Aulchenko *et al. BMC Bioinformatics* 2010, **11**:134
http://www.biomedcentral.com/1471-2105/11/134

BMC Bioinformatics

SOFTWARE Open Access

ProbABEL package for genome-wide association analysis of imputed data

Yurii S Aulchenko[1,2*], Maksim V Struchalin[1], Cornelia M van Duijn[1]

Abstract

Background: Over the last few years, genome-wide association (GWA) studies became a tool of choice for the identification of loci associated with complex traits. Currently, imputed single nucleotide polymorphisms (SNP) data are frequently used in GWA analyzes. Correct analysis of imputed data calls for the implementation of specific methods which take genotype imputation uncertainty into account.

Results: We developed the ProbABEL software package for the analysis of genome-wide imputed SNP data and quantitative, binary, and time-till-event outcomes under linear, logistic, and Cox proportional hazards models, respectively. For quantitative traits, the package also implements a fast two-step mixed model-based score test for association in samples with differential relationships, facilitating analysis in family-based studies, studies performed in human genetically isolated populations and outbred animal populations.

Conclusions: ProbABEL package provides fast efficient way to analyze imputed data in genome-wide context and will facilitate future identification of complex trait loci.

Background

Genome-wide association (GWA) studies became the tool of choice for the identification of loci associated with complex traits. In GWA analyses, association between a trait of interest and genetic polymorphisms (usually single nucleotide polymorphisms, SNPs) is studied using thousands of people typed for hundreds of thousands of polymorphisms. Several hundred loci for dozens of complex human disease and quantitative traits have been discovered thus far using this method [1].

For any given genetic polymorphism, association can be studied using standard statistical analysis methodology, such as fixed and mixed effects models. However, because of the large number of tests to be performed and the quantity of data to be stored in GWA studies, computational throughput and effective data handling are essential features of statistical analysis software to be used in this context. A number of specialized software packages, such as `PLINK` [2], `GenABEL` [3], `SNPTEST` [4] and `snpMatrix` [5] were developed for the statistical analysis of GWA data. Most of these packages were designed, and are fit for, the analysis of directly typed SNPs. When directly typed markers are studied, genotype calling is performed with a high degree of confidence for the vast majority of markers, resulting in four possible genotypes ("AA", "AB", "BB", and missing). This allows representation of each individual genotype using two-bit coding and consequently effective storage of the genotype data in RAM [3].

Recently, novel statistical tools for genotype imputations [4,6-9] and experimental techniques for high-throughput sequencing were developed. Implementation of these methods usually results in estimates of the posterior probability distributions $\mathbf{P}_g = (P_{AA}, P_{AB}, P_{BB})$ of the genotypes based on the available data. For many genomic loci, this distribution may be non-degenerate.

Several techniques can be applied to analysis of such "uncertain" data. The most simplistic approach would be to use the "best guess genotypes", that is to use the genotype with the highest posterior probability ($g = \max_g P_g$) for analysis as if it were a directly typed markers. This approach is equivalent to replacing the estimated probability distribution with a degenerate one where a probability of one is assigned to the genotype with the maximal posterior probability. From standard statistical theory it is known, however, that such a procedure results in biased estimates of the effects. A

* Correspondence: i.aoultchenko@erasmusmc.nl
[1]Department of Epidemiology, Erasmus MC, Postbus 2040, 3000 CA Rotterdam, The Netherlands

correct analysis can be achieved using a maximum likelihood approach. Under this approach the likelihood can be computed using the total probability formula in which summation is performed over the genotypes, whose true values are not known, but whose posterior probabilities can be estimated given the data. This approach is computationally demanding, as it requires summation over the underlying probability distribution and numerical maximization of the likelihood function. Alternatively, a regression approach in which the posterior genotypic probabilities are used as predictors, can be applied. The main advantage of this approach is that well-established regression analysis methodology, algorithms, and code can be used in its implementation. Most currently available packages for GWA analysis can not be directly used in this manner, as they assume degenerate genotypic distributions and do not provide a facility for the storage and analysis of real-number predictors (posterior genotypic probabilities).

In this work, we describe the `ProbABEL` package, which was designed to perform genome-wide regression on posterior genotypic probabilities estimated using imputation software, such as `MACH`[6] or `IMPUTE`[4,9]. In addition to standard linear and logistic regression, which is widely applied to the analysis of quantitative and binary outcomes in population-based GWA studies, we also implemented a Cox proportional hazards model. For quantitative traits, we implemented a fast two-step mixed model-based score test for association testing in studies with a high degree of confounding induced by differential relationships between study subjects (e.g. family-based studies, studies of human genetically isolated populations, and studies in outbred animal populations).

Implementation

Here, in the first few sub-sections, we will describe `ProbABEL` software, giving only the main outline of the underlying theory and with special emphasis on implementation and the options allowing to access specific analyzes within `ProbABEL`. In two last sub-sections, starting with the "Fixed effects model theory", we will give more in-depth review of the theory used by the package.

`ProbABEL` was implemented using code written in the C and C++ languages. The package consists of three executable files, used to perform linear, logistic, and Cox regressions, and a helper Perl script which facilitates the analysis of multiple chromosomes.

The package implements standard regression analysis methodology outlined in the section "Fixed effects model theory" and specific approximation to the mixed linear model described in the section "Two-step score test approximation to the mixed model". The key statistical tests performed by `ProbABEL` concern testing of the SNP effects. Here, we will describe the tests performed by `ProbABEL` using an example of linear regression; testing using other types of regression follows similar logic.

In linear regression, the expectation of the trait is described as

$$E[\mathrm{Y}] = \mathbf{X}\boldsymbol{\beta} = \mathrm{X}_x\boldsymbol{\beta}_x + \mathrm{X}_g\boldsymbol{\beta}_g$$

where $\mathbf{Y}$ is the vector of phenotypic values, $\mathbf{X}_g$ is the design matrix containing data about predictors of interest (these involving SNP data), and $\mathbf{X}_x$ is the design matrix containing other (nuisance) covariates. β_g and β_x are the vectors of corresponding fixed effects. The vector of phenotypes $\mathbf{Y}$ and the covariates matrix $\mathbf{X}_x$ are provided in the phenotype file. The genotypic data are read from the genotype (dose or probability) files and are analyzed one SNP at a time.

Our interest lies in testing the (components of) β_g . `ProbABEL` provides the estimates of the components of the vector β_g and corresponding standard errors, and, in most cases, the test of the general hypothesis concerning the involvment of the SNP, obtained by comparison of the estimated model to the null model formulated as $\beta_{g,0} = \mathbf{0}$, where $\mathbf{0}$ is the vector of zeros.

Under the general genotypic model, $\mathbf{X}_g$ is a matrix with the number of rows equal to the number of people under consideration and with two columns. Each row of the matrix contains the estimated probabilities that a person has genotype "AA" or "AB". Then, the vector of genotypic effects is described with two parameters: $\beta_g = (\beta_{AA}, \beta_{AB})$. Thus formulated, the model allows for the estimation of a general genotypic two-degree of freedom model. Further, a number of sub-models can be formulated by setting restrictions on these parameters. The "dominant B allele" model is formalized as $\beta_{AB} = 0$, "dominant A" (the same as "recessive B") as $\beta_{AA} = \beta_{AB}$, the additive model as $2 \cdot \beta_{AB} = \beta_{AA}$, and the over-dominant model as $\beta_{AA} = 0$. Note that the additive model is equivalent to performing linear regression on the estimated dose of allele "A" defined as $P_{AB} + 2 \cdot P_{AA}$. The latter model is tested when the allelic dosage file is provided as the input for `ProbABEL`, while the full range of described models is tested if the estimated probability files (option "`–ngpreds = 2`") are supplied.

`ProbABEL` can also test for interaction between a specified covariate and the set of SNPs; for that alternative, the interaction covariate should be specified using the "`–interaction N`" option, where N corresponds to the number of the column of the design matrix $\mathbf{X}_x$, which contains that covariate. If this option is used, the expectation of the trait is defined as

$$E[\mathrm{Y}] = \mathbf{X}\boldsymbol{\beta} = \mathrm{X}_x\boldsymbol{\beta}_x + \mathrm{X}_g\boldsymbol{\beta}_g + (\mathrm{X}_g^T\mathrm{W})^T\boldsymbol{\beta}_{\mathrm{gxe}}$$

where $\mathbf{W}$ is a diagonal matrix, whose diagonal elements are formed by substituting the interaction covariate to the matrix and β_{gxe} is the vector of interaction regression coefficients.

Analysis of population-based data

If the study subjects can be assumed to be genetically "independent", in the sense that they come from the general outbred population without a marked degree of stratification and that cryptic relatedness is absent, the data can be effectively analyzed using standard linear fixed effects regression methodology, as described in section "Fixed effects model theory". The (small) effects of confounding can be corrected posterior to analysis using the genomic control [10] procedure. If a marked degree of stratification is present, such methods as structured association analysis and EIGENSTRAT [11] can be combined with the standard methods.

Using standard methods, the estimates of the parameters can be obtained using the standard formula 1 (see "Fixed effects model theory" below), which provides maximum likelihood estimates if $(X^TX)^{-1}$ exists. The latter condition is fulfilled for virtually all analyses; practically, exceptions may occur for SNPs with very low minor allele frequencies or poor quality imputations.

The standard errors are computed as square roots of the diagonal elements of the parameter estimates' variance-covariance matrix. This matrix is computed using one of three different methods: the standard method, with residual variance estimated under the alternative (formula 2, see "Fixed effects model theory" below) or null hypothesis concerning SNPs (option "`--score`"), or using a "sandwich" estimator (formula 5, see "Fixed effects model theory"), resulting in robust standard errors (option "`--robust`"). The value of the global likelihood ratio test statistic, testing the joint significance of all terms involving SNP, is computed using the formula 3 (see "Fixed effects model theory"). In this test, the null model is formulated as $\beta_{g,0} = \mathbf{0}$, where $\mathbf{0}$ is the vector of zeros. If an interaction term is present, that is also set to zero under the null: $\beta_{gxe,0} = \mathbf{0}$. The likelihoods involved are computed using the formula 4 (see "Fixed effects model theory") with the values of the parameters fixed at the point of the maximum likelihood estimate obtained with 1 (see "Fixed effects model theory").

Analysis of data on subjects with differential relationships

In the case of a study involving subjects with markedly differential relationships (family-based designs, studies of human genetically isolated populations, studies in outbred animal populations), a mixed model approach may be used, in which a random effect ("heritability") accounts for similarities between the phenotypes of study subjects [12]. However, the estimation of the full mixed model using either maximum likelihood or the restricted maximum likelihood approach is computationally demanding, if not unfeasible, within the framework of GWAS [13], and therefore a two-step mixed model-based approach [13-15] is utilized in `ProbABEL`.

In this approach, the mixed model containing all terms but those involving SNP is first estimated by maximizing the likelihood function provided by the expression 7 (see section "Two-step score test approximation to the mixed model" for details). These estimates are then used in the second step to compute estimates of the SNP effects (formula 8 of "Two-step score test approximation to the mixed model") and the variance-covariance matrix of these estimates (formula 10, see "Two-step score test approximation to the mixed model"). These values can be used to perform a score test for association. The second step of a mixed-model based score test for association is available in `ProbABEL` using option "`--mmscore IVFile`", where `IVFile` is the name of a file containing the inverse of the variance-covariance matrix ($\mathbf{V}^{-1}_{\hat{h}^2,\hat{\sigma}^2}$ of formulas 8 and 10, see "Two-step score test approximation to the mixed model") evaluated at the point of the maximum likelihood estimates obtained in step one. The phenotypes analyzed in the second step are residuals (as specified by the formula 9, see "Two-step score test approximation to the mixed model") obtained by subtracting the trait values expected under the mixed model-based estimates of the fixed effects from the original trait values.

Step one of the regression procedure can be performed using our `GenABEL` software [3]. This software performs genomic data based estimation of the kinship matrix as described in section "Estimation of genomic kinship matrix" using the `ibs (...,weight="freq")` function, and performs maximum likelihood estimation of the step-one mixed model using the `polygenic()` function. The resulting object contains the inverse variance-covariance matrix (`object $InvSigma`), which can be saved as a text file and used in `ProbABEL` analysis. The residuals to be used as trait values in step two of the analysis can be accessed through `object$residualY`.

Input and output

The input consists of a phenotypic data file and a set of files describing the imputed genotypic data. The phenotypic file provides data on the outcome of interest and any additional covariates to be included in the analysis. The genotypic data files, at present, utilize the MACH imputation software output format. Minimally, a file with estimated probability distributions ("mlprob") or allelic dosages ("mldose") and the "mlinfo" file containing information about allele coding and overall imputation quality should be provided. Optionally, a map file in HapMap format, containing chromosome and location information, may be supplied. Information contained in the latter two files is not used in analysis, but is forwarded directly to the output. If the mixed-model based score test for association in related individuals is to be computed, a file containing the inverse matrix of variances and covariances between the phenotypes of study individuals should be supplied as a part of the input. The output of the program consists of one line for each SNP tested, containing information about the SNP supplied as part of the input, as well as the results from analysis (estimates of the coefficients of regression, standard errors of the coefficients, and test statistic values).

Fixed effects model theory

Most of the fixed effects model theory outlined here is standard and can be found in textbooks, such as "Generalized, Linear, and Mixed Models" [16]. Specific references are provided when this is not the case.

Linear regression assuming normal distribution

Standard linear regression theory is used to estimate coefficients of regression and their standard errors. We assume linear model with expectation

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

and variance-covariance matrix

$$\mathbf{V} = \sigma^2\mathbf{I}$$

where **Y** is the vector of phenotypes of interest, **X** is design matrix, β is the vector of regression parameters, σ^2 is variance and **I** is identity matrix.

The maximum likelihood estimates (MLEs) for the regression parameters is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad (1)$$

and MLE of the residual variance is

$$\hat{\sigma}^2 = \frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{N-r_X}$$

where N is the number of observations and r_X is rank of **X** (number of columns of the design matrix).

The variance-covariance matrix for the parameter estimates under alternative hypothesis can be computed as

$$\mathbf{var}_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1} \quad (2)$$

For the *j*-the element $\hat{\boldsymbol{\beta}}(j)$ of the vector of estimates the standard error under alternative hypothesis is given by the square root of the corresponding diagonal element of the above matrix, $\mathbf{var}_{\hat{\boldsymbol{\beta}}}(jj)$, and the Wald test can be computed with

$$T^2(j) = \frac{\hat{\boldsymbol{\beta}}(j)^2}{\mathbf{var}\,\hat{\boldsymbol{\beta}}(jj)}$$

which asymptotically follows the χ^2 distribution with one degree of freedom under the null hypothesis. When testing significance for more than one parameter simultaneously, several alternatives are available. Let us partition the vector of parameters into two components, $\beta = (\beta_g, \beta_x)$, and our interest is testing the parameters contained in β_g (SNP effects), while β_x (e.g. effects of sex, age, etc.) are considered nuisance parameters. Let us define the vector of the parameters of interest which are fixed to certain values under the null hypothesis as $\beta_{g,0}$ (usually, $\beta_{g,0} = \mathbf{0}$, vector of zeros).

The likelihood ratio test can be obtained with

$$LRT = 2 \cdot (logLik(\hat{\boldsymbol{\beta}}_g, \hat{\boldsymbol{\beta}}_x) - logLik(\boldsymbol{\beta}_{g,0}, \hat{\boldsymbol{\beta}}_x)) \quad (3)$$

which under the null hypothesis is asymptotically distributed as $\chi^2$with number of degrees of freedom equal to the number of parameters specified by β_g . Assuming the normal distribution, the log-likelihood of a model specified by the vector of parameters β and residual variance σ^2 can be computed as

$$logLik(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}(N.log_e\sigma^2 + (\mathbf{Y} - \boldsymbol{\beta}\mathbf{X})^T(\mathbf{I}/\sigma^2)(\mathbf{Y} - \boldsymbol{\beta}\mathbf{X})) \quad (4)$$

Secondly, the Wald test can be used; for that the inverse variance-covariance matrix of $\hat{\boldsymbol{\beta}}_g$ should be computed as

$$\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}_g} = \mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(g,g) - \mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(g,x)(\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(x,x))^{-1}\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(x,g)$$

where $\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}}(a,b)$ correspond to sub-matrices of the inverse of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, involving either only covariances between the parameters of interest (*g, g*), only the nuisance parameters (*x, x*) or between the parameters of interest and nuisance parameters, (*x, g*), (*g, x*).

The Wald test statistics is then computed as

$$W^2 = (\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g,0})^T\mathbf{var}^{-1}_{\boldsymbol{\beta}_g}(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g,0})$$

which asymptotically follows the χ^2 distribution with the number of degrees of freedom equal to the number of parameters specified by β_g . The Wald test generally is computationally easier than the LRT, because it avoids estimation of the model specified by the parameter's vector ($\beta_{g,0}$, $\hat{\boldsymbol{\beta}}_x$).

Lastly, similar to the Wald test, the score test can be performed by use of $\mathbf{var}_{\boldsymbol{\beta}=(\boldsymbol{\beta}_{g,0},\boldsymbol{\beta}_x)}$ instead of $\mathbf{var}_{\hat{\boldsymbol{\beta}}}$.

Logistic regression

For logistic regression, the procedure to obtain parameters estimates, their variance-covariance matrix, and tests are similar to these outlined above with several modifications.

The expectation of the binary trait is defined as expected probability of the event as defined by the logistic function

$$E(\mathbf{Y}) = \pi = \frac{1}{1+e^{-(\mathbf{X}\boldsymbol{\beta})}}$$

The estimates of the parameters are obtained not in one step, as is the case of the linear model, but using iterative procedure (iteratively re-weighted least squares). This procedure is not described here for the sake of brevity.

The log-likelihood of the data is computed using binomial probability formula:

$$logLik(\boldsymbol{\beta}) = \mathbf{Y}^T log_e \pi + (\mathbf{1}-\mathbf{Y})^T log_e(\mathbf{1}-\pi)$$

where $log_e\pi$ is a vector obtained by taking the natural logarithm of every value contained in the vector π.

Robust variance-covariance matrix of parameter estimates

For computations of robust variance-covariance matrix we use White's sandwich estimator [17,18], which is equivalent to the "HC0" estimator described by Zeilers and Lumley in "`sandwich`" package for R.

For linear model, the variance-covariance matrix of parameter estimates is computed using formula

$$\mathbf{var}_r = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{R}\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}$$

where **R** is a diagonal matrix containing squares of residuals of **Y**. The same formula may be used for "standard" analysis, in which case the elements of the **R** matrix are constant, namely mean residual sum of squares (the estimate of residual variance, $\hat{\sigma}^2$).

Similar to that, the robust matrix is computed for logistic regression with

$$\mathbf{var}_r = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{R}\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$$

where **W** is the diagonal matrix of "weights" used in logistic regression.

Cox proportional hazards model

The implementation of the Cox proportional hazard model used in `ProbABEL` is entirely based on the code of `R` library `survival` developed by Thomas Lumley (function `coxfit2`), and is therefore not described here.

Two-step score test approximation to the mixed model

The framework for analysis of data containing differential relationships follows the two-step logic developed in the works of Aulchenko et al. [13] and Chen and Abecasis [14]. General analysis model is a linear mixed model which defines the expectation of the trait as

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

identical to that defined for linear model. To account for possible correlations between the phenotypes of study subjects the variance-covariance matrix is defined to be proportional to the linear combination of the identity matrix **I** and the relationship matrix **Φ**:

$$\mathbf{V}_{\sigma^2,h^2} = \sigma^2(2h^2\mathbf{\Phi} + (1-h^2)\mathbf{I})$$

where h^2 is the heritability of the trait. The relationship matrix **Φ** is twice the matrix containing the coefficients of kinship between all pairs of individuals under consideration; its estimation is discussed in a separate section "Estimation of genomic kinship matrix".

Estimation of thus defined model is possible by numerical maximization of the likelihood function, however, the estimation of such model for large data sets is not computationally feasible for hundreds of thousands to millions of SNPs tested in the context of GWAS, as we have demonstrated previously [13].

Two-step score test for association

A two-step score test approach is therefore used to decrease the computational burden. Let us re-write the expectation of the trait by splitting the design matrix in two parts, the "base" part $\mathbf{X}_x$, which includes all terms not changing across all SNP models fit in GWAS (e.g. effects of sex, age, etc.), and the part including SNP information, $\mathbf{X}_g$:

$$E[\mathbf{Y}] = \mathbf{X}_x\boldsymbol{\beta}_x + \mathbf{X}_g\boldsymbol{\beta}_g$$

Note that the latter design matrix may include not only the main SNP effect, but e.g. SNP by environment interaction terms.

At the first step, linear mixed model not including SNP effects

$$E[\mathbf{Y}] = \mathbf{X}_x\boldsymbol{\beta}_x$$

is fitted. The maximum likelihood estimates (MLEs) of the model parameters (regression coefficients for the fixed effects $\hat{\boldsymbol{\beta}}_x$, the residual variance $\hat{\sigma}_x^2$ and the heritability $\hat{h}_x^2$) can be obtained by numerical maximization of the likelihood function

$$logLik(\boldsymbol{\beta}_x, h^2, \sigma^2) = -\frac{1}{2}(log_e \mid \mathbf{V}_{\sigma^2,h^2} \mid +(\mathbf{Y} - \boldsymbol{\beta}_x\mathbf{X}_x)^T \mathbf{V}^{-1}_{\sigma^2,h^2}(\mathbf{Y} - \boldsymbol{\beta}_x\mathbf{X}_x)) \quad (7)$$

where $\mathbf{V}^{-1}_{\sigma^2,h^2}$ is the inverse and $\mid \mathbf{V}_{\sigma^2,h^2} \mid$ is the determinant of the variance-covariance matrix.

At the second step, the estimates of the fixed effects of the terms involving SNP are obtained with

$$\hat{\boldsymbol{\beta}}_g = (\mathbf{X}_g^T \mathbf{V}^{-1}_{\hat{\sigma}^2,\hat{h}^2} \mathbf{X}_g)^{-1} \mathbf{X}_g^T \mathbf{V}^{-1}_{\hat{\sigma}^2,\hat{h}^2} \mathbf{R}_{\hat{\boldsymbol{\beta}}_x} \quad (8)$$

where $\mathbf{V}^{-1}_{\hat{\sigma}^2,\hat{h}^2}$ is the variance-covariance matrix at the point of the MLE estimates of $\hat{h}_x^2$ and $\hat{\sigma}_x^2$ and

$$\mathbf{R}_{\hat{\boldsymbol{\beta}}_x} = \mathbf{Y} - \hat{\boldsymbol{\beta}}_x \mathbf{X}_x \quad (9)$$

is the vector of residuals obtained from the base regression model. Under the null model, the inverse variance-covariance matrix of the parameter's estimates is defined as

$$\mathbf{var}_{\hat{\boldsymbol{\beta}}_g} = \hat{\sigma}_x^2 (\mathbf{X}_g^T \mathbf{V}^{-1}_{\hat{\sigma}^2,\hat{h}^2} \mathbf{X}_g)^{-1} \quad (10)$$

Thus the score test for joint significance of the terms involving SNP can be obtained with

$$T^2 = (\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g,0})^T \mathbf{var}^{-1}_{\boldsymbol{\beta}_g} (\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g,0})$$

where $\beta_{g,0}$ are the values of parameters fixed under the null model. This test statistics under the null hypothesis asymptotically follows the χ^2 distribution with the number of degrees of freedom equal to the number of parameters tested. The significance of an individual *j*-the elements of the vector $\hat{\boldsymbol{\beta}}_g$ can be tested with

$$T_j^2 = \frac{\hat{\boldsymbol{\beta}}_g^2(j)}{\mathbf{var}\hat{\boldsymbol{\beta}}_g(jj)}$$

where $\hat{\boldsymbol{\beta}}_g^2(j)$ is square of the *j*-th element of the vector of estimates $\hat{\boldsymbol{\beta}}_g$, and $\mathbf{var}_{\hat{\boldsymbol{\beta}}_g}(jj)$ corresponds to the *j*-th diagonal element of $\mathbf{var}^{-1}_{\hat{\boldsymbol{\beta}}_g}$. This statistics asymptotically follows χ_1^2.

Estimation of genomic kinship matrix

The relationship matrix **Φ** used in estimation of the linear mixed model is twice the matrix containing the coefficients of kinship between all pairs of individuals under consideration. This coefficient is defined as the probability that two gametes randomly sampled from each member of the pair are identical-by-descent (IBD), that is they are copies of exactly the same ancestral allele. The expectation of kinship can be estimated from pedigree data using standard methods, for example the kinship for two outbred sibs is 1/4, for grandchild-grandparent is 1/8, etc. However, in many situations, pedigree information may be absent, incomplete, or not reliable. Moreover, the estimates obtained using pedigree data reflect the expectation of kinship, while the true realization of kinship may vary around this expectation. In presence of genomic data it may therefore be desirable to estimate the kinship coefficient from these, and not from pedigree. It can be demonstrated that unbiased and positive semi-definite estimator of the kinship matrix [19] can be obtained by computing the kinship coefficients between individuals *i* and *j* with

$$\hat{K}_{ij} = \frac{1}{L}\sum_{l=1}^{L} \frac{(g_{l,i} - p_l)(g_{l,j} - p_l)}{p_l(1 - p_l)}$$

where L is the number of loci, p_l is the allelic frequency at *l*-th locus and $g_{l,\ j}$ is the genotype of *j*-th person at the *l*-th locus, coded as 0, 1/2, and 1, corresponding to the homozygous, heterozygous, and other type of homozygous genotype [11,15,19]. The frequency is computed for the allele which, when homozygous, corresponds to the genotype coded as "1'.

Results

To ensure the statistical correctness of the two-step procedure, we performed a small-scale simulation study. We used real data from the Erasmus Rucphen Family (ERF) study [20]. In simulations, we used genotypic data from 2,313 people who had high-density SNP genotyping data. The trait was simulated as a sum of four independent effects: two fixed effects explaining 10 and 5% of the total trait variance, a polygenic effect, and a residual random effect. The residual random effect was assumed to be distributed normally with mean zero and variance fixed at the value that explained 59.5% of total variance. To simulate the polygenic effect, similar to our previous work [15], we selected 200 random SNPs, and assigned these SNPs with fixed effects such that, in total, these SNPs explained 25.5% of total variance. Thus, the heritability of the trait when adjusted for the fixed effects was 30%.

The SNPs mimicking the polygenic effect were selected randomly from all autosomes but the second. To estimate type 1 error of the two-step procedure, we studied association of the trait with the second chromosome SNPs using real imputed data. Only SNPs with estimated minor allele frequencies greater than 1% were used in analysis (212,691 SNPs in total). We compared type 1 error rates for four different models: a linear model ignoring the relatedness structure (using both a standard and a robust covariance matrix) and our 2-step mixed-model based score test. For the latter, we adjusted for two fixed effect covariates in the first step (polygenic) analysis.

The results of these tests are summarized in Table 1. It is easy to see that when relationships between study individuals are not taken into account, the distribution of the test statistic is inflated, regardless of whether a robust or standard covariance matrix is used. In our previous work, we demonstrated that this inflation grows with increasing trait heritability, with more close relatives present in the sample [15] and with increasing sample size and can reach very high values. On the contrary, when two-step approximation to the mixed model is used ("Linear, mmscore" row of Table 1), the test statistic shows very good agreement to the $\chi^2_{df=1}$ distribution expected under the null.

Next, we measured CPU time required for particular ProbABEL analyses. To do this, we selected 500, 1000, and 1500 people from 2,313 genotyped individuals and measured the speed of different types of analysis using chromosome 2 imputed data on 220,833 SNPs. All analyses were ran on a Sun Fire X4640 server with an Intel Xeon CPU 5160 (3.00 GHz). Results are present in Table 2. From this table, it is clear that all population-based analyzes (these not involving the −mmscore option) scale roughly linearly with the number of people. Use of the −robust option increases the running time by only a small fraction. Based on these data, one would expect that a GWA analysis involving, for example, 2.5 millions SNPs imputed on HapMap2 release 22 in 1,500 individuals would take 1/2 hour for linear, 2 hours for logistic and 1 1/2 hours for Cox proportional hazards models.

Use of the −mmscore option to adjust for relationships between study subjects, however, induces a non-linear relationship between the number of study subjects and analysis time: while the time to analyze 500 people is 16 minutes, the time for analysis of 1500 people is ≈ 14 times longer. The time for a GWA with 1,500 people and 2.5 millions imputed SNPs is, therefore, estimated to be ≈ 43 hours.

Table 1 Mean values of the test statistics (Wald for Linear, score for mmscore), genomic control λ (median test statistic over 0.455), and type 1 error at different α for different models.

			α		
Model	**Mean(T^2)**	λ	**0.05**	**0.01**	**0.001**
Linear	1.206	1.224	0.073	0.018	0.0027
Linear, robust	1.210	1.228	0.073	0.018	0.0028
Linear, mmscore	0.984	1.007	0.047	0.009	0.0011

Tests were performed using a trait dependent on two covariates and with (adjusted) heritability of 30%. Only SNPs with estimated minor allele frequency greater than 0.01 (*n* = 212, 691) used. Linear: standard linear model; Linear, robust: linear models using with standard errors; Linear, mmscore: two-step approximation to mixed model, fixed effects included in step 1 of analysis.

Table 2 Time for analysis of chromosome 2 imputed data (220,833 SNPs).

Model	Option	No. people	CPU time
Linear	-	500	0 m 43 s
		1000	1 m 23 s
		1500	2 m 10 s
Linear	−robust	500	0 m 50 s
		1000	1 m 43 s
		1500	2 m 35 s
Linear	−mmscore	500	16 m 18 s
		1000	92 m 45 s
		1500	231 m 49 s
Logistic	-	500	3 m 20 s
		1000	6 m 38 s
		1500	10 m 8 s
Logistic	−robust	500	3 m 25 s
		1000	6 m 53 s
		1500	10 m 29 s
Cox PH	-	500	2 m 18 s
		1000	4 m 30 s
		1500	6 m 43 s

In all analyzes, 2 covariates were included in the model.

Discussion

Imputed SNP data are conventionally used for the analysis of GWA data; correct use of imputed data allows for higher power and location accuracy [21,22]. However, correct analysis of imputed data needs to account for the uncertainty surrounding estimated genotypic probability distributions. This can be done using approaches based on either likelihood or regression on estimated probabilities, as outlined in the “Background” and “Implementation”. A number of software packages are available for such analyses. SNPTEST implements a score test based on missing data likelihood [4] allowing for the study of both quantitative and binary outcomes. MACH2QTL and MACH2DAT implement regression models on estimated probabilities for quantitative and binary traits, respectively, in a manner similar to ProbABEL. ProbABEL extends the functionality available in these packages by allowing analysis under the Cox proportional hazards model. Further, while SNPTEST allows for testing interaction of a covariate with SNPs studied, it does not provide the value of the global significance test. Finally, ProbABEL is the only package that implements specific mixed-model based procedures for the study of association in samples with differential relationships, facilitating analysis in family-based studies, studies performed in human genetically isolated populations, and outbred animal populations.

In theory, the mixed model we have described can also be used to correct for population stratification in a study where a number of (population-based and family based) samples come from differentiated genetic populations [12,19]. However, given the different genetic and potentially different environmental compositions of such differentiated populations, similar heritabilities can not be assumed in all study populations. We speculate that, in practice, one should combine population-specific (fixed or mixed-model) approaches with structured association or similar methods. For example, one could identify sets of individuals coming from divergent genetic populations using either prior information or analysis of the principal components of the genomic kinship matrix [11]; perform standard analysis in population-based sets and mixed-model analysis in family based sets (or those exhibiting substantial cryptic relatedness), as described here; and finally combine the results using meta-analysis. The best strategy to analyze such complex studies is to be addressed elsewhere in more details.

The two-step mixed model-based score test implemented in `ProbABEL` is an extension of the family-based association score test suggested by Chen and Abecasis [14], and is similar in its logic to the GRAMMAR and GRAMMAR-GC tests described by Aulchenko et al. [13,15]. In the test procedure, the model is split into two parts (see the equation 6 in “Two-step score test approximation to the mixed model”), the first of which contains the effects of nuisance parameters, including random genetic effects, and the second includes the parameters of interest (SNP effects and SNP-interacting covariates). Estimation in the second step is performed based on the estimates obtained from fitting the first part. Strictly speaking, the test defined in this manner is correct if the distributions of covariates in the first and the second parts of the model are independent conditional on the estimated phenotypic variance-covariance matrix. This assumption is most likely to be true when the covariates included in the base model are environmental ones, and thus are not expected to exhibit conditional correlation with SNPs. However, when endogenous risk factors, such as body mass index, are included as the covariates in the base model, some SNPs are expected to exhibit covariance with this covariate. In such situations, the covariate should be included in the second step analysis. This, however, may violate the assumptions of the score test if the covariate explains a large proportion of trait variance. In such situation we expect that the test will become conservative and may be less powerful compared to the classical maximum likelihood analysis.

At present, GWA analysis of millions of imputed SNPs using the `–mmscore` option in `ProbABEL` takes a few days for samples of a few thousands of people. However, the relationship between CPU time and the number of subjects is not linear; as the number of subjects reaches 5,000 or more, the mixed-model based analysis will take too much time (weeks to months) when using a single CPU. A straightforward approach to solve this problem would be to use parallel computations. Still, the non-linear dependency of computational time on the number of subjects may become a major analysis bottleneck with larger and larger studies becoming available.

Other software packages which implement similar mixed-model functionality and are suitable for GWA analyses are MERLIN [23] and QxPak [24]. In particular, MERLIN implements the two-step score test [14], which is equivalent to our test in the absence of covariates. QxPak is a flexible tool for mixed modeling of quantitative traits, which implements classical full Maximum Likelihood and Restricted Maximum Likelihood estimation procedures. Neither MERLIN nor QxPak, however, allow for analyses of imputed data in the form of regression onto estimated genotype probabilities. Both packages assume that pedigree structure is known, and estimate kinship based on that.

On the contrary, the input required by ProbABEL consists of the inverse matrix of estimated variances and covariances between the phenotypes of study individuals. This matrix can be obtained in a number of different ways; our standard approach is to estimate it using `GenABEL`’s `polygenic()` function based on kinship estimated from genomic data, as computed with the `ibs(..., weight="freq")` function.

However, it is possible and straightforward to use kinship estimated from pedigree data as well (using, e.g., “`kinship`“ library of `R`) in the `polygenic()` procedure. The latter approach is preferable in a study where no genome-wide data is available for estimation of genomic kinship (such as a candidate gene or region study).

Presently, there is no package (including `ProbABEL`), which allows for genome-wide association analysis of binary traits or time-till-event outcomes under a mixed model or an approximation to a mixed model accounting for relatedness, and providing the correct estimates of Odds or Hazards Ratios. With the growing number of GWA scans performed in families and genetically isolated populations, this gap needs to be filled.

For population-based analyses using fixed effects models, `ProbABEL` computes Maximum Likelihood estimates of the parameters and the standard errors under the alternative hypothesis, allowing a Wald test for every parameter under consideration. The global SNP significance test is implemented using the Likelihood Ratio Test. Theoretically, the Wald test can be used for the same purpose, thereby avoiding the need to re-estimate the null model with respect to each SNP. However, in GWAS with imputed data, where full information is available for all SNPs, the null model estimation needs to be performed only once, and can then used for testing all SNPs. Thus the overhead related to re-estimation of the null model is minimal, and, for that reason, we did not implement the global SNP significance Wald test.

We should emphasise that, in general, the `ProbABEL` software can be used to do massive regression analyzes using any type of real-type outcomes and predictors. As such, `ProbABEL` is not restricted to SNP, or even, more generally, to genetic analyzes and can be used for any analyzes requiring regression of a dependent variable on a very large number of independent variables in turn. For example, `ProbABEL` may be use to perform association testing among traits and Copy Number Polymorphisms [25].

The practical applicability of `ProbABEL` for the analysis of GWAS is confirmed by the fact that the early versions of the package were successfully used for analysis of multiple data sets, including already published genome-wide analyzes of such various traits as height [26,27], gout [28], waist circumference [29], smoking initiation [30], and others.

Conclusions

We developed the ProbABEL software package, which facilitates fast genome-wide association analysis of imputed data under linear, logistic and Cox proportional hazards models. For quantitative traits, the package also implements a two-step mixed model-based score test for association in samples with differential relationship, facilitating analysis in family-based studies, studies performed in human genetically isolated populations, and outbred animal populations.

Availability and requirements

Project name: ProbABEL

Project home page: http://mga.bionet.nsc.ru/~yurii/ABEL/ (source code and binaries for various platforms), http://r-forge.r-project.org/projects/genabel/ (project development page)

Operating system(s): source code was successfully compiled and used on Windows, Mac OS X, Linux, SUN Solaris

Programming language: C, C++, Perl

Other requirements: make

License: GNU GPL

Any restrictions to use by non-academics: None

Acknowledgements

We thank Dr. Aaron Isaacs, Prof. Tatiana I. Axenovich, Dr. Wlliam Astle, Prof. David J. Balding and two anonymous reviewers for valuable suggestions and discussion. We also appreciate the efforts of Thomas Lumley and other member of R community to provide free statistical computing code. Above all we appreciate help of `ProbABEL` and `GenABEL` users who provided valuable comments about the functionality and helped detecting errors in the `ProbABEL` code.

The work was supported by the Centre for Medical Systems Biology (CMSB), the Netherlands Organization for Scientific Research (NWO) and Russian Foundation for Basic Research (NWO-RFBR 047.017.043) and the Horizon program of the Netherlands Genomics Initiative (93518029).

Author details

[1]Department of Epidemiology, Erasmus MC, Postbus 2040, 3000 CA Rotterdam, The Netherlands. [2]Institute of Cytology and Genetics SD RAS, Novosibirsk, 630090, Russia.

Authors' contributions

YSA developed the original idea, methodology, and code for the fixed effects part. MVS contributed the code for the interaction testing and two-step mixed-model based procedures. CvD provided ERF study data. All co-authors contributed to writing of the manuscript.

Received: 19 October 2009 Accepted: 16 March 2010
Published: 16 March 2010

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci USA* 2009, **106(23)**:9362-9367.
2. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81(3)**:559-575.
3. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics* 2007, **23(10)**:1294-1296.
4. Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes.** *Nat Genet* 2007, **39(7)**:906-913.
5. Clayton D, Leung HT: **An R package for analysis of whole-genome association studies.** *Hum Hered* 2007, **64**:45-51.
6. Li Y, Abecasis G: **Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference.** *American Journal of Human Genetics* 2006, **S79**:2290.
7. Li Y, Willer C, Sanna S, Abecasis G: **Genotype imputation.** *Annu Rev Genomics Hum Genet* 2009, **10**:387-406.
8. Servin B, Stephens M: **Imputation-based analysis of association studies: candidate regions and quantitative traits.** *PLoS Genet* 2007, **3(7)**:e114.
9. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.** *PLoS Genet* 2009, **5(6)**:e1000529.
10. Devlin B, Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55(4)**:997-1004.
11. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, 38(8):904-909.
12. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.** *Nat Genet* 2006, **38(2)**:203-208.
13. Aulchenko YS, de Koning DJ, Haley C: **Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis.** *Genetics* 2007, **177**:577-585.
14. Chen WM, Abecasis GR: **Family-based association tests for genomewide association scans.** *Am J Hum Genet* 2007, **81(5)**:913-926.
15. Amin N, van Duijn CM, Aulchenko YS: **A genomic background based method for association analysis in related individuals.** *PLoS One* 2007, **2(12)**:e1274.

16. McCulloch CE, Searle SR: *Generalized, Linear, and Mixed Models* John Wiley & Sons, Inc 2001.
17. White H: **A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity.** *Econometrica* 1980, **48**:817-838.
18. Zeileis A: **Econometric Computing with HC and HAC Covariance Matrix Estimators.** *Journal of Statistical Software* 2004, **11**:1-17.
19. Astle W, Balding DJ: **Population structure and cryptic relatedness in genetic association studies.** *Statistical Science* .
20. Pardo LM, MacKay I, Oostra B, van Duijn CM, Aulchenko YS: **The effect of genetic drift in a young genetically isolated population.** *Ann Hum Genet* 2005, **69(Pt 3)**:288-295.
21. Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, Morris AP: **Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms.** *Am J Hum Genet* 2008, **83**:112-119.
22. Hao K, Chudin E, McElwee J, Schadt EE: **Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies.** *BMC Genet* 2009, **10**:27.
23. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: **Merlin-rapid analysis of dense genetic maps using sparse gene flow trees.** *Nat Genet* 2002, **30**:97-101.
24. Perez-Enciso M, Misztal I: **Qxpak: a versatile mixed model application for genetical genomics and QTL analyses.** *Bioinformatics* 2004, **20(16)**:2792-2798.
25. Kim W, Gordon D, Sebat J, Ye KQ, Finch SJ: **Computing power and sample size for case-control association studies with copy number polymorphism: application of mixture-based likelihood ratio test.** *PLoS One* 2008, **3(10)**:e3475.
26. Axenovich TI, Zorkoltseva IV, Belonogova NM, Struchalin MV, Kirichenko AV, Kayser M, Oostra BA, van Duijn CM, Aulchenko YS: **Linkage analysis of adult height in a large pedigree from a Dutch genetically isolated population.** *Hum Genet* 2009, **126(3)**:457-471.
27. Estrada K, Krawczak M, Schreiber S, van Duijn K, Stolk L, van Meurs JBJ, Liu F, Penninx BWJH, Smit JH, Vogelzangs N, Hottenga JJ, Willemsen G, de Geus EJC, Lorentzon M, von Eller-Eberstein H, Lips P, Schoor N, Pop V, de Keijzer J, Hofman A, Aulchenko YS, Oostra BA, Ohlsson C, Boomsma DI, Uitterlinden AG, van Duijn CM, Rivadeneira F, Kayser M: **A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation.** *Hum Mol Genet* 2009, **18(18)**:3516-3524.
28. Woodward OM, Koettgen A, Coresh J, Boerwinkle E, Guggino WB, Koettgen M: **Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout.** *Proc Natl Acad Sci USA* 2009, **106(25)**:10338-10342.
29. Heard-Costa NL, Zillikens MC, Monda KL, Johansson A, Harris TB, Fu M, Haritunians T, Feitosa MF, Aspelund T, Eiriksdottir G, Garcia M, Launer LJ, Smith AV, Mitchell BD, McArdle PF, Shuldiner AR, Bielinski SJ, Boerwinkle E, Brancati F, Demerath EW, Pankow JS, Arnold AM, Chen YDI, Glazer NL, McKnight B, Psaty BM, Rotter JI, Amin N, Campbell H, Gyllensten U, Pattaro C, Pramstaller PP, Rudan I, Struchalin M, Vitart V, Gao X, Kraja A, Province MA, Zhang Q, Atwood LD, Dupuis J, Hirschhorn JN, Jaquish CE, O'Donnell CJ, Vasan RS, White CC, Aulchenko YS, Estrada K, Hofman A, Rivadeneira F, Uitterlinden AG, Witteman JCM, Oostra BA, Kaplan RC, Gudnason V, O'Connell JR, Borecki IB, van Duijn CM, Cupples LA, Fox CS, North KE: **NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium.** *PLoS Genet* 2009, **5(6)**:e1000539.
30. Vink JM, Smit AB, de Geus EJC, Sullivan P, Willemsen G, Hottenga JJ, Smit JH, Hoogendijk WJ, Zitman FG, Peltonen L, Kaprio J, Pedersen NL, Magnusson PK, Spector TD, Kyvik KO, Morley KI, Heath AC, Martin NG, Westendorp RGJ, Slagboom PE, Tiemeier H, Hofman A, Uitterlinden AG, Aulchenko YS, Amin N, van Duijn C, Penninx BW, Boomsma DI: **Genome-wide association study of smoking initiation and current smoking.** *Am J Hum Genet* 2009, **84(3)**:367-379.

doi:10.1186/1471-2105-11-134
Cite this article as: Aulchenko *et al.*: ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 2010 **11**:134.