ProbABEL manual

Yurii Aulchenko, Maksim Struchalin
Erasmus MC Rotterdam

September 22, 2009

Contents

1 Motivation

Many statistical and experimental techniques, such as imputations and high-throughput sequencing, generate the 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 contract, when dealing with imputed or high-throughput sequencing data, for many of genomic loci we are quite uncertainty about genotypic status of the person; instead of known genotypes we deal rather with probability distribution; that is based on observed information, we have estimates that true underlying genotype is either AA, AB or BB; the degree of confidence about the real status is measured with 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 solution which generates unbiased estimates of association parameters and takes probability distribution into account is achieved by performing association analysis by means of regression of the outcome of interest onto estimated genotypic probabilities.

`ProbABEL` package was designed to perform such regression in 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. MLINFO file of MACH), file with genome- or chromosome-wide predictor information (e.g. MLDOSE or MLPROB file of MACH), and a file containing phenotype of interest and covariates.

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

2.1 SNP information file

In the simplest scenario, 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" = average maximum posterior probability and "Rsq" – 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 under simplest scenario this is a 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 MLPROB file, two columns for each SNP correspond to posterior probability that person has one or two 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 MACH input files). The second column contains method (e.g. "MLDOSE") keyword.

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 `ProbABEL/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.

Thus, 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

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 is 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`; example is given in `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 analysis which will consider -999.9 as indeed true measurements of the trait/covariates.

In 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 `ProbABEL/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 MLDOSE and

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

Clearly, you can for example include `sex x 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 example few lines of phenotypic information file designed for logistic regression analysis follow here (also to be found in `ProbABEL/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 $chd \sim \mu + sex + age + othercov$.

In case of 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 event occurred and zero if censoring). Columns from four inclusive specify covariates to be included into analysis. An example few lines of phenotypic information file designed for Cox proportional hazards model analysis follow here
(also to be found in `ProbABEL/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 first 10 people, event happens for three of them, while for the other seven there is no event during follow-up time, as indicated by "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
$Surv(fuptime\_chd, chd) \sim sex + age + othercov$.

2.4 Optional map file

If you would like that 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 analysis

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

There are in total 11 command line options you can specify to `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)
--interaction_only: like previos but without covariate acting in
                    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 regression analysis.

These options are `--dose` (or `-d`), specifying 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 analysis

of height by

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

Output from analysis will be directed to "regression.out.csv" file.
You can run analysis of binary trait "chd" 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 analysis options.

To run analysis with MLPROB files, you need specify the MLPROB file with -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

Option `--interaction` allows you to include interaction between SNP and any covariate. If e.g. your model is `trait sex + age + SNP`, running the program with option `--interaction=2` will model `trait sex + age + SNP + age*SNP`.

Option `--robust` allows you to compute so-called "robust" (aka "sandwich", aka Hubert-White) standard errors (see section "Methodology" for details).

With option `--mmscore` score test for association between a trait and genetic polymorphism in samples of related individuals is performed. File with inverse of variance-covarince matrix goes as input parameter with that key. Like `--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. Phenotype file in case of using key –mscore may contain any amount of covariates (opposed to previous versions). The first is id names, the second - trait. Otheres are covariates.

An example how 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 essential proportion of variance should be fit as part of the `GenABEL`'s `polygenic` procedure.

Option `--interaction_only` is like `--interaction` but does not include in the model the main effect of the covariate, which is acting in interaction with SNP. This option is useful when running `--mmscore`, in whch case the main effect should normally estimated in the polygenic model and only the interaction term in the `ProbABEL` analysis.

3.3 Running multiple analyses at once: probabel.pl

Perl script `bin/probabel.pl_example` represents a handy wraper for ProbABEL functions. To start using it you have to change config file `bin/probabel_config.cfg_example`. Configuration file consists of 5 columns. Each column except of the first is pattern for files produced by `MACH` (imputation software). Column named "cohort" is name of population ("ERGO" in this example), column "mlinfo_path" – full path to mlinfo files and pattern of name where chromosome number has been replaced by "_._chr_._". Columns "mldose_path", "mlprobe_path" and "legend_path" are paths and patterns for "mldose", "mlprob" and "legend" files. Probably you also have to change variable `$config` in sript to point full path to configuration file and variable `@anprog` to point full path to ProbABEL scripts.

4 Output file format

Let us consider what comes out of the linear regression analysis described in the above section. After you have run the analysis, 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, I show only three first lines of output. Note that lines starting with "..." are actually the ones continuing the previous line – I just have wrapped this output so we can see these long lines.

The header provides short description of what can be found in the specific column. The first column provides the SNP name and next six are descriptives which were brought directly from the SNP information file. Thus these describe allele frequencies and quality in your total imputations, not necessarily in the data under analysis.

On the 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 warned 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 you estimated frequency of the predictor allele in subjects with complete phenotypic data.

If "–chrom" option was used, in the next column you will find the value specified by this option. If "–map" option was used, in next column you will find map location brought from the map-file. Next columns provide coefficients of regression of the phenotype onto genotype, corresponding standard errors, and the χ^2 of the Likelihood Ratio Test for deviation from zero.

5 Preparing input files

In the `ProbABEL/bin` directory you can find `perepare_data.R` file – an R script which 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, genome-wide scan is completed in less that an hour for linear model with 1-2 covariates and overnight for logistic regression or Cox proportional hazards model.

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

7 Methodology

7.1 Linear regression

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

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

and variance-covaraince matrix

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

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

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

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

and MLE of the 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 rank of $\mathbf{X}$ (number of columns of the design matrix).

Standard errors for the j-th parameter can be obtained as

$$s.e.(\hat{\beta}_j) = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj} \tag{4}$$

where $(\mathbf{X}^T\mathbf{X})^{-1}_{jj}$ stands for the j-th diagonal element of the inverse of matrix $(\mathbf{X}^T\mathbf{X})$.

7.2 Logistic regression

Standard methodology based on iteratively re-weighted least squares is used to obtain parameters' estimates.

7.3 Cox proportional hazards model

The code of this section is entirely based on the code of `R` library `survival` code developed by Thomas Lumley (function `coxfit2`).

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

7.4 Robust standard errors

These are computed using formula

$$((\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{V}\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1})_{jj}$$

where $\mathbf{V}$ is a diagonal marix containing residuals. The same formula may be used for "standard" analysis, in which case the elements of the $\mathbf{V}$ matrix are constant, namely mean residual sum of squares (the estimte of σ^2).

8 How to cite

As for May 2008, we have not yet published `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 `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.

Proper reference may look like

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

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

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