

# **lodGWAS: A software package for genome-wide association analysis of biomarkers with a limit of detection**

## **User Manual**

*Version 1.0-0*

*February 2015*

### **Contents**

|                             |          |
|-----------------------------|----------|
| <b>Introduction.....</b>    | <b>2</b> |
| <b>Input Formats .....</b>  | <b>2</b> |
| Genotype Files .....        | 2        |
| Phenotype File .....        | 4        |
| <b>Getting started.....</b> | <b>6</b> |
| Function lod_QC .....       | 6        |
| Function lod_GWAS .....     | 7        |
| <b>Output format.....</b>   | <b>9</b> |

## Introduction

Statistical analysis of a biomarker is often complicated because the detection range of the assay of the biomarker is restricted. The limits of detection (LOD) are floor and/or ceiling values of the biomarker that can be accurately measured by a particular assay type. Any value of the biomarker beyond the range of LOD, either smaller than the lower LOD or larger than the upper LOD, cannot be determined accurately. Those observations, so-called non-detects (NDs), cannot simply be excluded from the analysis, because NDs are not '*missing at random*'. They can be considered as censored data, and can therefore be best analyzed by using statistical methods for survival analysis.

lodGWAS is a flexible R package that is capable of performing a genome-wide association study (GWAS) accommodating the problem of LOD. It treats NDs as censored data, either left- or right-censored or both, and performs a parametric survival analysis on the phenotype of interest that includes both measured and censored values.

## Input Formats

A GWAS analysis with the lodGWAS R package requires two files for the genotypes and one phenotype file. Their formats are described below. All input files (genotype and phenotype files) can be either space or tab delimited. The input files may be compressed in the gzip format (extension .gz).

### Genotype Files

lodGWAS uses the [PLINK dosage format](#) for the genotype data. This means that two files are needed: one with the genotypes (*genotype dosage file*) and one with the locations of the genetic variants (*map file*).

### Genotype Dosage File

The genotype dosage file should contain a header line. The header line (first line) should be:

```
SNP A1 A2 FID1 IID1 FID2 IID2 ... FIDn IIDn
```

The first three columns must appear before the dosage data. The following columns are the family identifier (FID) and the individual identifier (IID) of individuals 1 to n. Thus, the number of columns of the header line should be exactly:  $3+(2 \times n\_individuals)$ .

The next lines contain the actual genetic data per individual, with each row corresponding to a genetic variant. The PLINK dosage format can be any of three formats: dosage, two-probabilities, or three-probabilities (see below). lodGWAS is capable of handling all three plink dosage formats. It will automatically recognize whether there is one (dosage), two (two-probabilities), or three (three-probabilities) columns per individual. In case of any other format it will report that it cannot recognize the data format and will not run.

**Dosage format:** a dosage is provided in one column per individual. Each dosage is a number between 0 and 2. A dosage of 0, 1, or 2 means that the individual is homozygous for the A2 allele, heterozygous, or homozygous for the A1 allele, respectively. When the genetic dataset is expanded using imputation, non-integer values are also possible, and are defined as the weighted sum of genotype probabilities (i.e.  $0 \times \text{Prob}(A2/A2) + 1 \times \text{Prob}(A1/A2) + 2 \times \text{Prob}(A1/A1)$ ). The number of columns of the (non-header) lines in a genotype file in dosage format should be exactly  $3+n\_individuals$ .

Example of the dosage format:

| SNP    | A1 | A2 | FID1 | IID1 | FID2 | IID2 | FID3 | IID3 |
|--------|----|----|------|------|------|------|------|------|
| rs0001 | A  | C  | 0.08 |      | 0.72 |      | 1.99 |      |

**Two-probabilities format:** two numbers, representing the probabilities of the A1/A1 and A1/A2 genotypes, respectively. The probability of A2/A2 can be determined by 1 minus the sum of Prob(A1/A1) and Prob(A1/A2). Each probability is a number between 0 and 1. The number of columns of the (non-header) lines in a genotype file in two-probabilities format should be exactly  $3+(2 \times n\_individuals)$ .

Example of the two-probabilities format:

| SNP    | A1 | A2 | FID1 | IID1 | FID2 | IID2 |
|--------|----|----|------|------|------|------|
| rs0001 | A  | C  | 0.97 | 0.02 | 0.88 | 0.10 |

**Three-probabilities format:** three numbers, representing the probabilities of the A1/A1, A1/A2, and A2/A2 genotypes, respectively. Each probability is a number between 0 and 1, and the three probabilities per genetic variant per individual should add up to 1. The number of columns of the (non-header) lines in a genotype file in three-probabilities format should be exactly  $3+(3 \times n\_individuals)$ .

Example of the three-probabilities format:

| SNP    | A1 | A2 | FID1 | IID1 | FID2 | IID2           |
|--------|----|----|------|------|------|----------------|
| rs0001 | A  | C  | 0.97 | 0.02 | 0.01 | 0.88 0.10 0.02 |

## Genotype Map File

The genotype map file contains the locations of the genetic variants, with each row of the file corresponding to a variant. It must contain four columns:

1. Chromosome (1-22, X, Y or 0 if unspecified)
2. Marker ID (identifier of the genetic variant)
3. Genetic distance (Morgan, this is not used by lodGWAS; so the actual value doesn't matter)
4. Physical position (base-pair position)

**Note :** Unlike the dosage file, the map file has no header line.

Example of a genotype map file:

```
1 rs0001 0 111111
1 rs0002 0 222222
1 rs0003 0 333333
2 rs0004 0 121212
2 rs0005 0 343434
```

## Phenotype File

The phenotype file is a text file containing the non-genetic data with each row of the file corresponding to an individual.

- This file should have a header line.
- The phenotype file should contain at least the following variables: family ID, individual ID, phenotype, and type of value (i.e. whether or not the phenotype is outside the LODs). Other columns, e.g. for covariates, are optional.
- The header (name) of columns for family ID, for individual ID, and for type of value should exactly be “FID”, “IID”, and “outsideLOD”, respectively (note that R is case sensitive). The other columns (phenotype and cov1 to covN) can have any arbitrary name.
- The order of columns is not important.
- The order of the rows (samples) is not important.

Example of a phenotype file:

| FID | IID | Phenotype | outsideLOD | cov1 | cov2 | ... | covN    |
|-----|-----|-----------|------------|------|------|-----|---------|
| 1   | 1   | 0.1       | 2          | 43   | 1    | ... | 0.025   |
| 2   | 2   | 0.1       | 2          | 38   | 2    | ... | 0.036   |
| 3   | 3   | 1.34      | 1          | 62   | 2    | ... | -0.008  |
| 4   | 4   | 1.70      | 1          | 56   | 1    | ... | 0.124   |
| 5   | 5   | 2.85      | 1          | 45   | 2    | ... | -0.148  |
| 6   | 6   | 27        | 0          | 51   | 1    | ... | 0.001   |
| 7   | 7   | 27        | 0          | 55   | 1    | ... | 0.00189 |

In this example, we assume the lower and upper LOD of the assay are 0.1 and 27, respectively.

### Column descriptions of phenotype file:

**FID:** family identifier of the individual. It should exactly match with FIDs of the genotype dosage file. The header (name) of this column should exactly be “FID”.

**IID:** the unique identifier of the individual within each family. It should exactly match with IIDs of the genotype dosage file. The header (name) of this column should exactly be “IID”.

**Phenotype:** the phenotype or trait of interest, which can be any numeric value. Please see below for [a few considerations regarding the phenotype](#).

**outsideLOD:** The variable outsideLOD indicates whether the phenotype value is within or beyond the range of LOD. It needs to be coded as ‘0’ if phenotype > upper LOD, ‘1’ if phenotype is within the

detection interval, and '2' if phenotype < lower LOD. Values other than '0', '1', or '2' are not accepted. The header (name) of this column should exactly be "outsideLOD".

**cov1 to covN:** covariate 1 to covariate N. The phenotype file can contain as many covariates as necessary. Some examples are: age, sex, BMI, smoking status, medication, population stratification parameters (principal components), dosage data of a particular genetic variant (for conditional analysis), study center, etc. The user can specify the analysis model by including or excluding covariates. Different models can be analyzed with different runs of lodGWAS. The user should be aware of missing values in the covariate columns. If one of the covariate columns contains a missing value, and that covariate is included in the analysis, the corresponding individual (entire row) will automatically be excluded from the analysis.

### *A few considerations regarding the phenotype*

Please pay particular attention to instructions below, as failing to heed them may cause invalid results.

1. The user must carefully distinguish between two types of missing phenotype, i.e. *missing* and *censored* values. Any mix-up between these two types will yield incorrect results.
2. **Missing phenotype values** are those phenotypes that are missing due to any reason other than being beyond the LOD. They are considered as real missing (at random). All missing values of the phenotype due to any reason, other than the problem of LOD, should be coded as "NA" for both columns of "Phenotype" and "outsideLOD".
3. **Censored phenotype values** are NDs, i.e. values of the phenotype that fall beyond the LOD of the assay, either below the lower LOD or above the upper LOD. NDs are not real missing values, since they do provide a piece of information about the distribution of the phenotype. Any ND that is below the lower LOD should be coded as the lower LOD itself, and not to any other number (and the corresponding outsideLOD value should be coded as '2'). Any ND that is above the upper LOD should be coded as the upper LOD itself, and not to any other number (and the corresponding outsideLOD value should be coded as '0'). NDs should NOT be coded as missing values (e.g. "NA", ".", "-9", etc.). LodGWAS can handle multiple lower and upper LOD levels, e.g. as a result from different assays used to measure the biomarker. In that case the phenotype variable for an ND of an individual should be exactly coded as the lower or upper LOD level of the appropriate assay type used for that individual. In the following example dataset two assays were used: one with a lower LOD of 0.5 and an upper LOD of 10, and another with a lower LOD of 0.1 and an upper LOD of 27.

| FID | IID | Phenotype | outsideLOD |
|-----|-----|-----------|------------|
| 1   | 1   | 0.1       | 2          |
| 2   | 2   | 0.5       | 2          |
| 3   | 3   | 0.5       | 1          |
| 4   | 4   | 1.3       | 1          |
| 5   | 5   | 0.4       | 1          |
| 6   | 6   | 15.0      | 1          |
| 7   | 7   | 10.0      | 0          |
| 8   | 8   | 27.0      | 0          |

In this example, individuals 2-2 (FID=2; IID=2) and 7-7 were measured with the former assay; individuals 1-1, 3-3, 5-5, 6-6, and 8-8 were measured with the latter assay; for individual 4-4 either assay could have been used.

4. The column phenotype can be either raw or transformed values of the phenotype. Some examples of transformation of phenotype are: normalization, standardization, log-transformation, residualization, kinship adjustment, or any combination of these transformations, such as 'log-transformed-kinship-adjusted' values. Please take care that NDs that have been coded as the LOD must also be transformed appropriately.
5. Column 'phenotype' can be any numeric value, including zero and negative values.

## Getting started

The lodGWAS package provides two functions. The first one, `lod_QC`, allows the user to check the quality of the phenotype data with respect to the LODs, and the second one, `lod_GWAS`, performs a GWAS analysis by applying a parametric censored survival analysis.

### Function `lod_QC`

As mentioned above it is important that the phenotype and outsideLOD values in the phenotype file are coded correctly (see [A few considerations regarding the phenotype](#)). The function `lod_QC` checks the quality of the phenotype file, and provides a number of descriptive statistics about the phenotype, as well as warnings on suspected problems.

As a quality check, the user can specify values for the lower and upper LOD in the function `lod_QC`. It will then compare the values provided in the column 'phenotype' with those values conditional on the values in the column 'outsideLOD'. Phenotype values of measurements that have been coded as being lower than the lower LOD (outsideLOD=2) should be equal to the specified lower LOD. Similarly, phenotype values of measurements that have been coded as being larger than the upper LOD (outsideLOD=0) should be equal to the specified upper LOD. Phenotype values of measurements that have been coded as being truly measured (outsideLOD=1) should be larger than the specified lower LOD and smaller than the specified upper LOD.

Regardless whether LODs are specified, a second check will be performed to test whether the phenotype values for the NDs are in line with the phenotype values of the measured observations. That is, `lod_QC` will check if there is a gap between the smallest/largest values within the LOD (truly measured values; outsideLOD=1), and the coded values for NDs (censored values; outsideLOD=0 or 2). If the coded values for NDs are far smaller/larger than the smallest/largest measured values, `lod_QC` will give a warning. An example of such an erroneous coding of NDs is when the phenotype values below the lower LOD have been set to -9, while the measured values are >0.

The results of these quality checks by `lod_QC` will be recorded in a text file.

**Note 1:** The function `lod_QC` just checks and reports. It does not correct the phenotype or outsideLOD values.

**Note 2:** The function `lod_QC` assumes that there is a single lower and/or upper LOD. If multiple LODs are used in the file, the function produces several warnings when both NDs and real phenotypes do not match the specified limits, but these may be ignored. In that case it is up to the user to ensure the quality of the phenotype file.

### *Function call*

```
lod_QC (phenofile, pheno_name,  
        filedirectory = getwd(), outputfile = "lod_QC",  
        lower_limit = NA, upper_limit = NA)
```

### *Input arguments*

**phenofile:** either the full name of the phenotype file (including the file extension), or the name of a dataset that has already been loaded into R. For the required format of this file/dataset see section [Input Formats](#).

**pheno\_name:** the name of the column in `phenofile` containing the phenotype values.

**filedirectory:** the directory that contains the phenotype file, and where the output file will be saved. The default setting is current R working directory.

**outputfile:** the name for the output file.

**lower\_limit** and **upper\_limit:** two numeric values specifying the lower and upper LOD of the assay. Defaults are NA, in which case `lod_QC` only checks for gaps between measured values and coded phenotype values of the NDs.

### **Function lod\_GWAS**

The main function `lod_GWAS` enables the user to perform a GWAS analysis accommodating the problem of LOD. This function performs a parametric survival analysis on the phenotype of interest, which contains both measured and censored values, for each genetic variant in the genotype file with the option of including covariates in the model.

The function `lod_QC` is automatically called within `lod_GWAS`, and its quality report will be saved in a separate text file.

**Note :** GWAS analysis will not be performed: 1) on rare genetic variants (with allele frequency  $<0.001$  or  $>0.999$ ), and 2) on badly imputed genetic variants (with imputation quality score  $<0.01$ ). Those genetic variants will be included in the output file, but the association results will be “NA”.

### *Function call*

```
lod_GWAS (phenofile, pheno_name,  
          basic_model = NULL, dist = "gaussian",  
          mapfile, genofile,  
          outputfile, filedirectory = getwd(),  
          outputheader = "QCGWAS",  
          gzip_output = TRUE,  
          lower_limit = NA, upper_limit = NA)
```

### *Input arguments*

**phenofile:** either the full name of the phenotype file (including the file extension), or the name of a dataset that has already been loaded into R. For the required format of this file/dataset see section [Input Formats](#).

**pheno\_name:** the name of the column in `phenofile` containing the phenotype values.

**basic\_model:** a formula describing the basic model without the genetic variant to be fitted to the phenotype. The covariates to be included into the analysis are mentioned within quotation marks separated by plus signs; such as `basic_model="sex+age"`. Please note that covariate names should exactly match the appropriate column names of phenotype file. Default is `NULL`, in which case the association of genetic variants with the phenotype is tested without adjustment for covariates.

**dist:** assumed distribution of the (raw or transformed) phenotype. The options are `"weibull"`, `"exponential"`, `"gaussian"`, `"logistic"`, `"lognormal"` and `"loglogistic"`. Default is `"gaussian"`.

**mapfile:** the file name of the genotype map file (including the file extension). For the required format of this file see section [Input Formats](#).

**genofile:** the file name of the genotype dosage file (including the file extension). For the required format of this file see section [Input Formats](#).

**outputfile:** the name for the output file.

**filedirectory:** the directory that contains the phenotype and genotype files and where the output files will be saved. The default setting is current R working directory.

**outputheader:** the optional output format of the analysis results file, to make it compatible with different downstream software packages. The options are `"QCGWAS"`, `"GWAMA"`, `"PLINK"`, `"META"`, and `"GenABEL"`. Default is `"QCGWAS"`.

**gzip\_output:** an optional logical value that determines if the output file is compressed. Default is `TRUE`.

**lower\_limit** and **upper\_limit:** arguments passed to `lod_QC`. Defaults are `NA`. As already mentioned, GWAS analysis will be performed on the values of the phenotype file that will not be modified by `lod_QC` based on the values of these arguments. These arguments are just for the quality checks.



## Output format

Column descriptions of the output file (as per default: `outputheader="QCGWAS"`) are as the following:

**MARKER:** marker ID (identifier of the genetic variant) as specified in the genotype input files.

**CHR:** chromosome as specified in the genotype map file.

**POSITION:** physical position (base-pair position) as specified in the genotype map file.

**OTHER\_ALL:** non-effect allele (non-coded allele)

**EFFECT\_ALL:** effect allele (coded allele)

**N\_TOTAL:** total sample size, including all NDs as well as valid measured values

**N\_VALID:** the sample size of valid measured values (excluding all NDs). This is useful if the user wants to know the percentage of NDs to the total sample size.

**EFF\_ALL\_FREQ:** effect allele frequency

**EFFECT:** effect size (beta) of effect allele

**STDERR:** standard error of effect allele

**PVALUE:** p-value of association

**IMP\_QUALITY:** imputation quality of the genetic variant

If another output format is chosen, the same columns will be presented in the output file, but with header names as required by the specified software program.