METAINTER
–
Meta-analysis tool for multiple regression models in genome-wide association studies allowing for interaction

Tim Becker
Dmitriy Drichel
Christine Herold
André Lacour
Vitalia Schüller
Tatsiana Vaitsiakhovich

German Center for Neurodegenerative Diseases (DZNE), Bonn

Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn

Bonn
May 5, 2014

Contents

Chapter 1

Introduction

METAINTER is a stand-alone software written in C/C++ to perform meta-analysis of summary statistics obtained from a series of related studies. The special feature of METAINTER is the ability to meta-analyze the results of multiple linear and logistic regression models, broadly used in genome-wide association studies (GWAS).

It is assumed that a unique pre-defined model is used in multiple studies to test for association of SNP tuples with a particular phenotype. SNP coding and parameter coding in the regression models have to follow the standards specified in (Cordell and Clayton, 2002), see also Section 3.1. As input, the analysis results of the individual studies have to be provided in tabulated format. METAINTER supports the output format of the genetic interaction analysis software INTERSNP (`http://intersnp.meb.uni-bonn.de/`) as well as any freely defined format, Section 3.3.5.

The main meta-analysis method implemented in METAINTER is the method of the synthesis of regression slopes (MSRS), Section 2.4, suggested in (Becker and Wu, 2007). MSRS requires, in addition to model parameters estimates and their standard error, the availability of the covariance matrix. The covariance matrix of model parameters is provided, for instance, by INTERSNP tool (Herold *et al.*, 2009). Note that in case of tests with just one parameter, MSRS is equivalent to the standard fixed effects meta-analysis method. Within MSRS framework, METAINTER can be used to test the homogeneity of studies results, and to obtain the common parameter estimates of multiple regression models in the joint sample.

Since the covariance matrix of model parameters is not always available, METAINTER provides three further meta-analysis methods: the Fisher's method, the Stouffer's method with weights, the Stouffer's method with weights and effect directions, see Chapter 2. Thereby, METAINTER enables meta-analysis of single-marker association tests, global haplotype tests and tests for and under gene-gene interaction.

Chapter 2

Methods

There are four meta-analysis methods implemented in METAINTER:

- Fisher's method;
- Stouffer's method with weights;
- Stouffer's method with weights and effect directions;
- Method of synthesis of regression slopes (MSRS).

The first two methods are based on combining p-values of individual studies participating in meta-analysis and can be applied for summarizing the results of any association test. The Stouffer's method with weights and effect directions represents a p-value combination approach, where the consistency of effect directions across the studies is involved, and can be used to meta-analyze the results of multiple regression models. The fourth method, based on multivariate generalized least squares estimation, can be applied to synthesize the results of multiple regression models. MSRS involves model parameter estimates and their correlation and provides the overall meta-analysis p-values together with meta-analytic estimates of the regression slopes.

Assume that Study 1, ... , Study k were conducted to test a particular hypothesis H_1 versus the null H_0. Let p_j be a p-value, n_j be a sample size and w_j be a study specific weight of Study j, $j = 1, \dots, k$.

2.1 Fisher's method

The test statistic of the Fisher's method (Fisher, 1932) has the form

$$T = -2\sum_{j=1}^{k} \log p_j$$

and is χ^2-distributed under H_0 with $2k$ degrees of freedom (df).

2.2 Stouffer's method with weights

According to the Stouffer's method with weights (Stouffer *et al.*, 1949), (Lipták, 1959), a combined p-value can be found as

$$p = 1 - \Phi\left(Z\right), \quad \text{with} \quad Z = \frac{\sum\limits_{j=1}^{k} w_j \Phi^{-1}(1 - p_j)}{\sqrt{\sum\limits_{j=1}^{k} w_j^2}},$$

where Φ and Φ^{-1} denote the standard normal cumulative distribution function and its inverse, and where w_j are study specific weights, e.g. $w_j = \sqrt{n_j}$.

2.3 Stouffer's method with weights and effect directions

In case of multiple regression models the Stouffer's method with weights can be modified in order to include the information on the consistency of effect directions across the studies.

Assume that the same regression model is used in all studies, and consider two of them, Study 1 and Study 2. Let exemplarily a model equation be

$$\operatorname{logit} Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_P X_P, \tag{2.1}$$

and assume that it is tested versus $\operatorname{logit} Y = \beta_0$. To compare the effect directions between two studies, we suggest the following criterion:

Two studies are said to have the *same effect directions*, if and only if the dot product of two vectors

$$\left[\frac{\widehat{\beta}_{11}}{\operatorname{se}(\widehat{\beta}_{11})}, \ldots, \frac{\widehat{\beta}_{1P}}{\operatorname{se}(\widehat{\beta}_{1P})}\right] \quad \text{and} \quad \left[\frac{\widehat{\beta}_{21}}{\operatorname{se}(\widehat{\beta}_{21})}, \ldots, \frac{\widehat{\beta}_{2P}}{\operatorname{se}(\widehat{\beta}_{2P})}\right]$$

is positive in $\mathbb{R}^P$. Here, $\widehat{\beta}_{jl}$ are estimates of β_l with the standard error $\operatorname{se}(\widehat{\beta}_{jl})$ in Study j, $j = 1, 2$, $l = 1, \ldots, P$.

In case, when the model equation (2.1) is tested versus $\operatorname{logit} Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_S X_S$, $1 \leq S < P$, we have to include into consideration the predictor estimates $\widehat{\beta}_{jl}$ with $j = 1, 2$ and $l = S+1, \ldots, P$ only.

According to the Stouffer's method with weights and effect directions a combined p-value can be found as

$$p = 2\left(1 - \Phi\left(|Z|\right)\right), \quad \text{with} \quad Z = \frac{\sum\limits_{j=1}^{k} \delta_{j1} w_j \Phi^{-1}(1 - p_j/2)}{\sqrt{\sum\limits_{j=1}^{k} w_j^2}},$$

where $\delta_{j1} = 1$ if Study j and Study 1 have the same direction, and $\delta_{j1} = -1$, otherwise.

2.4 Method of synthesis of regression slopes

We describe the method of synthesis of regression slopes for a multiple *linear* regression model (Becker and Wu, 2007). The method works analogously for a multiple *logistic* regression model.

Assume that in k studies a multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_P X_P \tag{2.2}$$

was tested versus $Y = \beta_0$ to find the effect of P predictors on the outcome variable Y. The aim of the meta-analysis is to combine the results of k studies to obtain the overall p-value and the common estimate of the slope vector

$$\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_P]'.$$

Let $[\widehat{\beta}_{j0}, \widehat{\beta}_{j1}, \ldots, \widehat{\beta}_{jP}]'$ be an estimate of $\boldsymbol{\beta}$ in Study j, $j = 1, \ldots, k$. The slopes estimates vectors obtained in k studies are stacked to a vector

$$\mathbf{b} = [\widehat{\beta}_{10}, \widehat{\beta}_{11}, \ldots, \widehat{\beta}_{1P}, \ldots, \widehat{\beta}_{k0}, \widehat{\beta}_{k1}, \ldots, \widehat{\beta}_{kP}]'.$$

The method relies mutually on the availability of the covariance matrix

$$\boldsymbol{\Sigma}_j = \mathrm{cov}([\beta_{j0}, \beta_{j1}, \ldots, \beta_{jP}]') = \begin{bmatrix} \mathrm{var}\beta_{j0} & \mathrm{cov}(\beta_{j0}, \beta_{j1}) & ... & \mathrm{cov}(\beta_{j0}, \beta_{jP}) \\ \mathrm{cov}(\beta_{j0}, \beta_{j1}) & \mathrm{var}\beta_{j1} & ... & \mathrm{cov}(\beta_{j1}, \beta_{jP}) \\ ... & ... & ... & ... \\ \mathrm{cov}(\beta_{j0}, \beta_{jP}) & \mathrm{cov}(\beta_{j1}, \beta_{jP}) & ... & \mathrm{var}\beta_{jP} \end{bmatrix}$$

or at least its estimate $\widehat{\boldsymbol{\Sigma}}_j$ in each Study j. Combining k studies, an estimate of the joint covariance matrix $\boldsymbol{\Sigma}$ has a form of a block diagonal matrix

$$\widehat{\boldsymbol{\Sigma}} = \mathrm{diag}\left[\widehat{\boldsymbol{\Sigma}}_1, \ldots, \widehat{\boldsymbol{\Sigma}}_k\right].$$

The slopes estimates are modeled as

$$\mathbf{b} = \mathbf{W}\boldsymbol{\beta} + \mathbf{e},$$

i.e. as a function of $\boldsymbol{\beta}$ and a design matrix $\mathbf{W}$, where the covariance matrix of $\mathbf{e}$ has to satisfy the condition $\mathrm{cov}(\mathbf{e}) = \boldsymbol{\Sigma}$. The design matrix $\mathbf{W}$ is composed of zeros and ones that identify which slopes are estimated in each Study. In case, when all studies examine the same set of P predictors, a stack of k identity matrices of the dimension $(P+1)\times(P+1)$ serves as $\mathbf{W}$. The generalized least squares approach gives the following common estimates of the slope vector $\boldsymbol{\beta}$ and its covariance matrix $\mathrm{cov}(\boldsymbol{\beta})$:

$$\widehat{\boldsymbol{\beta}} = \left[\mathbf{W}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{W}\right]^{-1}\mathbf{W}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{b}, \quad \mathrm{cov}(\widehat{\boldsymbol{\beta}}) = \left[\mathbf{W}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{W}\right]^{-1}.$$

With large samples and under typical regularity conditions $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathrm{cov}(\widehat{\boldsymbol{\beta}}))$, the confidence intervals for each element of $\boldsymbol{\beta}$ are available, namely $\widehat{\beta}_i \pm Z_{1-\alpha/2}\sqrt{C_{ii}}$, where $Z_{1-\alpha/2}$ is the upper tail $1-\alpha/2$ critical value of the standard normal distribution, C_{ii} is the ith diagonal element of $\mathrm{cov}(\widehat{\boldsymbol{\beta}})$ matrix, $i = 0, \ldots, P$.

There are several tests available:

- **Test of model fit** (or a test of homogeneity of model parameters across the studies)

  $$\beta_{10} = \ldots = \beta_{k0}, \ldots, \beta_{1P} = \ldots = \beta_{kP},$$

  is given by a test statistic

  $$T = \left(\mathbf{b} - \mathbf{W}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{b} - \mathbf{W}\widehat{\boldsymbol{\beta}}\right),$$

  which is χ^2-distributed with $(k-1)(P+1)$ df.

- **Test of the composite hypothesis**

  $$\boldsymbol{\beta} = 0$$

  is given by a test statistic

  $$T = \widehat{\boldsymbol{\beta}}'\left(\mathrm{cov}(\widehat{\boldsymbol{\beta}})\right)^{-1}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'\left(\mathbf{W}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{W}\right)\widehat{\boldsymbol{\beta}},$$

  which is χ^2-distributed with $P+1$ df.

In general situation, where the model equation (2.2) is tested versus

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_S X_S \tag{2.3}$$

in each study, $1 \le S < P$, the method can be modified by including predictor slopes $\beta_{S+1}, \ldots, \beta_P$ in $\boldsymbol{\beta}$, $\mathbf{b}$, $\boldsymbol{\Sigma}$ and thus reducing the dimension of $\mathbf{W}$. Then a test of model fit

$$\beta_{1(S+1)} = \ldots = \beta_{k(S+1)}, \ldots, \beta_{1P} = \ldots = \beta_{kP}$$

will have $(k-1)(P-S)$ df, and a test of the composite hypothesis

$$\beta_{S+1} = \ldots = \beta_P = 0$$

will have $(P-S)$ df.

To perform the meta-analysis with MSRS, METAINTER requires the input of the elements of the covariance matrix $\widehat{\Sigma}_j$ from each Study $j, j = 1, \ldots, k$. We underline the corresponding elements in the matrices below.

- In case, when (2.2) is tested versus $Y = \beta_0$ in the primary analysis, the input of upper triangle elements elements of $\widehat{\Sigma}_j$ is required, i.e.

$$\widehat{\mathbf{\Sigma}}_j = \mathrm{cov}([\widehat{\beta}_{j0}, \ldots, \widehat{\beta}_{jP}]') = \begin{bmatrix} \underline{\mathrm{var}\widehat{\beta}_{j0}} & \underline{\mathrm{cov}(\widehat{\beta}_{j0}, \widehat{\beta}_{j1})} & ... & \underline{\mathrm{cov}(\widehat{\beta}_{j0}, \widehat{\beta}_{jP})} \\ \mathrm{cov}(\widehat{\beta}_{j0}, \widehat{\beta}_{j1}) & \underline{\mathrm{var}\widehat{\beta}_{j1}} & ... & \underline{\mathrm{cov}(\widehat{\beta}_{j1}, \widehat{\beta}_{jP})} \\ ... & ... & ... & ... \\ \mathrm{cov}(\widehat{\beta}_{j0}, \widehat{\beta}_{jP}) & \mathrm{cov}(\widehat{\beta}_{j1}, \widehat{\beta}_{jP}) & ... & \underline{\mathrm{var}\widehat{\beta}_{jP}} \end{bmatrix}$$

- In case, when (2.2) is tested versus (2.3) in the primary analysis, the adjusted covariance matrix is used, and the input of the following elements is required

$$\widehat{\mathbf{\Sigma}}_j = \mathrm{cov}([\widehat{\beta}_{j0}, \widehat{\beta}_{j(S+1)}, \ldots, \widehat{\beta}_{jP}]') = \begin{bmatrix} \underline{\mathrm{var}\widehat{\beta}_{j0}} & \underline{\mathrm{cov}(\widehat{\beta}_{j0}, \widehat{\beta}_{j(S+1)})} & ... & \underline{\mathrm{cov}(\widehat{\beta}_{j0}, \beta_{jP})} \\ \mathrm{cov}(\widehat{\beta}_{j0}, \widehat{\beta}_{j(S+1)}) & \underline{\mathrm{var}\widehat{\beta}_{j(S+1)}} & ... & \underline{\mathrm{cov}(\widehat{\beta}_{j(S+1)}, \widehat{\beta}_{jP})} \\ ... & ... & ... & ... \\ \mathrm{cov}(\widehat{\beta}_{j0}, \widehat{\beta}_{jP}) & \mathrm{cov}(\widehat{\beta}_{j1}, \widehat{\beta}_{jP}) & ... & \underline{\mathrm{var}\widehat{\beta}_{jP}} \end{bmatrix}$$

2.5 Meta-analysis methods: summary table

We summarize the meta-analysis methods implemented in METAINTER in the table below. Note that we refer here to the regression models (2.1), (2.2) that are tested versus $\operatorname{logit} Y = \beta_0$ or $Y = \beta_0$, respectively.

Method	Primary model	Inputs	Outputs
Fisher's method	arbitrary	p_j	p_*
Stouffer's method with weights	arbitrary	p_j, w_j	p_*
Stouffer's method with weights and effect directions	regression	$p_j, w_j, \widehat{\beta}_{ji}, \mathrm{se}(\widehat{\beta}_{ji})$	p_*
Method of synthesis of regression slopes	regression	$\widehat{\beta}_{ji}, \widehat{\mathbf{\Sigma}}_j$	$p_*, \widehat{\beta}_i^*, \mathrm{se}(\widehat{\beta}_i^*), \widehat{\mathbf{\Sigma}}^*$

p_j is a p-value of Study j;
w_j is a weight of Study j, e.g. square root of the sample size, $w_j = \sqrt{n_j}$;
$\widehat{\beta}_{ji}$ is an estimate of slope β_i in Study j;
$\mathrm{se}(\widehat{\beta}_{ji})$ is a standard error of an estimate of the slope β_i in Study j;
$\widehat{\mathbf{\Sigma}}_j$ is an estimate of the covariance matrix $\mathbf{\Sigma}_j$ in Study j;
p_* is a meta-analysis p-value;
$\widehat{\beta}_i^*$ is a meta-analysis estimate of the slope β_i;
$\mathrm{se}(\widehat{\beta}_i^*)$ is a standard error of $\widehat{\beta}_i^*$;
$\widehat{\mathbf{\Sigma}}^*$ is a meta-analysis estimate of the covariance matrix $\widehat{\mathbf{\Sigma}} = \operatorname{diag}\left[\widehat{\mathbf{\Sigma}}_1, \ldots, \widehat{\mathbf{\Sigma}}_k\right]$;
$i = 0, \ldots, P$, $j = 1, \ldots, k$.

Chapter 3

Usage

In this chapter we describe the prerequisites needed to run METAINTER. We start with SNP and model parameters coding in individual studies. Then we discuss some technical issues, the user have to be aware of during the preparation of the input files. After that we present a list of options available in METAINTER, and describe how to create a configuration file. We continue with the description of the METAINTER output files. At the end of the chapter, we give three examples.

METAINTER is written in C/C++ and can be operated from the command line.

Compilation: g++ metainter.cpp -o metainter -lm -O3.

Run: ./metainter configurationfile.txt

3.1 Within study parameter coding

Consider basic models interpreting different genetic effects of genotypes AA, Aa and aa at a single locus in genome on a given phenotype, where A is a susceptibility allele.

To model genetic effects on a quantitative trait a standard linear regression equation for an outcome variable y is used:

$$y = \beta_0 + \beta x + \beta_D x_D, \tag{3.1}$$

where x, x_D are two genetic predictor variables corresponding to the additive and the dominance effects of a SNP and coded according to the number of copies of the susceptibility allele, e.g. $2, 1, 0$. There are different ways to define values of the predictor variables (coding scheme). We work with the coding scheme (Cordell and Clayton, 2002)

$$x = \begin{cases} 1 & \text{for } AA, \\ 0 & \text{for } Aa, \\ -1 & \text{for } aa, \end{cases} \qquad x_D = \begin{cases} -0.5 & \text{for } AA, \\ 0.5 & \text{for } Aa, \\ -0.5 & \text{for } aa, \end{cases}$$

The coefficients β_0, β, $\beta_D \in \mathbb{R}$ are real numbers, β represents a magnitude of the additive effect defined as a half of the difference between two homozygote genotypic values, β_D is a magnitude of the dominance effect defined as the difference between the heterozygote genotypic value and the intercept parameter β_0, having a particular biological meaning depending on the choice of the coding scheme. All three coefficients have to be estimated.

To model genetic effects on a qualitative trait in e.g. case-control studies a logistic regression model is used. Let p denote the probability of expressing a phenotype, and let $0 < p < 1$. In the logistic regression the logarithm of odds $\log \frac{p}{1-p} =: \operatorname{logit} p$ is modeled as

$$\operatorname{logit} p = \beta_0 + \beta x + \beta_D x_D \tag{3.2}$$

with the same predictor variables as before and with the coefficients β, $\beta_D \in \mathbb{R}$ presenting the logarithm of the odds ratios (or genotype relative risks) that describe association between disease and genotypes.

The linear and logistic regression models (3.1), (3.2) for a single locus can easily be adjusted to the two-locus models. Moreover, they can be modified for modeling pair-wise statistical interaction. An extension of (3.2) to a model allowing for pair-wise interaction is represented by a logistic regression equation with interaction terms:

$$\begin{aligned} \operatorname{logit} p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1D} x_{1D} + \beta_{2D} x_{2D} + \\ + \gamma_{12} x_1 x_2 + \gamma_{1,2D} x_1 x_{2D} + \gamma_{1D,2} x_{1D} x_2 + \gamma_{1D,2D} x_{1D} x_{2D}. \end{aligned} \tag{3.3}$$

This equation models the additive x_i, and dominance effects x_{iD}, $i = 1, 2$, at two loci as well as interaction effects between them; $x_1 = 1, 0, -1$ and $x_{1D} = -0.5, 0.5, -0.5$ for the genotypes AA, Aa, aa; $x_2 = 1, 0, -1$ and $x_{2D} = -0.5, 0.5, -0.5$ for the genotypes BB, Bb, bb, respectively, where A and B are susceptibility alleles. The coefficients β_i, $\beta_{iD} \in \mathbb{R}$, represent the logarithm of the genotype relative risks at the locus i, $i = 1, 2$. The coefficients γ_{12}, $\gamma_{1,2D}$, $\gamma_{1D,2}$, $\gamma_{1D,2D} \in \mathbb{R}$ reflect the magnitude of the interaction effects. In the same manner the linear regression equation with interaction terms can be defined as a generalization of (3.1). For more details on linear and logistic regression models used in Genetic Epidemiology see (Cordell and Clayton, 2002)

3.2 SNP order, allele reference and strand issues

METAINTER currently does not allow different order in listing SNPs across studies. For example, if a SNP pair is specified as (rs1,rs2) in study A, but is given in the order (rs2,rs1) in study B, the pair will **not** be meta-analyzed. Note that INTERSNP output tuples are typically ordered by the genomic location and the issue of different orders will not occur with INTERSNP files (unless the usage of different genome builds caused flips in the genomic order). For other input file formats, the lines affected by the inconsistent ordering have to be re-edited. Note that it is **not** sufficient to flip the columns with the respective SNP names. Changing the SNPs order has an impact on the sign of parameter estimates and the entries of the covariance matrix. The sign changes depending on the parameter type (additive or dominance variation term), and re-adjustment can become tricky when more than two SNPs are involved.

METAINTER takes care of the varying allele references across the studies. Suppose that for a C/T SNP the alleles in input file are given in the order C/T for Study 1, but given in the order T/C for Study 2. In this case all parameter estimates of log-additive type that depend on the SNP in the underlying regression model in Study 2 will be multiplied by -1 to unify the reference. The procedure will be repeated for all SNPs of the tuple under consideration with diverging reference. In addition, all entries of the covariance matrix (if available) that depend on log-additive terms of the SNP will be multiplied by -1. Note that the diagonal elements of the covariance matrix depend twice on the same parameter. Therefore, the diagonal elements remain unchanged, which is self-understood as these elements of the covariance matrix are just the variances of the model parameters. For parameters of entire dominance variation type no modification is needed. Note that allele reference is automatically handled by INTERSNP.

METAINTER also attempts to solve strand flips. If a SNP is given as C/T polymorphism in Study 1, and as G/A polymorphism in Study 2, METAINTER assumes that C↔G and T↔A. If the alleles of Study 2 occur in the order A/G instead, the SNP will undergo in addition the procedure described in the previous paragraph. C/G polymorphism, of course, will not be flipped by METAINTER. For such polymorphisms, strand consistency across studies has to be established prior to analysis.

3.3 Options

Parameters and options of the meta-analysis have to be specified in a configuration file.

3.3.1 Top level keywords

METAINTER uses two top level keywords:

`GENERAL`

This **obligatory** keyword indicates that all the following lines of a configuration file, until the occurrence of the other top level keyword `NEW_STUDY`, specify either general options or options for all studies. It is obligatory to specify the keywords `METHOD` and `OUTPUT` under `GENERAL`.

`NEW_STUDY`

This **obligatory** keyword indicates that all the following lines of a configuration file, until the occurrence of the line with the keyword `NEW_STUDY`, refer to the same study. Studies will be enumerated according to the number of occurrences of `NEW_STUDY`. Under each `NEW_STUDY`, the keyword `FILE` has to be specified. The first `NEW_STUDY` must occur after the `GENERAL` keyword, and the `GENERAL` keyword cannot re-occur after a `NEW_STUDY`. Keywords specified under `NEW_STUDY` overwrite values that were set under `GENERAL`. The keywords `INTERSNP` and `INTERSNP_SINGLE` and the keywords that are set by them cannot be overwritten.

3.3.2 General keywords

`OUTPUT` <string>

This **obligatory** keyword is used to specify the path and the name of the output files. The value of <string> is a name tag. All output file names begin with this tag.

`METHOD` <string>

This **obligatory** keyword is used to specify meta-analysis method that shall be applied. The methods are coded as 1 = Fisher's method, 2 = Stouffer's method with weights, 3 = Stouffer's method with weights and effect directions, 4 = Method of synthesis of regression slopes. Several methods can be chosen in one run.

Examples:

`METHOD` 1;3; // do methods 1 and 3
`METHOD` 1-4; // do all four methods

`pFILTER` <r>

This **optional** keyword is used to set a p-value cut-off level r. METAINTER produces two output files, one with all results, another one with those results that reached a particular p-value cut-off. The default value for r is 1.0×10^{-6}.

3.3.3 INTERSNP format keywords

INTERSNP format keywords can be used when the primary analysis of all studies was performed by INTERSNP. The keywords `INTERSNP` and `INTERSNP_SINGLE` are special keywords that indicate that input files from all studies were generated with INTERSNP and therefore have the same format. The keywords `INTERSNP` or `INTERSNP_SINGLE` have to be specified under `GENERAL`. When they are used, specification of model parameters becomes redundant. Only the INTERSNP test used in the primary analysis has to be specified.

INTERSNP <n>
This **optional** keyword indicates that input files from all studies were generated with INTERSNP, and a multi-marker test was used in the primary analysis. Here, n is the test indicator of an INTERSNP two- or three-marker test. The indicators are those used with the INTERSNP keyword TEST (-> link zu Doku IS).

INTERSNP_SINGLE <n>
This **optional** keyword indicates that input files from all studies were generated with INTERSNP, and a single-marker test was used in the primary analysis. Here, n is the test indicator of an INTERSNP single-marker test. The indicators are those used with the INTERSNP keyword SINGLE_MARKER (-> link zu Doku IS).

3.3.4 Study-specific keywords

FILE <filename>
This **obligatory** keyword specifies the path and the name of the input file for a current study. It has to be used under each NEW_STUDY.

STUDYWEIGHT <r>
This **optional** keyword specifies the weight of a study. It is needed in methods 2 and 3. The keyword has to be specified under each NEW_STUDY.

For instance, square root of the sample size can be chosen as a study weight (Zaykin, 2011).

3.3.5 Keywords for free format input files

In case when "free" input file format is used, several additional keywords are obligatory. They can be specified either under GENERAL and will refer then to all studies, or under NEW_STUDY to set them for a current study. Values specified under GENERAL can be modified for a particular study by re-defining the keywords in the corresponding NEW_STUDY block. File formats are allowed to differ across studies.

HEADERLINES <n>
This **optional** keyword specifies the number of header lines in a study (all studies, when specified under GENERAL). The default is 0.

nSNPs <n>
This **obligatory for free format** keyword specifies the number of SNPs in the analysis model. It has to be specified under GENERAL.

nPARAM <n>
This **obligatory for free format** keyword specifies the number of parameters in the primary analysis model. It has to be specified under GENERAL.

PARAMREFERENCE <string>
This keyword is **obligatory for free format**, when methods 3 or 4 are selected. For each model parameter, PARAMREFERENCE indicates, which SNP this parameter refers to. The keyword has to be specified under GENERAL.

We explain how to use the last two keywords by example. Suppose that the full genotype model defined by 2 SNPs (`nSNPs` 2) is used in case-control studies. In this case a logistic regression model

$$\operatorname{logit} p = \beta_0+\beta_1 x_1+\beta_{1D}x_{1D}+\beta_2 x_2+\beta_{2D}x_{2D}+\gamma_{12}x_1x_2+\gamma_{1,2D}x_1x_{2D}+\gamma_{1D,2}x_{1D}x_2+\gamma_{1D,2D}x_{1D}x_{2D}$$

is tested versus

$$\operatorname{logit} p = \beta_0,$$

or, equivalently, the null hypothesis

$$H_0 :\ \beta_1 = \beta_{1D} = \beta_2 = \beta_{2D} = \gamma_{12} = \gamma_{1,2D} = \gamma_{1D,2} = \gamma_{1D,2D} = 0$$

has to be verified. Here `nPARAM` 8 has to be set. For `PARAMREFERENCE` the following specification is needed:

`PARAMREFERENCE` 1;1;2;2;1+2;1+2;1+2;1+2;

For each parameter, the `PARAMREFERENCE` string clarifies, which SNP (indicated by its number) the parameter depends on. The first parameter is x_1, it depends only on SNP 1 ("1;"). The second parameter is x_{1D}, it also depends only on SNP 1. Then it comes to parameters that depend on SNP 2 ("2;") only. Parameters 5 to 8 are interaction terms. They depend on both SNPs, hence we use "1+2;" for them. Models with more SNPs can be defined analogously.

`PARAMTYPE` <string> This keyword is **obligatory for free format**, when methods 3 or 4 are selected. For each parameter, `PARAMTYPE` indicates the type of the parameter (A for (log-) additive or D for dominance variation). The keyword has to be specified under `GENERAL`.

We again explain the keyword by the 2-SNP full genotype 8 df model example. The proper usage of the keyword in this case is

`PARAMTYPE` A;D;A;D;A+A;A+D;D+A;D+D;

Parameters 1 and 3 depend on one SNP and are log-additive parameters, hence we use "A;". Parameters 2 and 4 depend on one SNP and are dominance variation parameters, hence we use "D;". Parameter 5 depends on two SNPs and corresponds to the interaction term x_1x_2, where the first and the second component are log-additive ("A+A;"). Analogously, for x_1x_{2D} term we have "A+D;", for $x_{1D}x_2$ we use "D+A;", and for $x_{1D}x_{2D}$ we write "D+D;".

In a 5-SNP model including a parameter for the interaction term $x_1x_3x_{5D}$, we would specify `PARAMREFERENCE` "1+3+5;" and `PARAMTYPE` "A+A+D;" for this parameter.

`pCOL` <n> This keyword is **obligatory for free format**, when methods 1, 2 or 3 are selected. The keyword indicates the column with p-values in each study, and can be specified both under `GENERAL` and `NEW_STUDY`.

`SNPCOLS` <string> This **obligatory for free format** keyword indicates the columns with SNPs IDs. The keyword can be specified both under GENERAL and `NEW_STUDY`.
Example:
`SNPCOLS` 3;7; // Columns 3 and 7 contain SNPs IDs

`CHRCOLS` <string> This **optional** keyword indicates the columns with SNPs chromosomes. The keyword can be specified both under `GENERAL` and `NEW_STUDY`.

Example:
CHRCOLS 2;6; // Columns 2 and 6 contain chromosomes of the SNPs (the same SNPs order as in SNPCOLS is assumed)

POSCOLS <string>

This **optional** keyword indicates the columns with SNPs positions (bp). The keyword can be specified both under GENERAL and NEW_STUDY.
Example:
POSCOLS 4;8; // Columns 4 and 8 contain position in bp of the SNPs (the same order as in SNPCOLS is assumed)

ALLELECOLS <string>

This keyword is **obligatory for free format**, when methods 3 or 4 are selected, and indicates the columns with SNPs alleles. The keyword can be specified both under GENERAL and NEW_STUDY. Note that the number of alleles that have to be specified is twice the number of SNPs.
Example:
ALLELECOLS 12;13;14;15; // Columns 12 to 15 contain SNPs alleles.

Remark: Two alleles of the first SNP are given in column 12 and 13, those of the second SNP are given in columns 14 and 15. The sign of the parameter estimates refers to the alleles in columns 12 (SNP 1) and columns 14 (SNP 2). In other words, it is assumed that the alleles in columns 12, 14 were coded as "1" and that the alleles in columns 13, 15 were coded as "-1" in the regression analysis. This rule coincides with that of PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/), in SNPTEST (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) the coding is the other way round (!).

BETACOLS <string>

This keyword is **obligatory for free format**, when methods 3 or 4 are selected, and indicates the columns with parameter (beta) estimates. The keyword can be specified both under GENERAL and NEW_STUDY. One column for each parameter (as defined by nPARAM) is needed.

SECOLS <string>

This keyword is **obligatory for free format**, when methods 3 or 4 are selected, and indicates the columns with standard errors. The keyword can be specified both under GENERAL and NEW_STUDY. One column for each parameter (as defined by nPARAM) is needed.

COVCOLS <string>

This keyword is **obligatory for free format**, when method 4 is selected, and indicates the columns with the entries of the covariance matrix sigma. More precisely, only columns for the entries of the upper triangle (including the diagonal) of the covariance matrix have to be indicated. The keyword can be specified both under GENERAL and NEW_STUDY. The number of the required columns is (nPARAM+2)*(nPARAM+1)/2, see Section 2.4.
Example:
COVCOLS 32-76; // columns 32 to 76 contain the entries of the upper triangle of the covariance matrix (model with nPARAM=8)

3.3.6 METAINTER keywords table

An overview of all METAINTER keywords described above is given in the following table:

Top level keywords	Status	Description
GENERAL	obligatory	To specify the general options
NEW_STUDY	obligatory	To specify the options for a current study
General keywords		
OUTPUT <string>	obligatory	To specify the path and the name of the output files
METHOD <string>	obligatory	To specify the method(s) of the meta-analysis to be applied 1=Fisher's method; 2=Stouffer's method with weights; 3=Stouffer's method with weights and effect directions; 4=Method of synthesis of regression slopes
pFILTER <r>	optional	To specify the p-value cut-off. By default r$= 1.0 \times 10^{-6}$
INTERSNP format keywords		
INTERSNP <n>	optional	To indicate that input files from all studies were generated with INTERSNP and that a two- or three-marker test was used; n is the argument of the keyword TEST in INTERSNP
INTERSNP_SINGLE <n>	optional	To indicate that input files from all studies were generated with INTERSNP and that a single-marker test was used; n is the argument of the keyword SINGLE_MARKER in INTERSNP
Study-specific keywords		
FILE <filename>	obligatory	To specify the path and the filename of the input file for a current study
STUDYWEIGHT <r>	optional	To specify the weight for a current study
Arbitrary input files keywords		
HEADERLINES <n>	optional	To indicate the number of header lines of a project
nSNPS <n>	obligatory	To indicate the number of SNPs in the initial analysis model Has to be specified under GENERAL
nPARAM <n>	obligatory	To indicate the number of parameters in the primary analysis model. Has to be specified under GENERAL
PARAMREFRERENCE <string>	obligatory*	To indicate for each parameter, which SNP it depends on Has to be specified under GENERAL
PARAMTYPE <string>	obligatory*	To indicate for each parameter, wether it is additive or dominance variance parameter Has to be specified under GENERAL
pCOL <n>	obligatory**	To indicate the column with p-values Can be specified both under GENERAL and NEW_STUDY
SNPCOLS <string>	obligatory	To indicate the columns with SNP names

		Can be specified both under GENERAL and NEW_STUDY
CHRCOLS <string>	optional	To indicate the columns with SNP chromosomes Can be specified both under GENERAL and NEW_STUDY
POSCOLS <string>	optional	To indicate the columns with SNP positions Can be specified both under GENERAL and NEW_STUDY
ALLELECOLS <string>	obligatory*	To indicate the columns with SNP alleles Can be specified both under GENERAL and NEW_STUDY
BETACOLS <string>	obligatory*	To indicate the columns with parameter (beta) estimates Can be specified both under GENERAL and NEW_STUDY
SECOLS <string>	obligatory*	To indicate the columns with standard errors Can be specified both under GENERAL and NEW_STUDY
COVCOLS	obligatory***	To indicate the columns with the entries of the upper triangle of the covariance matrix Can be specified both under GENERAL and NEW_STUDY

*obligatory for METHOD 3, 4

**obligatory for METHOD 1, 2, 3

***obligatory for METHOD 4

3.4 METAINTER output files

We assume that "test" has been specified as output name tag. The corresponding line in the configuration file is

`OUTPUT` test;

METAINTER creates the following output files:

- test_Log.txt

  The log file re-states the selected keywords and provides some basic summary statistics. These should be self-explanatory.

- test_Result.txt

  The main output file is tab-separated and contains all results. The majority of columns headings should be self-explanatory. The column "minimalPlausibility" indicates, wether the meta-analysis p-value is smaller than the smallest p-value observed in any of the studies. The column "consistency" shows if the regression slopes of Study 1 (more precisely, the first non-missing study) and Study j have the same direction in the `nPARAM`-dimensional space, see Section 2.3. Studies with missing values are indicated by "x". The column "consistency" contains meaningful values only when `METHOD` 3 is selected.

- test_TopResult.txt

  This file has the same format as the main output file test_Result.txt, but contains only those lines for which at least one meta-analysis method has a p-value below the cut-off `pFILTER` (by default, 1.0×10^{-6}).

- test_NoMeta.txt

  This file lists the SNP tuples for which no meta-analysis was conducted. Possible reasons are:

  a) The tuple was found in one study only;

  b) Allele codes were inconsistent across studies;

  c) The p-value was missing or had invalid value in some studies.

The main output file can contain tuples with no valid meta-analysis. This can happen, for instance, in case of missing or invalid standard errors (value < 0), then methods 3 and 4 cannot be conducted.

3.5 Example of a configuration file. Input files generated by INTERSNP

Example. Consider a project, where:

- The primary analysis was performed with INTERSNP, `TEST` 4. This is genotypic test under interaction with two marginal effects. In case-control GWAS, `TEST` 4 is given by logistic regression model, where

$$\begin{aligned}\operatorname{logit} p = \beta_0 + \beta_1 x_1 + \beta_{1D} x_{1D} + \beta_2 x_2 + \beta_{2D} x_{2D} \\ + \gamma_{1,2} x_1 x_2 + \gamma_{1,2D} x_1 x_{2D} + \gamma_{1D,2} x_{1D} x_2 + \gamma_{1D,2D} x_{1D} x_{2D}\end{aligned}$$

  is tested versus $\operatorname{logit} p = \beta_0$, see Section 3.1;

- Three studies participate in meta-analysis; assume 3000 subjects in Study 1, 2000 subjects in Study 2 and 1000 subjects in Study 3 (to define weights);

- Meta-analysis has to be performed by all methods available in METAINTER.

Let the output files with the results from the INTERSNP run be titled as
Study1_IS.txt,
Study2_IS.txt,
Study3_IS.txt
for Study 1 to 3, respectively. A configuration file to perform the meta-analysis with METAINTER in this example has to be organized as follows, see also http://metainter.meb.uni-bonn.de/Howto.html:

Keyword	Parameter	Comment
`GENERAL`		// general options valid for all studies
`INTERSNP`	4	// to indicate that input files from all studies were generated with INTERSNP, two-marker TEST 4
`METHOD`	1; 2; 3; 4;	// to specify the method(s) of meta-analysis to be applied: 1=Fisher's method, 2=Stouffer's method with weights, 3=Stouffer's method with weights and effect directions, 4=Method of synthesis of regression slopes
`pFILTER`	0.0001	// p-value cut-off
`OUTPUT`	MA_IS	// the path and the name of the output files
`NEW_STUDY`		// options for a current study
`FILE`	Study1_IS.txt	// the path and the name of the input file for a current study
`STUDYWEIGHT`	55	// weight for a current study, here 55 is appr. square root of the sample size 3000
`NEW_STUDY`		
`FILE`	Study2_IS.txt	
`STUDYWEIGHT`	45	
`NEW_STUDY`		
`FILE`	Study3_IS.txt	
`STUDYWEIGHT`	32	

3.6 Examples of configuration files. Free format input files

In this section two examples of configuration files for free format input files are presented. In the first example, it is assumed that the results of the primary analysis are organized in the same manner in **all** studies, i.e. all input files have the same structure. The amount of columns, the order, in which they appear, etc. have to be consistent in all studies. The second example refers to the case, when the results of the primary analysis organized differently in different studies.

Example 1. Consider a project, where:

- The primary analysis was performed by logistic regression model to test genotypic interaction of two SNPs in case-control GWAS. The model equation

$$\begin{aligned}\operatorname{logit} p = {} & \beta_0 + \beta_1 x_1 + \beta_{1D} x_{1D} + \beta_2 x_2 + \beta_{2D} x_{2D} \\ & + \gamma_{1,2} x_1 x_2 + \gamma_{1,2D} x_1 x_{2D} + \gamma_{1D,2} x_{1D} x_2 + \gamma_{1D,2D} x_{1D} x_{2D}\end{aligned}$$

  was tested versus

$$\operatorname{logit} p = \beta_0 + \beta_1 x_1 + \beta_{1D} x_{1D} + \beta_2 x_2 + \beta_{2D} x_{2D}.$$

- Three studies participate in meta-analysis; assume 3000 subjects in Study 1, 2000 subjects in Study 2 and 1000 subjects in Study 3 (to define weights);
- **The results of the primary analysis are given in tabulated form, the same in all studies;**
- Meta-analysis has to be performed by all methods available in METAINTER.

Let the input files with the results of the primary analysis are titled as
Study1_FF1.txt,
Study2_FF1.txt,
Study3_FF1.txt.
A configuration file to perform the meta-analysis with METAINTER in this example has to be organized as follows, see also http://metainter.meb.uni-bonn.de/Howto.html:

Keyword	Parameter	Comment
GENERAL		// general options valid for all studies
OUTPUT	MA_FF1	// the path and the name of the output files
METHOD	1;2;3;4;	// to specify the method(s) of meta-analysis to be applied: 1=Fisher's method, 2=Stouffer's method with weights, 3=Stouffer's method with weights and effect directions, 4=method of synthesis of regression slopes
pFILTER	0.0001	// p-value cut-off
HEADERLINES	1	// number of header lines of a project
nSNPS	2	// number of SNPs in the primary analysis model
nPARAM	4	// number of parameters in the primary analysis model
PARAMREFERENCE	1+2;1+2;1+2;1+2;	// to indicate for each parameter, which SNP it depends on
PARAMTYPE	A+A;A+D;D+A;D+D	// to indicate for each parameter, wether it is additive or dominance variance parameter
pCOL	10	// column with p-values
SNPCOLS	3;7;	// columns with SNP names
CHRCOLS	2;6;	// columns with SNP chromosomes
POSCOLS	4;8;	// columns with SNP positions
ALLELECOLS	11-14;	// columns with SNP alleles

BETACOLS	15;17;19;21;	// columns with parameter (beta) estimates
SECOLS	16;18;20;22;	// columns with standard error
COVCOLS	23-37;	// columns with the entries of the upper triangle of the covariance matrix
NEW_STUDY		// options for a current study
FILE	Study1_FF1.txt	// the path and the name of the input file for a current study
STUDYWEIGHT	55	// weight for a current study, here 55 is appr. square root of the sample size 3000
NEW_STUDY		
FILE	Study2_FF1.txt	
STUDYWEIGHT	45	
NEW_STUDY		
FILE	Study3_FF1.txt	
STUDYWEIGHT	32	

Example 2. Consider a project, where:

- The primary analysis was performed by logistic regression model to test genotypic interaction of two SNPs in case-control GWAS, see Example 1.
- Three studies participate in meta-analysis; assume 3000 subjects in Study 1, 2000 subjects in Study 2 and 1000 subjects in Study 3 (to define weights);
- **The results of the primary analysis are given in tabulated form, which varies from study to study;**
- Meta-analysis has to be performed by all methods available in METAINTER.

Let the input files with the results of the primary analysis are titled as
Study1_FF2.txt,
Study2_FF2.txt,
Study3_FF2.txt.
A configuration file to perform the meta-analysis with METAINTER in this example has to be organized as follows, see also `http://metainter.meb.uni-bonn.de/Howto.html`:

Keyword	Parameter	Comment
GENERAL		// general options valid for all studies
OUTPUT	MA_FF2	// the path and the name of the output files
METHOD	1;2;3;4;	// to specify the method(s) of meta-analysis to be applied: 1=Fisher's method, 2=Stouffer's method with weights, 3=Stouffer's method with weights and effect directions, 4=method of synthesis of regression slopes
pFILTER	0.0001	// p-value cut-off
nSNPS	2	// number of SNPs in the primary analysis model
nPARAM	4	// number of parameters in the primary analysis model

```
PARAMREFERENCE  1+2;1+2;1+2;1+2;  // to indicate for each parameter, which SNP it depends on
PARAMTYPE       A+A;A+D;D+A;D+D   // to indicate for each parameter, wether it is additive or dominance variance parameter

NEW_STUDY                         // options for a current study

FILE            Study1_FF2.txt    // the path and the name of the input file for a current study
HEADERLINES     1                 // number of header lines of a project
pCOL            10                // column with p-values
SNPCOLS         3;7;              // columns with SNP names
CHRCOLS         2;6;              // columns with SNP chromosomes
POSCOLS         4;8;              // columns with SNP positions
ALLELECOLS      11-14;            // columns with SNP alleles
BETACOLS        15;17;19;21;      // columns with parameter (beta) estimates
SECOLS          16;18;20;22;      // columns with standard error
COVCOLS         23-37;            // columns with the entries of the upper triangle of the covariance matrix
STUDYWEIGHT     55                // weight for a current study, here 55 is appr. square root of the sample size 3000

NEW_STUDY

FILE            Study2_FF2.txt
HEADERLINES     1
pCOL            11
SNPCOLS         1;2;
CHRCOLS         3;4;
POSCOLS         5;6;
ALLELECOLS      7-10;
BETACOLS        12-15;
SECOLS          16-19;
COVCOLS         20-34;
STUDYWEIGHT     2000

NEW_STUDY
FILE            Study3_FF2.txt
HEADERLINES     0
pCOL            11
SNPCOLS         3;8;
CHRCOLS         1;6;
POSCOLS         2;7;
ALLELECOLS      4;5;9;10;
BETACOLS        12;14;16;18;
SECOLS          13;15;17;19;
COVCOLS         20-34;
STUDYWEIGHT     1000
```

Bibliography

Becker, B.J., Wu, M.-J. (2007) The synthesis of regression slopes in meta-analysis. *Stat. Sci.* **22**, 414-429.

Cordell, H.J., Clayton, D.G. (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case-control or family data: application to HLA in type 1 diabetes. *Am. J. Hum. Genet.* **70**, 124–141.

Fisher, R. (1932) *Statistical methods for research workers.* Oliver and Boyd, Edinburgh.

Herold, C., Steffens, M. et al. (2009) INTERSNP: genome-wide interaction analysis guided by a priori information, *Bioinformatics* **25**, 3275-3281, doi: 10.1093/bioinformatics/btp596.

Herold, C., Mattheisen, M. et al. (2012) Integrated genome-wide pathway association analysis with INTERSNP. *Hum. Hered.* **73**, 63-72, doi: 10.1159/000336196.

Lipták, T. (1959) On the combination of independent tests, *Publ. Math. Inst. Hungar. Acad. Sci.* **3**, 171-197.

Stouffer, S., DeVinney, L. et al. (1949) *The American soldier: Adjustment during army life*. Vol. 1. Princeton University Press, Princeton, US.

Zaykin, D.V. (2011) Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis, *J. Evol. Biol.* **24**, 1836-1841, doi: 10.1111/j.1420-9101.2011.02297.x.