ensemble classifier. The (e)QTL test procedure is based on a generalized Mann-Whitney test that is tailored for directional alternatives, as they are present in an (e)QTL analysis.

Results: Our package GenomicTools provides functions to determine SNP combinations that have the highest accuracy for a MDR classification problem. It also provides functions to combine the best MDR results to a joined ensemble classifier for improved classification results. Further, the (e)QTL analysis is based on a solid statistical theory. In addition, informative visualizations of the results are provided.

Conclusion: The here presented new class of tests and methods have an easy to apply syntax, so that also researchers inexperienced in R are able to apply our proposed methods and implementations. The package creates publication ready Figures and hence could be a valuable tool for genomic data analysis.

1. Introduction

The R-package *GenomicTools* adds another tool to the statistical toolbox for the analysis of genomic data. With the advent of next-generation data and the perpetual price decline, full genome RNA- and DNA-seq data is widely available for many researchers and the demand for sound and easy to use methods is high. Already during the peak of the SNP-chip and microarray era, the need to combine Single Nucleotide Polymorphism (SNP) data with gene expression data was high and resulted in the development of analysis methods for Quantitative Trait Loci (QTL) [1] respective expression Quantitative Trait Loci (eQTL) [2]. The most commonly used command-line tool to calculate (e)QTLs is plink [3], but also commercial software like, e.g. Goldenhelix SVS8 or CLC workbench provide functions to calculate (e)QTLs. For R [4], there is the R/qtl [5] project active, for details see [6]. Also the package *MatrixEQTL* [7] is tailored for eQTL analysis.

In [8] we presented a robust, non-parametrical test for directional alternatives as they are also present in the (e)QTL analysis and the basic testing functions are available in the R-package gMWT [9]. The here presented R-package *GenomicTools* uses the testing functions and places them into a user-friendly framework for genomic data analysis. Besides the identification of significant associations between SNPs and gene expressions respective with phenotypes, the package also provides functions to visualize the results.

The second focus of our package is the Multifactor Dimensionality Reduction (MDR), for a review on MDR see [10] or for a more recent one [11]. An MDR analysis can be used to identify interactions among (discrete) variables to predict a target variable. The discrete variables can be either SNPs, or any other categorical variable, like diabetes, smoking, etc. The predicted target variable is then usually binary like for example in case of a case/control study. There are already existing tools to calculate the MDR, e.g. there is a stand-alone Java tool available to calculate the MDR [12]. Within R, the package *MDR* [13] provides functions to calculate the MDR. Further, the R-package *mdmdr* [14] provides also functions to apply

E-mail address: daniel.fischer@luke.fi
URL: http://genomictools.danielfischer.name

http://dx.doi.org/10.1016/j.cmpb.2017.08.012
0169-2607/

MDR for binary traits in case-control studies as well as for quantitative traits for unrelated individuals.

2. Models and computational methods

2.1. MDR

The MDR is a method for the analysis of the interaction of two or more categorical variables with a target variable and was introduced by Ritchie et al. [15]. The MDR is especially useful for the analysis of the interaction of SNPs with certain traits or with other genes. The gene-gene interaction is also called *epistasis* and is assumed to play a major role in the genetics of common diseases [16]. Hence, the detection of gene-gene interaction is an important step in the characterization of such diseases. The MDR method was developed to identify such interactions and is described e.g. in [17].

The basic principle of an MDR analysis is to create for k-tuples of factors k-dimensional frequency tables, assign the observations into the corresponding cells and use them to determine risk classes for each cell. For the sake of simplicity we will describe the case of a two-fold interaction, cases with larger k's are then easily derived. Let's consider a set of measured genotypes at L loci G_l with $l = 1, \ldots, L$. Further, let $\mathbb{X}_h$ be a set of H healthy individuals and $\mathbb{X}_c$ a similar set of cases of size C. For each individual in $\mathbb{X}_h$ and $\mathbb{X}_c$ the genotype information at each loci G_l are available.

In case of a two-fold interaction, all pairwise combinations $G_{l,l^*} = (G_l, G_{l^*})$ with $l \neq l^*$ are then considered. For each combination G_{l,l^*} of two loci their corresponding three possible genotypes *AA*, *Aa* and *aa* form two-dimensional 3×3 frequency tables T_{l,l^*}. The individuals from $\mathbb{X}_h$ and $\mathbb{X}_c$ are then assigned to separate frequency tables $T_{l,l^*;h}$ and $T_{l,l^*;c}$, according to the corresponding pairwise genotype information. Then, the cell-wise ratios between $T_{l,l^*;h}$ and $T_{l,l^*;c}$ are calculated and cells respective genotype combinations, are labelled either as 'high-risk' or 'low-risk', depending on the value of the ratio of cases and healthy individuals in the cell. Usually, if the value of the cell ratio is larger than the ratio of the sizes C/H of the two sets, a cell is labelled to be of high risk, indicating an over representation of cases for that particular genotype combination.

Based on the high-/low-risk labels of each cell, all individuals are then classified and compared to the real class labels. This way, for each combination G_{l,l^*} we get the amount of false positives (FP), false negative (FN), true positive (TP) and true negative (TN) classifications. Further we can calculate the sensitivity (true positive rate) as TPR=TP/(TP+FN) and the specificity (true negative rate) as TNR=TN/(TN+FP). A useful overall criteria is then the accuracy (ACC) as an average of TPR and TNR calculated as ACC = (TPR+TNR)/2. The tuple G_{l,l^*} with the highest accuracy is then reported as the best possible combination c.

The concept of MDR can be further extended and can be used as an ensemble classifier. For that, not only the best tuple is considered, but a larger set of the e best k-tuples, denoted as $\mathbb{C} = (c_1, \ldots, c_e)$. A new set of observations is then not only classified by the best tuple, but from every tuple in $\mathbb{C}$, providing e classifications. For the e classifications different statistics can then be calculated, like e.g. the average over all classification results. This average can then be interpreted as likelihood of being diseased. This concept works in a similar way as decision trees are combined in a random forest and a trained MDR ensemble might be a powerful classifier. To our knowledge, MDR has not yet been used as ensemble classifier and further theoretical considerations are required.

2.2. (e)QTL

The (e)QTL analysis is a standard procedure in genomic data analysis to associate genotypes with either gene expressions (eQTL) or with phenotypes (QTL) and a couple of different statistical methods have been proposed to calculate such associations. However, the most widely used single-marker method is to fit a linear model to the data. In an eQTL analysis the expression values of a target gene are assigned to the three genotype groups *AA*, *Aa* and *aa* of each SNP $s = 1, \ldots, S$ in a surrounding area of that gene. For an eQTL analysis, let $\mathbf{x}_g = (x_{1;g}, x_{2;g}, \ldots, x_{n;g})'$ be the vector of n expression values of a specific gene g. In case of a QTL, $\mathbf{x}_g$ are the numerical representations of some phenotype. We define then the vectors $\mathbf{x}_{0;g,s} = (x_{01;g,s}, \ldots, x_{0n_0;g,s})'$, $\mathbf{x}_{1;g,s} = (x_{11;g,s}, \ldots, x_{1n_1;g,s})'$ and $\mathbf{x}_{2;g,s} = (x_{21;g,s}, \ldots, x_{2n_2;g,s})'$, as the vectors of expression values of gene g assigned to the three genotype groups $AA(=0), Aa(=1)$ and $aa(=2)$ of one particular SNP s. The numerical representations 0,1,2 of the genotypes account then for the number of wild-type alleles, but is rather an arbitrary assignment. Further, denote the cumulative density functions (cdf) of $\mathbf{x}_{0;\,g,\,s}$, $\mathbf{x}_{1;\,g,\,s}$ and $\mathbf{x}_{2;\,g,\,s}$ as $F_{0;\,g,\,s}$, $F_{1;\,g,\,s}$ and $F_{2;\,g,\,s}$.

The dimensions of the three vectors $\mathbf{x}_{0;\,g,\,s}$, $\mathbf{x}_{1;\,g,\,s}$ and $\mathbf{x}_{2;\,g,\,s}$ correspond to the amount of individuals n_0, n_1 and n_2 that have the certain genotype in that particular SNP s and it is $n_0 + n_1 + n_2 = n$. The values of n_0, n_1 and n_2 naturally depend on s, but to keep the notation concise we skipped that additional index.

The null hypothesis for each pair g and s we have then usually in mind is

$$H_0 : F_{0;g,s} \equiv F_{1;g,s} \equiv F_{2;g,s}.$$

Different approaches test now for different alternatives with certain assumptions. In the commonly used linear model $F_{0;\,g,\,s}$, $F_{1;\,g,\,s}$, $F_{2;\,g,\,s}$ are assumed to be cdf's of $N(\mu, \sigma^2)$, $N(\mu + \Delta, \sigma^2)$ and $N(\mu + 2\Delta, \sigma^2)$ and the testing problem at hand is then H_0 : $\Delta = 0$ vs. H_1: $\Delta \neq 0$.

In case of an ANOVA these assumptions are relaxed and $F_{0;\,g,\,s}$, $F_{1;\,g,\,s}$, $F_{2;\,g,\,s}$ are cdf's of $N(\mu_0, \sigma^2)$, $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$. Here, the testing problem is $H_0 : \mu_0 = \mu_1 = \mu_2$ vs. H_1: $\mu_i \neq \mu_j$ for at least one pair $i \neq j$.

Our fully-non-parametrical approach based on a generalized Mann–Whitney test [8] has no further assumptions about parameters of $F_{0;\,g,\,s}$, $F_{1;\,g,\,s}$ and $F_{2;\,g,\,s}$ and we use the null hypothesis that we had initially in mind. The two different alternatives we consider are then

$$H_1 : F_{0;g,s} <_{st} F_{1;g,s} <_{st} F_{2;g,s}$$

$$H_2 : F_{2;g,s} <_{st} F_{1;g,s} <_{st} F_{0;g,s},$$

where $<_{st}$ refers to stochastical ordering of cdf's. These directional alternatives are a natural approach, as in case of an (e)QTL the expression values of the heterozygous genotype group *Aa* are clearly assumed to be between those of the homozygous cases *AA* and *aa*. A classical test that would lead almost to the same results is the Jonckheere-Terpstra test. The test we use here for the eQTL is implemented in our R-package *gMWT* [9] that is available on CRAN as well. The test statistic is based on the triple sum

$$TS_{g,s} = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} \mathrm{I}(x_{0i;g,s} < x_{1j;g,s} < x_{2k;g,s})$$

of the expression values of the three genotype groups, where $\mathrm{I}(\cdot)$ is the indicator function that is 1 if the condition $(\cdot)$ is true and 0 if not.

This test is then performed for every SNP in a window around the target gene. Depending on the window size of the surrounding area, eQTLs are denoted as *cis*- or *trans*-eQTL. Hence, the amount of performed tests per gene can vary between hundreds in the *cis*-case to hundreds of thousands in the *trans*-case. Our method uses by default permutation type tests and hence is slower to compute than the popular linear model or ANOVA approach. This drawback