ProbABEL manual

Yurii Aulchenko, Maksim Struchalin, Lennart Karssen
Erasmus MC Rotterdam

October 11, 2010

Contents

1 Motivation

Many statistical and experimental techniques, such as imputations and high-throughput sequencing, generate data which are informative for genome-wide association analysis and are probabilistic in the nature.

When we work with directly genotyped markers using such techniques as SNP or microsatellite typing, we would normally know the genotype of a particular person at a particular locus with very high degree of confidence, and, in case of biallelic marker, can state whether genotype is AA, AB or BB.

On the contrary, when dealing with imputed or high-throughput sequencing data, for many of the genomic loci we are quite uncertain about the genotypic status of the person. Instead of dealing with known genotypes we work with a probability distribution that is based on observed information, and we have estimates that true underlying genotype is either AA, AB or BB. The degree of confidence about the real status is measured with the probability distribution $\{P(AA), P(AB), P(BB)\}$.

Several techniques may be applied to analyse such data. The most simplistic approach would be to pick up the genotype with highest probability, i.e. $\max_g[P(g = AA), P(g = AB), P(g = BB)]$ and then analyse the data as if directly typed markers were used. The disadvantage of this approach is that it does not take into account the probability distribution – i.e. the uncertainty about the true genotypic status. Such analysis is statistically wrong: the estimates of association parameters (regression coefficients, odds or hazard ratios, etc.) are biased, and the bias becomes more pronounced with greater probability distribution uncertainty (entropy).

One of the solutions that generate unbiased estimates of association parameters and takes the probability distribution into account is achieved by performing association analysis by means of regression of the outcome of interest onto estimated genotypic probabilities.

The `ProbABEL` package was designed to perform such regression in a fast, memory-efficient and consequently genome-wide feasible manner. Currently, `ProbABEL` implements linear, logistic regression, and Cox proportional hazards models. The corresponding analysis programs are called `palinear`, `palogist`, and `pacoxph`.

2 Input files

`ProbABEL` takes three files as input: a file containing SNP information (e.g. the MLINFO file of MACH), a file with genome- or chromosome-wide predictor information (e.g. the MLDOSE or MLPROB file of MACH), and a file containing the phenotype of interest and covariates.

Optionally, the map information can be supplied (e.g. the "legend" files of HapMap).

The dose/probbaility file may be supplied in filevector format in which case `ProbABEL` will operate much faster, and in low-RAM mode (approx. 128 MB). See the R libraries GenABEL and DatABEL on how to convert MACH and IMPUTE files to filevector format (functions: `mach2databel()` and `impute2databel()`, respectively).

2.1 SNP information file

In the simplest scenario, the SNP information file is an MLINFO file generated by MACH. This must be a space or tab-delimited file containing SNP name, coding for allele 1 and 2 (e.g. A, T, G or C), frequency of allele 1, minor allele frequency and two quality metrics ("Quality", the average maximum posterior probability and "Rsq", the proportion of variance decrease after imputations).

Actually, for `ProbABEL`, it does not matter what is written in this file – this information is just brought forward to the output. However, **it is critical** that the number of columns is seven and the number of lines in the file is equal to the number of SNPs in the corresponding DOSE file (plus one for the header line).

The example of SNP information file content follows here (also to be found in `ProbABEL/example/test.mlinfo`)

Note that header line is present in the file. The file describes five SNPs.

2.2 Genomic predictor file

Again, in the simplest scenario this is an MLDOSE or MLPROB file generated by MACH. Such file starts with two special columns plus, for each of the SNPs under consideration, a column containing the estimated allele 1 dose (MLDOSE). In an MLPROB file, two columns for each SNP correspond to posterior probability that person has two ($P_{A_1A_1}$) or one ($P_{A_1A_2}$) copies of allele 1. The first "special" column is made of the sequential id, followed by an arrow followed by study ID (the one specified in the MACH input files). The second column contains the method keyword (e.g. "MLDOSE").

An example of the few first lines of an MLDOSE file for five SNPs described in SNP information file follows here (also to be found in the file `../example/test.mldose`)

The order of SNPs in the SNP information file and DOSE-file must be the same. This should be the case if you just used MACH outputs.

Therefore, by all means, the number of columns in the genomic predictor file must be the same as the number of lines in the SNP information file plus one.

2.3 Phenotypic file

The phenotypic data file contains phenotypic data, but also specifies the analysis model. There is a header line, specifying the variable names. The first column should contain personal study IDs. It is assumed that **both the total number and the order of these IDs are exactly the same as in the genomic predictor (MLDOSE) file described in previous section**. This is not difficult to arrange using e.g. `R`; an example is given in the `ProbABEL/examples` directory.

Missing data should be coded with 'NA', 'N' or 'NaN' codes. Any other coding will be converted to some number which will be used in analysis! E.g. coding missing as '-999.9' will result in an analysis which will consider -999.9 as indeed a true measurements of the trait/covariates.

In the case of linear or logistic regression (programs `palinear` and `palogist`, respectively), the second column specifies the trait under analysis, while the third, fourth, etc. provide information on covariates to be included into analysis. An example few lines of phenotypic information file designed for linear regression analysis follow here (also to be found in `../example/height.txt`)

```
id height sex age
id636728 174.429795159687 0 56.5664877162697
id890314 168.176943059097 0 74.8311971509938
id102874 178.612190619767 1 45.2478051768211
id200949 171.770230117638 0 46.7362651142108
id336491 185.941629656499 1 61.2743318817997
id988766 173.159286450017 1 43.9794924518567
id21999 167.478282481124 0 64.842094190157
id433893 168.33178468379 1 49.2526444099125
id688932 171.691587811178 0 50.3954417563357
id394203 173.491495887183 1 71.6498502881161
```

Note again that the order of IDs is the same between the MLDOSE file

and the phenotypic data file. The model specified by this file is $height \sim \mu + \text{sex} + \text{age}$, where μ is the intercept.

Clearly, you can for example include `sex` $\times$ `age` interaction terms by specifying another column having a product of sex and age here.

For logistic regression, it is assumed that in the second column cases are coded as "1" and controls as "0". An couple of example lines of a phenotypic information file designed for logistic regression analysis follow here (also to be found in `../example/logist_data.txt`)

```
id chd sex age othercov
id636728 0 0 56.5664877162697 -0.616649220436139
id890314 0 0 74.8311971509938 0.695315865158652
id102874 1 1 45.2478051768211 -0.919192364890525
id200949 0 0 46.7362651142108 -0.623212536893650
id336491 0 1 61.2743318817997 -0.0835744351009496
id988766 0 1 43.9794924518567 -0.360419162609288
id21999 1 0 64.842094190157 -0.180940346913155
id433893 0 1 49.2526444099125 0.126374731789777
id688932 0 0 50.3954417563357 1.06437576032067
id394203 1 1 71.6498502881161 -1.18226498491599
```

You can see that in the first 10 people, there are three cases, as indicated by "chd" equal to one. The model specified by this file is chd $\sim \mu + \text{sex} + \text{age} + \text{othercov}$.

In case of the Cox proportional hazards model, the composition of the phenotypic input file is a bit different. In the second column and third column, you need to specify the outcome in terms of follow-up time (column two) and event (column three, "1" if an event occurred and zero if censored). Columns starting from four (inclusive) specify covariates to be included into the analysis. An example few lines of a phenotypic information file designed for the Cox proportional hazards model analysis follow here (also to be found in `../example/coxph_data.txt`)

```
id fupt_chd chd sex age othercov
id636728 3.187930645 0 0 56.56648772 -0.61664922
id890314 2.099691952 0 0 74.83119715 0.695315865
id102874 9.133488079 1 1 45.24780518 -0.919192365
id200949 7.525406804 0 0 46.73626511 -0.623212537
id336491 6.798229522 0 1 61.27433188 -0.083574435
id988766 6.149545358 0 1 43.97949245 -0.360419163
id21999 1.013546103 1 0 64.84209419 -0.180940347
```

```
id433893 1.282853098 0 1 49.25264441 0.126374732
id688932 8.340206657 0 0 50.39544176 1.06437576
id394203 3.392345681 1 1 71.64985029 -1.182264985
```

You can see that for the first ten people, the event occurs for three of them, while for the other seven there is no event during the follow-up time, as indicated by the "chd" column. Follow-up time is specified in the preceding column. The covariates included into the model are age (presumably at baseline), sex and "othercov"; thus the model, in terms of `R/survival` is $\text{Surv}(\text{fuptime\_chd}, \text{chd}) \sim \text{sex} + \text{age} + \text{othercov}$.

2.4 Optional map file

If you would like map information (e.g. base pair position) to be included in your outputs, you can supply a map file. These follow HapMap "legend" file format. For example, for the five SNPs we considered the map-file may look like

The order of the SNPs in the map file should follow that in the SNP information file. Only information from the second column – the SNP location – is actually used to generate the output.

3 Running an analysis

To run linear regression, you should use program called `palinear`; for logistic analysis use `palogist`, and for the Cox proportional hazards model use `pacoxph` (to be found in the `ProbABEL/bin/` directory after you have compiled the program).

There are in total 11 command line options you can specify to the `ProbABEL` analysis functions `linear` or `logistic`. If you run either program without any argument, you will get a short explanation to command line options:

```
user@server:~$ palogist

Usage: ../bin/palogist options

Options:
--pheno       : phenotype file name
--info        : information (e.g. MLINFO) file name
--dose        : predictor (e.g. MLDOSE/MLPROB) file name
--map         : [optional] map file name
--nids        : [optional] number of people to analyse
```

```
--chrom       : [optional] chromosome (to be passed to output)
--out         : [optional] output file name (default is regression.out.txt)
--skipd       : [optional] how many columns to skip in predictor
    (dose/prob) file (default 2)
--ntraits     : [optional] how many traits are analysed (default 1)
--ngpreds     : [optional] how many predictor columns per marker
    (default 1 = MLDOSE; else use 2 for MLPROB)
--separat     : [optional] character to separate fields (default is space)
--score       : use score test
--no-head     : do not report header line
--allcov      : report estimates for all covariates (large outputs!)
--interaction : which covariate to use for interaction with SNP
  (default is no ineraction, 0)
--mmscore     : score test for association between a trait and genetic
                    polymorphism, in samples of related individuals
--robust      : report robust (aka sandwich, aka Hubert-White) standard
                    errors
--help        : print help
```

3.1 Basic analysis options

However, for a simple run you can use only three, which specify the necessary files needed to run the regression analysis.

These options are `--dose` (or `-d`), specifying the genomic predictor/MLDOSE file described in sub-section 2.2; `--pheno` (or `-p`), specifying the phenotypic data file described in sub-section 2.3; and `--info` (or `-i`), specifying the SNP information file described in sub-section 2.1.

If you change to the `ProbABEL/example` directory you can run an analysis of height by running

```
user@server:~/ProbABEL/example/$ ../bin/palinear -p height.txt
-d test.mldose -i test.mlinfo
```

Output from the analysis will be directed to the `regression.out.csv` file.

The analysis of a binary trait "chd" can be run with

```
user@server:~/ProbABEL/example/$ ../bin/palogist -p logist_data.txt
                                 -d test.mldose -i test.mlinfo
```

To run a Cox proportional hazards model, try

```
user@server:~/ProbABEL/example/$ ../bin/pacoxph -p coxph_data.txt
                                 -d test.mldose -i test.mlinfo
```

Please have a look at the shell script files `example_qt.sh`, `example_bt.sh` and `example_all.sh` to have a better overview of the analysis options.

To run an analysis with MLPROB files, you need specify the MLPROB file with the `-d` option and also specify that there are two genetic predictors per SNP, e.g. you can run linear model with

```
user@server:~/ProbABEL/example/$ ../bin/palinear -p height.txt
                                 -d test.mlprob -i test.mlinfo
                                 --ngpreds=2
```

3.2 Advanced analysis options

The option `--interaction` allows you to include interaction between SNP and any covariate. If for example your model is

$$\text{trait} \sim \text{sex} + \text{age} + \text{SNP},$$

running the program with the option `--interaction=2` will model

$$\text{trait} \sim \text{sex} + \text{age} + \text{SNP} + \text{age} \times \text{SNP}.$$

The option `--robust` allows you to compute so-called "robust" (a.k.a. "sandwich", a.k.a. Hubert-White) standard errors (cf. section "Methodology" for details).

With the option `--mmscore` a score test for association between a trait and genetic polymorphism in samples of related individuals is performed. A file with the inverse of the variance-covarince matrix goes as input parameter with that key, e.g. `--mmscore <filename>`. The file has to contain the first column with id names exactly like in phenotype file, BUT OMITTING people with no measured phenotype. The rest is a matrix. The phenotype file in case of using the `--mscore` argument may contain any amount of covariates (this is different from previous versions). The first is id names, the second - trait. The others are covariates.

An example of how a polygenic object estimated by GenABEL can be used with ProbABEL is provided here: `../example/mmscore.R`

Though technically `--mmscore` allows for inclusion of multiple covariates, these should be kept to minimum as this is score test. We suggest that any covariates explaining an essential proportion of variance should be fit as part of `GenABEL`'s `polygenic` procedure.

3.3 Running multiple analyses at once: probabel.pl

The Perl script `bin/probabel.pl_example` represents a handy wraper for `ProbABEL` functions. To start using it the configuration file `bin/probabel_config.cfg_example` needs to be edited. The configuration file consists of five columns. Each column except the first is a pattern for files produced by `MACH` (imputation software). The column named "cohort" is an identifying name of a population ("ERGO" in this example), the column "mlinfo_path" is the full path to mlinfo files, including a pattern where the chromosome number has been replaced by `_._chr_._`. The columns "mldose_path", "mlprobe_path" and "legend_path" are paths and patterns for "mldose", "mlprob" and "legend" files, respectively. These also need to include the pattern for the chromosome as used in the column for the "mlinfo" files. Probably you also have to change the variable `$config` in the script to point to the full path of the configuration file and the variable `@anprog` to point full path to the `ProbABEL` scripts.

4 Output file format

Let us consider what comes out of the linear regression analysis described in the previous section. After the analysis has run, in the output file you will find something like

```
name A1 A2 Freq1 MAF Quality Rsq n Mean_predictor/2 chrom position
...beta_mu beta_sex beta_age beta_SNP
......sebeta_mu sebeta_sex sebeta_age sebeta_SNP
.........sigma2 SNP_Z SNP_chi2
rs7247199 G A 0.5847 0.415 0.9299 0.8666 182 0.56444 19 204938
...171.113 10.559 -0.054599 -0.218693
......2.18584 0.974984 0.0340798 0.734966
.........41.938 -0.297555 0.0885391
rs8102643 C T 0.5847 0.415 0.9308 0.8685 182 0.564412 19 207859
...171.112 10.559 -0.0545991 -0.218352
......2.1855 0.974987 0.0340799 0.734214
.........41.938 -0.297396 0.0884444
```

Here, only the first three lines of output have been shown. Note that lines starting with `...` are actually the ones continuing the previous line – they have just been wrapped this output so we can see these long lines.

The header provides a short description of what can be found in a specific column. The first column provides the SNP name and next six are descriptions which were brought directly from the SNP information file. Therefore,

these describe allele frequencies and the quality in your total imputations, not necessarily in the data under analysis.

In contrast, starting with the next column, named `n`, the output concerns the data analysed. Column 8 (`n`) tells the number of subjects for whom complete phenotypic information was available. At this point, unless you have complete measurements on all subjects, you should feel alarmed if the number here is exactly the number of people in the file – this probably indicates you did not code missing values according to `ProbABEL` format ('NA', 'NaN', or 'N').

The next column, nine ("Mean_predictor_allele"), gives the estimated frequency of the predictor allele in subjects with complete phenotypic data.

If the `--chrom` option was used, in the next column you will find the value specified by this option. If `--map` option was used, in the subsequent column you will find map location brought from the map-file. The Next columns provide coefficients of regression of the phenotype onto genotype corresponding standard errors, and log-likelihood of the model at the point of MLEs.

5 Preparing input files

In the `ProbABEL/bin` directory you can find the `perepare_data.R` file – an R script that arranges phenotypic data in right format. Please read this script for details.

6 Memory use and performance

Maximum likelihood regression is implemented in `ProbABEL`. With 6,000 people and 2.5 millions SNPs, a genome-wide scan is completed in less that an hour for linear model with 1-2 covariates and overnight for logistic regression or the Cox proportional hazards model.

Memory is an issue with `ProbABEL` – large chromosomes, such as chromosome one consumed up to 5 GB of RAM with 6,000 people.

7 Methodology

7.1 Analysis of population-based data

7.1.1 Linear regression assuming normal distribution

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

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

and variance-covariance matrix

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

where $\mathbf{Y}$ is the vector of phenotypes of interest, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of regression parameters, σ^2 is the variance and $\mathbf{I}$ is identity matrix.

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

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2}$$

and the MLE of the residual variance is

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})}{N - r_X}, \tag{3}$$

where N is the number of observations and r_X is the rank of $\mathbf{X}$ (i.e. the number of columns of the design matrix).

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

$$\mathbf{var}_{\hat{\beta}} = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{4}$$

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

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

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 first 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 $\beta_{g,0}$.

Firstly, the likelihood ratio test can be obtained with

$$LRT = 2 \cdot (logLik(\hat{\beta}_g, \hat{\beta}_x) - logLik(\beta_{g,0}, \hat{\beta}_x))$$

which under the null hypothesis is asymptotically distributed as χ^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

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

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

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

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

The Wald test statistics is then computed as

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

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 $(\beta_{g,0}, \hat{\beta}_x)$.

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

7.1.2 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

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

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:

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

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

7.1.3 Robust variance-covariance matrix of parameter estimates

For linear model, these are 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 $\mathbf{R}$ is a diagonal matrix containing squares of residuals of $\mathbf{Y}$. The same formula may be used for "standard" analysis, in which case the elements of the $\mathbf{R}$ matrix are constant, namely mean residual sum of squares (the estimate of σ^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 $\mathbf{1}$ is the vector of ones and $\mathbf{W}$ is the diagonal matrix of "weights" used in logistic regression.

7.1.4 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.

Many thanks to Thomas for making his code available under GNU GPL!

7.2 Analysis of pedigree data

The framework for analysis of pedigree data follows the two-step logic developed in the works of Aulchenko et al. (2007) and Chen and Abecasis (2007).

General analysis model is a linear mixed model which defines the expectation of the trait as

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

identical to that defined for linear model (1). To account for correlations between the phenotypes of relatives which may be induced by family relations the variance-covariance matrix is defined to be proportional to the linear combination of the identity matrix $\mathbf{I}$ and the relationship matrix $\mathbf{\Phi}$:

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

where h^2 is the heritability of the trait. The relationship matrix $\mathbf{\Phi}$ is twice the matrix containing the coefficients of kinship between all pairs of individuals under consideration; its estimation is discussed in a separate section "7.2.2" (7.2.2).

Estimation of thus defined model is possible by numerical maximization of the likelihood function, however, the estimation of this model for large pedigrees is laborious, and is not computationally feasible for hundreds of thousands to millions of SNPs to be tested in the context of GWAS, as we have demonstrated previously (Aulchenko et al., 2007).

7.2.1 Two-step score test for association

A two-step score test approach is therefore used to decrease the computational burden. Let us first re-define 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, $\mathbf{X_g}$:

$$E[\mathbf{Y}] = \mathbf{X}_x\beta_x + \mathbf{X}_g\beta_g$$

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

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

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

$$logLik(\beta_x, h^2, \sigma^2) = -\frac{1}{2}(log_e|\mathbf{V}_{\sigma^2,h^2}| + (\mathbf{Y} - \beta_x\mathbf{X}_x)^T\mathbf{V}^{-1}_{\sigma^2,h^2}(\mathbf{Y} - \beta_x\mathbf{X}_x))$$

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

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

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

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

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

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

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

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 $\hat{\beta}_g$ can be tested with

$$T^2_j = \hat{\beta}^2_g(j) \mathbf{var}^{-1}_{\hat{\beta}_g}(jj)$$

where $\hat{\beta}^2_g(j)$ is square of the j-th element of the vector of estimates $\hat{\beta}_g$, and $\mathbf{var}^{-1}_{\hat{\beta}_g}(jj)$ corresponds to the j-th diagonal element of $\mathbf{var}^{-1}_{\hat{\beta}_g}$. The latter statistics asymptotically follows χ^2_1.

7.2.2 Estimation of the kinship matrix

The relationship matrix $\mathbf{\Phi}$ used in estimation of the linear mixed model for pedigree data 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. For an outbred person, the kinship coefficient is 1/2 – that is two gametes sampled from this person at

random are IBD only if the same gamete is sampled. However, if the person is inbred, there is a chance that a maternal and paternal chromosomes are also IBD. The probability of this is characterized by kinship between individual's parents, which is defined as the individual's inbreeding coefficient, F. In this case, the kinship coefficient for the individual is $F + 1/2$. Similar logic applies to computation of the kinship coefficient for other types of pairs in inbred pedigrees.

The kinship matrix can be computed using the pedigree data using standard methods. However, in many cases, pedigree information may be absent, incomplete, or not reliable. Moreover, the estimates obtained using pedigree data reflect the expectation of the 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 can be obtained (Astle and Balding, 2010; Amin *et al.*, 2007) 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. The frequency is computed for the allele which, when homozygous, corresponds to the genotype coded as "1".

8 How to cite

As of May 2008, we have not yet published a `ProbABEL` paper. If you used `ProbABEL` for your analysis please give a link to the `ABEL` home page

`http://mga.bionet.nsc.ru/~yurii/ABEL/`

and cite the `GenABEL` paper to give us some credit:

> Aulchenko YS, Ripke S, Isaacs A, van Duijn CM.
> *GenABEL: an R library for genome-wide association analysis.*
> Bioinformatics. 2007 23(10):1294-6.

A proper reference may look like

> For the analysis of imputed data, we used the `ProbABEL` package from the ABEL set of programs (Aulchenko *et al.*, 2007).

If you have used the Cox proportional hazard model, please mention the `R` package `survival` by Thomas Lumley. Additionally to the above citation, please tell that

> The Cox proportional hazards model implemented in `ProbABEL` makes use of the source code of the `R` package "`survival`" as implemented by T. Lumley.