INTERSNP
–
Genome-wide Interaction Analysis

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

Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE)
German Center for Neurodegenerative Diseases
Bonn

Institute for Medical Biometrics, Informatics and Epidemiology (IMBIE)
University of Bonn

Bonn, July 2, 2015

Contents

1 Introduction **2**

2 Using INTERSNP **3**
2.1 Usage . . . 3
2.1.1 Command line . . . 3
2.1.2 Selection file . . . 3
2.2 Input files . . . 4
2.2.1 SNP/Pedigree/Genotype files . . . 4
2.2.2 Annotation file . . . 5
2.2.3 Pathway file . . . 5
2.2.4 Covariate file . . . 6
2.2.5 Model file . . . 6
2.2.6 Combi file . . . 7
2.2.7 SNP file . . . 7
2.3 Options . . . 8
2.4 Output files . . . 14
2.4.1 Output single marker . . . 14
2.5 Produce SNP-Proxy-Files for Meta-Analysis with YAMAS . . . 16

3 Examples **18**
3.1 Genome-wide Interaction Analysis with Pre-test . . . 18
3.2 Strategy I . . . 18
3.3 Strategy II . . . 18
3.4 Strategy III . . . 19
3.5 Strategy IV . . . 19
3.6 Strategy V . . . 19
3.7 Genome-wide Haplotype Analysis . . . 20

4 Tests in INTERSNP **21**
4.1 Single-marker tests . . . 21
4.2 Two- and three-marker analysis . . . 21
4.2.1 χ^2-test for full genotype contingency table (`TEST 1`) . . . 21
4.2.2 Test for interaction using a log-linear model (`TEST 2`) . . . 21
4.2.3 Logistic regression (`TESTS 3-12`) . . . 22
4.2.3.1 Parameter coding . . . 22
4.2.3.2 Likelihoods . . . 22
4.2.4 Linear regression . . . 23
4.2.4.1 Tests (2 SNPs) . . . 23
4.2.4.2 Tests (3 SNPs) . . . 23
4.2.4.3 X-chromosome . . . 24
4.2.4.4 Covariates . . . 24
4.2.4.5 User-defined regression models . . . 24

4.2.5 Pre-tests . . . 24
4.2.5.1 Pre-test allelic interaction (logistic regression) . . . 24
4.2.5.2 Pre-test genotypic interaction (logistic regression) . . . 24
4.2.5.3 Pre-test genotypic interaction (linear regression) . . . 25
4.2.5.4 Pre-test allelic interaction (linear regression) . . . 25

5 Case-only analysis **26**
5.1 Case-only allelic interaction test (`TEST 15`) . . . 26
5.2 Case-only SNP-SNP interaction test (`TEST 16`) . . . 27
5.3 Selection file for case-only analysis . . . 27

6 Pathway association analysis **28**
6.1 Pathway scores . . . 29
6.2 Simulation . . . 29
6.3 Selection file PAA . . . 30

7 Test for Supra-Multiplicativity of SNP Sets **31**
7.1 Selection file and output . . . 31

8 Stratified analysis **32**

9 INTERSNP-RARE: Genome-Wide Rare Variant Analysis **34**
9.1 Introduction . . . 34
9.2 Options . . . 36
9.3 Statistical Tests . . . 38
9.3.1 FR . . . 38
9.3.2 CMAT . . . 38
9.3.2.1 Weights . . . 39
9.3.3 COLL . . . 39
9.3.3.1 Odds Ratios . . . 40
9.3.4 REG . . . 40
9.3.5 FRACREG . . . 40
9.3.6 COLLREG . . . 40
9.4 Variable Threshold . . . 41
9.5 Rare Pretests . . . 41
9.6 Binning Methods . . . 42
9.6.1 Count-based Binning . . . 42
9.6.2 User-Supplied Intervals . . . 42
9.6.3 Interval Modification . . . 43
9.6.3.1 Merging of Input Intervals . . . 43
9.6.3.2 Flanking . . . 43
9.6.3.3 Close Gaps . . . 43
9.6.3.4 Concatenating Intervals . . . 43
9.6.4 Size of Bins: Quality Control . . . 43
9.6.4.1 Filter Small Bins . . . 43
9.6.4.2 Maximum Number of Rare Variants . . . 43
9.7 Imputation of Minor Allele Frequencies for Missing Genotypes . . . 43
9.8 Output . . . 44
9.8.1 Association Testing Results: *Rare.txt . . . 44
9.8.2 Modified Interval File: *_modified.txt . . . 44

Bibliography **45**

Chapter 1

Introduction

INTERSNP (Herold et al., 2009) is a software for GWAS multi-marker analysis, especially interaction analysis (GWIA), of case-control SNP data and quantitative traits. SNPs are selected for joint analysis using a priori information. Sources of information to define meaningful strategies can be statistical evidence (single marker association at a moderate level, computed on the fly from the data) and genetic/biologic relevance (genomic location, function class or pathway information). Our software product implements a logistic regression framework as well as log-linear models for joint analysis of multiple SNPs, automatic handling of SNP annotation and pathway information and methods to account for multiple testing, in particular, Monte-Carlo simulations (MC-simulations) to judge genome-wide significance. Additionally, different pathway association tests are provided and it is also possible to perform a genome-wide haplotype analysis (GWHA).

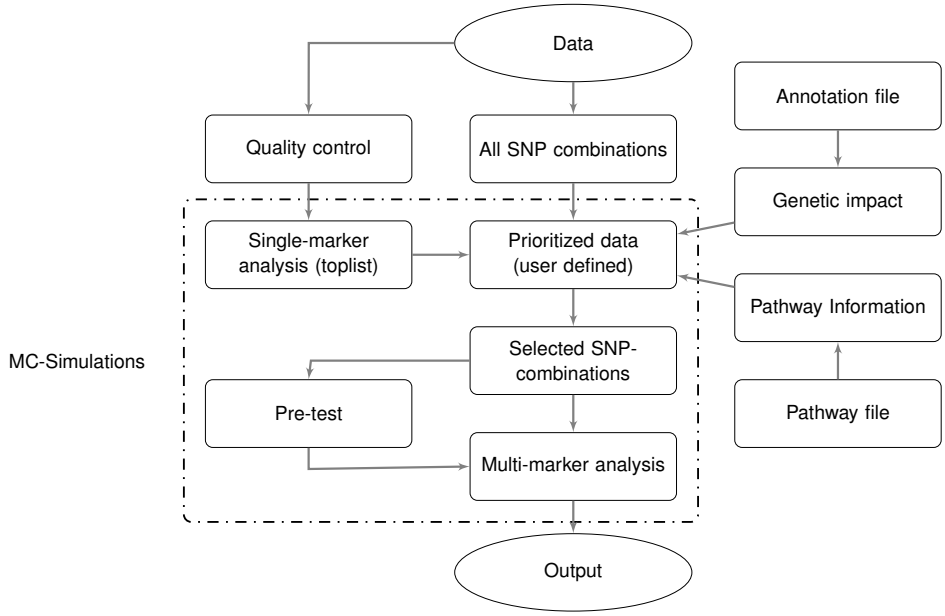

Figure 1.1: Program flow

NOTE: Single-marker analysis is always carried out by default.

Chapter 2

Using INTERSNP

2.1 Usage

2.1.1 Command line

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

Compile: g++ intersnp.cpp -o intersnp -O3.

Compile parallel version: g++ intersnp.cpp -o intersnpParallel -O3 -fopenmp.

There are two main loops which can be parallelized in INTERSNP. To compile a parallel version of INTERSNP it's necessary to modify the source code and to change either "`#define PARALLELN 0`" or "`#define PARALLELA 0`". In order to parallelize the n-loop used for MC-simulations, in particular pathway association analysis, change "`#define PARALLELN 0`" to "`#define PARALLELN 1`". To do a full Genome-wide interaction analysis (GWIA) it's recommended to parallelize the A-loop, change "`#define PARALLELA 0`" to "`#define PARALLELA 1`".

Run: ./intersnp selectionfile.txt
Run parallel version: ./intersnpParallel selectionfile.txt

2.1.2 Selection file

Various analysis options (statistical test, priorities, ...) can be specified by the user in a selection file. The file consists of three columns: keyword, parameter, and optionally comments, and is separated by whitespaces (tab, blank). Further columns are ignored.

To get familiar with the selection file, first of all a short example is introduced which will be described in detail.

```
Keyword Parameter Comment
TPED    ./test.tped  // path to the .tped file
TFAM    ./test.tfam  // path to the .tfam file
ANNOTATIONFILE  ./annotation_file.txt  // path to the annotation file
HWE_P_CASE  0.001  // QC-threshold for HWE in cases
HWE_P_CONTROL  0.01  // QC-threshold for HWE in controls
MRDIFF  1  // QC-threshold missing rate (mr): individuals and SNPs
              worse than the average mr + MRDIFF will be deleted
MAF        // QC-threshold minor allele frequency
SINGLE_MARKER  1  // Armitage's trend test
TWO_MARKER  1  // two-marker analysis
```

```
TEST  5  // logistic regression model:
            test for additive interaction
SINGLETOP  50000 // top 50000 single-marker results
M_WITH_SINGLETOP  2  // all pairs of SNPs which are among
                        the top 50000 single-marker results
GENETIC_IMPACT  3  // genetic impact: within coding region
M_WITH_GENETIC_IMPACT  2  // all pairs of SNPs which lie in a
                            coding region of a gene
OUTPUTNAME      ./myproject/test  // specify the output name
END
```

TPED is the path to the .tped input file and TFAM is the path to the .tfam file. In this example we also use an annotation file (ANNOTATIONFILE) to get the genetic information. After the input files we define some quality control (QC) criteria to delete individuals and SNPs not satisfying the QC-criteria of HWE_P_CASE and HWE_P_CONTROL, MRDIFF (missing rate) and MAF (minor allele frequency). Every analysis starts with a single-marker test which can be specified by SINGLEMARKER. In our example we use Armitage's trend test (SINGLEMARKER 1). Furthermore, we would like to do a two-marker analysis, so we select TWO_MARKER 1 and choose TEST 5, which is a test for additive interaction (see section Model table for case-control data4.1). To reduce the number of tests we use the statistic and genetic criteria. SINGLETOP 50000 and M_WITH_SINGLETOP 2 means that we analyze only pairs of SNPs which are among the top 50000 single-marker results. In addition, these pairs should also be in a coding region of a gene (GENETIC_IMPACT 2, M_WITH_GENETIC_IMPACT 3). Finally, we specify the output name (OUTPUTNAME).

2.2 Input files

2.2.1 SNP/Pedigree/Genotype files

Possible data input formats are the PLINK (Purcell et al., 2007) formats (.ped/.map; .tped/.tfam; .bed/.bim/.fam).

For instance the .tped file (TPED) contains SNP and genotype information, where each row represents a SNP and two columns represent genotypes per individual. The first four columns are chromosome, marker ID, genetic distance and base pair position.

```
16 rs8466895 0 37354 T T T T T T C C C T
16 rs216590 0 41263 G A G A A A A A G A
16 rs216596 0 45320 A G G G G G G G A G
16 rs2541594 0 45444 C T T T C T C C C C
16 rs8466 0 49427 T T T T T T C C C T
16 rs216590 0 52259 G A G A A A A A G A
...
```

NOTE: To improve running time it it recommended to sort the file by chromosome and position.
The .tfam file (TFAM) contains family information, each row represents a person. To create a .tfam file use the first six columns of a .ped file: Family ID, Individual ID, Father, Mother, Sex and Affection status.

```
co1 co1 0 0 1 1
co2 co2 0 0 2 1
...
ca1 ca1 0 0 2 2
ca2 ca2 0 0 2 2
...
```

For more details on the data formats visit the PLINK website http://pngu.mgh.harvard.edu/ purcell/-plink/data.shtml.

2.2.2 Annotation file

To work with genetic criteria a SNP annotation file (`ANNOTATIONFILE`) is needed. We use the same format as in the Illumina Human610 chip annotation file. The first columns represent rs number, chromosome, base pair position and genome_build number; all the other columns show detailed information. For our different categories the following columns are important:

- 1st column: rs_number;
- 2nd: chromosome;
- 5th column: gene_id;
- 8th column: gene_location: coding, intron, 3UTR, 5UTR, UTR, flanking_3UTR, flanking_5UTR;
- 9th column: location_relative_to_gene: numbers below zero: distance to the nearest gene, location within a gene: [461/456], since the exact position is not used you may write [1/1] if the SNP is in the gene, but not in the intron(!);
- 10th column: SNP coding_status: -1, NULL, SYNON, NONSYNON.

The file can be downloaded from our homepage (http://intersnp.meb.uni-bonn.de/ Downloads).

```
name;chr;coordinate;genome_build;gene_symbol;gene_id;accession;location;
location_relative_to_gene;coding_status;amino_acid_change;id_with_mouse;
phast_conservation
rs12354060;1;10004;36.2;LOC653635;653635;XR_017611.1;intron;-1762;
NULL;NULL;NULL;NULL
rs2691310;1;46844;36.2;LOC642894;642894;XR_016145.1;flanking_5UTR;-672;
NULL;NULL;NULL;NULL
rs2531266;1;59415;36.2;OR4F5;79501;NM_001005484.1;coding;[461/456];
SYNON;A154A (NP_001005484.1);0.64;0.979
rs4124251;1;97215;36.2;LOC727901;727901;XR_015157.1;3UTR;[3300/491];
NULL;NULL;NULL;NULL
rs8179466;1;224176;36.2;LOC728481;728481;XR_015292.1;intron;-42;NULL;
NULL;NULL;NULL
...
```

NOTE: `GENETIC_IMPACT` and `M_WITH_GENETIC_IMPACT` have to be selected.
NOTE: To improve running time it is recommended to sort the file by chromosome and position.

2.2.3 Pathway file

To include pathway information a pathway file (`PATHWAYFILE`) is needed. A file containing KEGG (Kanehisa, et al.), (O'Dushlaine et al., 2009) pathways[1] can be downloaded from our homepage (Downloads). We use the KEGG_2_SNP_229.txt file with genes annotated by us. The columns present pathway name, rs number and gene.

```
Pathway rsNo     Gene
hsa00010 rs61487361 HK2
hsa00010 rs2286168  ALDH3B1
hsa00010 rs41275697 ADH1B
hsa00010 rs191970 PGM1
hsa00010 rs7583259 GALM
...
```

[1] We wish to thank Colm O'Dushlaine (codushlaine@gmail.com) for providing the KEGG pathway file

NOTE: PATHWAY has to be selected (1).
NOTE: To improve running time it is recommended to sort the file by pathway name.

2.2.4 Covariate file

To include own covariates into the analysis use the covariate file (COVARIATEFILE). The first two columns represent Family ID and Person ID and then up to 10 covariates are allowed.

```
FID PID COV1 COV2
co1 co1 1.31 5.75
co2 co2 7.24 5.97
...
ca1 ca1 1.85 5.12
ca2 ca2 2.36 6.42
...
```

NOTE: Covariates are selected with the COVARIATES keyword. If all covariates shall be used, write 1-10; (semicolon is important)
NOTE: Missing data value is "–" or "x".

2.2.5 Model file

For user-defined logistic/linear regression models that are not included in the list of tests, the model file (MODELFILE) is needed. The first column presents the name of the parameter followed by 0/1 indicators for the first and the second likelihood. The indicators define which of the compared likelihoods L1 and L2 are used.

```
PARAMETER L1  L2
x1 1 0
x1D 0 0
x2  0 0
x2D 0 0
x1x2  0 0
x1x2D 0 0
x1Dx2 0 0
x1Dx2D  0 0
x3  0 0
x3D 0 0
x1x3  0 0
x1x3D 0 0
x1Dx3 0 0
x1Dx3D  0 0
x2x3  0 0
x2x3D 0 0
x2Dx3 0 0
x2Dx3D  0 0
x1x2x3  0 0
x1x2x3D 0 0
x1x2Dx3 0 0
x1x2Dx3D  0 0
x1Dx2x3 0 0
x1Dx2x3D  0 0
x1Dx2Dx3  0 0
x1Dx2Dx3D 0 0
```

NOTE: TEST has to be set to M.

2.2.6 Combi file

To analyse only user-specified SNP pairs or triples a combi file (`COMBILIST`) is required. Each row represents one combination.

```
rs11248850 rs7190878
rs7404049 rs4984707
rs11649498 rs1040499
...
```

NOTE: `COMBILIST` has to be set to 1.

2.2.7 SNP file

To analyse all combinations of user-specified SNPs from the .tped file a SNP file (`SNPFILE`) is required. Each row represents one SNP.

```
rs11248850
rs7190878
rs7404049
rs4984707
rs11649498
...
```

NOTE: `SNPLIST` has to be set to 1.

2.3 Options

Keyword	Example	Default	Description
# data files			
FILE			Path to SNP/Pedigree/Genotype files. The input format is the standard PLINK format (.ped/.map) (section :TP 2.2.1)
TFILE			Path to SNP/Pedigree/Genotype files. The input format is the transposed PLINK format (.tped/.tfam) (section :TP 2.2.1)
BFILE			Path to SNP/Pedigree/Genotype files. The input format is the binary PLINK (format (.bed/.bim/.fam) section :TP 2.2.1)
FAM / TFAM / BFAM	./test.fam		Path to the .fam/.tfam file
MAP	./test.map		Path to the .map file
BMAP / BIM	./test.bim		Path to the .bim file
PED	./test.ped		Path to the .ped file
TPED	./test.tped		Path to the .tped file
BPED / BED	./test.bed		Path to the .bed file
ANNOTATIONFILE	./annotationfile.txt		Path to the annotation file (section :TP 2.2.2). Only needed if genetic criteria is used NOTE: GENETIC_IMPACT and M_WITH_GENETIC_IMPACT have to be selected NOTE: To improve running time it is recommended to sort the file by chromosome and position
PATHWAYFILE	./pathwayfile.txt		Path to the pathway file (section :TP 2.2.3) NOTE: PATHWAY has to be switched on 1 NOTE: To improve running time it is recommended to sort the file by pathway name
COVARIATEFILE	./covariatefile.txt		Path to the covariate file (section :TP 2.2.4). Only needed if covariates are included into the analysis NOTE: Covariates are selected with the COVARIATES keyword (see below). If all covariates are to be used, write 1-10; NOTE: Missing data value is "–" or "x"

MODELFILE	./modelfile.txt		Path to the model file (sections :TP 2.2.5, :TP 4.2.3). Only needed for user-defined logistic regression models that are not included in the list of tests NOTE: TEST has to be switched on M
COMBIFILE	./combifile.txt		Path to the combi file (section :TP 2.2.6) NOTE: COMBILIST has to be put on 1
SNPFILE	./snpfile.txt		Path to the SNP file (section :TP 2.2.7) NOTE: SNPLIST has to be put on 1
# Selection/ reduction of data			
SNPLIST		0	Only SNPs from the SNP file will be analyzed: 1=yes, 0=no
COMBILIST		0	Only SNPs pairs/triples form the combi file will be analyzed: 1=yes, 0=no
ONLY_MALE		0	Only male persons will be analyzed: 1=yes, 0=no
ONLY_FEMALE		0	Only female persons will be analyzed: 1=yes, 0=no
POSCHOICE	chr4; chr12,123000-160000; chr24;		Selection of certain regions (ONLY the stated regions will be analyzed) NOTE: Use the notation shown in the example
NEGCHOICE	chr4; chr12,123000-160000; chr24;		Exclusion of certain regions (The stated regions will NOT be analyzed) NOTE: Use the notation shown in the example
MATCHINGCHOICE	chr4; chr12,123000-160000; chr24;		Selection of certain regions for pairwise matching NOTE: Use the notation shown in the example
QT		0	Use the analysis of quantitative traits: 1=yes, 0=no (case-control)
MISSING_PHENO	-9	-99/0	Specify the missing phenotype value By default it is -99 for QT 1, and is 0 for QT 0 NOTE: When using QT 1, make sure that the missing phenotype value is defined
LD_DISTANCE			Define the LD distance between SNP pairs, in base pairs (bp)

# Quality control			
HWE_P_CASE	0.001	1	QC-threshold for HWE in cases (p-value cutoff)
HWE_P_CONTROL	0.01	1	QC-threshold for HWE in controls (p-value cutoff)
MRDIFF	0.02	1	QC-threshold missing rate (mr): individuals and SNPs worse than the average mr + `MRDIFF` will be deleted (iterative algorithm)
MAF		0	SNPs with lower MAF will be deleted
DOIBS		0	Calculate average IBS-value of all possible pairs: `1`=yes and stop afterwards, 2=yes and proceed with analysis, `0`=no
IBS_SD_RELATIVES		4.0	Limit in standard deviations for high/low-IBS. Pairs that exceed this limit are considered to be relatives
IBS_SD_OUTLIER		4.0	Limit in standard deviations. People whose low-IBS-count exceed this limit are considered to be outlier
# Single- and multi-marker tests			
SINGLE_MARKER			`1`=Armitage's trend test, `2`=genotype test with 2 d.f., `3`=logistic regression with 1 d.f., `4`=logistic regression with 2 d.f.
TWO_MARKER		0	Change parameter from `0` to `1` to perform two-marker analysis NOTE: It's not possible to select two-marker and three-marker analysis together
THREE_MARKER		0	Change parameter from `0` to `1` to perform three-marker analysis NOTE: It's not possible to select three-marker and two-marker analysis together
PRETEST		1	Pre-test is selected: `1`=yes, `0`=no
PRETEST_CUTOFF		0.05	Pre-test threshold. If Pre-test yields p-value <= cutoff, `TEST` is conducted
TEST			Multi-marker tests: `1`=chi-square test on contingency table (8 d.f.), `2`=log-linear model (4 d.f., test FOR interaction), `3-12`=logistic regression models, `15`=case-only allelic interaction, `16`=case-only SNP-SNP interaction, M=User-defined logistic regression model (`MODELFILE`). Detailed description are presented in the model table

			NOTE: tests 1-2 are possible for two-marker and three-marker analysis, tests 3-8 are used only for two-marker analysis and tests 9-12 are only for three-marker analysis
# Covariates			
COVARIATES	1-5;7;8;		Select the covariates which are to be considered, separate by ";", condense with "-". Numbers must be put in ascending order. NOTE: It's only possible to use this option if the path to the covariate file (keyword: `COVARIATEFILE`) is declared.
SEXCOV			Sex as a covariate: `1`=yes, `0`=no. Can be selected without a covariate file
COVARIANCE_MATRIX		0	Change parameter from 0 to 1 to get the upper triangle elements of the covariance matrix of model parameters in output files. NOTE: relevant only for regression models.
# Priorities			
SINGLETOP	1000		Based on these single-marker p-values (Armitage's trend test or genotype test with 2 d.f.), a list of n top SNPs will be computed. The length can be specified with this option
M_WITH_SINGLETOP	2		Number of SNPs (0,1,2,3) which have to be taken from the single-marker p-values toplist NOTE: `SINGLETOP` has to be specified NOTE: `M_WITH_SINGLETOP 3` is only possible when `THREE_MARKER 1`
GENETIC_IMPACT	3		Genetic impact: `0`=none (gene desert), `1`=close to gene, `2`=within gene, `3`=within coding region, `4`=non-synonymous NOTE: To use this option the path to the annotation file (keyword: `ANNOTATIONFILE`) has to be declared
M_WITH_GENETIC_IMPACT	2		Number of SNPs (0,1,2,3) involved in the analysis which shall have at least the selected genetic impact NOTE: `GENETIC_IMPACT` has to be specified

		NOTE: `M_WITH_GENETIC_IMPACT 3` is only possible when `THREE_MARKER 1`
SNP1		Fix first SNP for analysis, specify rs number
SNP2		Fix second SNP for analysis, specify rs number
SNP3		Fix third SNP for analysis (possible only when `THREE_MARKER 1`), specify rs number
# Pathway Association Analysis		
PATHWAY		Pathway information is used as a prior for interaction analysis: `1`=yes, `0`=no NOTE: To use this option the path to the pathway file (keyword: `PATHWAYFILE`) has to be declared
PATHWAYANALYSIS	0	Pathway association analysis with one of the six pathway association tests (chapter chapter6) NOTE: Choose, for instance, a number of SNP, multiply it by 0.05 ($50000 \times 0.05 = 25000 \rightarrow$ `SINGLETOP 25000`). Simulations should be put on at least 999 NOTE: To use this option the path to the pathway file (keyword: `PATHWAYFILE`) has to be declared
PATHWAYTEST	2	`1`=SNP ratio, `2`=Fisher score (all significant SNPs), `3`=gene ratio (ALIGATOR), `4`=Fisher Max (best SNP per gene), `5`=Fisher MaxPlus, `6`=interaction ratio NOTE: `PATHWAYANALYSIS` has to be selected. NOTE: To use this option the path to the pathway file (keyword: `PATHWAYFILE`) has to be declared
P_INTERRATIO		p-value cutoff for `TEST 6`
# Genome-wide Haplotype Analysis		
HAPLO		Genome-wide Haplotype Analysis (GWHA) for all SNPs pairs/triples less than `HAPLO_DIST` apart NOTE: To use two-marker select `TWO_MARKER 1`, for three-marker select `THREE_MARKER 1`

			NOTE: To use this option the following options have to be selected: `HAPLO 1` and `HAPLO_DIST` have to be specified
HAPLO_DIST	50000	0	To define the haplotype distance. Here all pairs/triples less than 50000 bp apart
DOHAPFILE	0		Produce proxy-file for YAMAS (section :`TP` 2.5)
# MC-simulation			
SIMULATION			Number of MC-simulations (0=no simulations will be conducted)
MC_WITH_SM		0	`1`=MC-simulations account for multi-marker AND single-marker tests. `0`=MC-simulations account only for multi-marker tests
PRINTTOP	100		Best n multi-marker p-values are printed
# Test for supra-multiplicativity of SNP sets			
LIABILITY		0	Test for supra-multiplicativity of SNP sets (1=SMT, 2=ESMT)
SNP_COVARIATES	rs11248850;		reported SNPs NOTE: `LIABILITY` has to be specified.
SNP_RISK	G;T;A;C;R;		risk alleles (R=risk allele shall be derived from the data) NOTE: `LIABILITY` has to be specified.
# Stratified analysis			
STRATIFY_STRUCT			PAIR, GROUP or CLUSTER
STRATIFY_VALID		VICINITY	CLUSTER or VICINITY
GROUP_TEST			SQUARE or LINEAR
CLUSTER_COVARS		0	0=off, 1=on
# Output			
PFILTER			p-value cutoff for multi-marker analysis
OUTPUTNAME	./myproject/test		Specify the output name
ANNOTATE			Add annotation information in the output files
GENECOL		5	Specify column number of gene in the annotation information

2.4 Output files

The following output files are created by the program:

- singlemarker.txt: results of single-marker-analysis
- singelemarkerTop.txt: top results of single-marker-analysis
- bestMarkerCombi2.txt: results of two-marker-analysis
- bestMarkerCombi2Details.txt: contingency tables of two-marker-analysis and logistic/linear regression parameters
- bestMarkerCombi3.txt: results of three-marker-analysis
- bestMarkerCombi3Details.txt: contingency tables of three-marker-analysis and logistic/linear regression parameters
- toplistMC.txt: p-values from MC-Simulation
- mibs.txt: IBS-matrix
- ibsRelatives.txt: persons who are defined as relatives
- ibsOutlier.txt: persons who are defined as outliers
- deletedSnps.txt: SNPs deleted after QC
- deletedPerson.txt: persons deleted after QC
- errorMessage.txt: errors that occurred during the program flow/execution
- logfile.txt

2.4.1 Output single marker

The parameters in the output file depend on the used single-marker-test (Armitage's trend test, genotype test with 2 d.f., logistic regression with 1 d.f., logistic regression with 2 d.f.). The following list shows all possible parameters.

- **No** Line number in the input file
- **Chr** Chromosome
- **rs_No** rs number
- **Position** Position
- **Gene** Gene
- **minor** A-allele = allele listed first = minor allele
- **major** B-allele = allele listed second = major allele
- **MAF(All/Co/Ca)** Minor allele frequency
- **SNP_MR** Missing rate
- **HWE_Ca** Hardy-Weinberg Equilibrium cases
- **HWE_Co** Hardy-Weinberg Equilibrium controls

- **P_Single-marker** single-marker p-value
- **P_Corr** corrected single-marker p-value (Sidak-Correction)
- **A_Ca_N** Number of A-alleles in cases (minor allele)
- **B_Ca_N** Number of B-alleles in cases (major allele)
- **A_Co_N** Number of A-alleles in controls (minor allele)
- **B_Co_N** Number of B-alleles in controls (major allele)
- **A_Ca** Allele frequency of allele A in cases (minor allele)
- **B_Ca** Allele frequency of allele B in cases (major allele)
- **A_Co** Allele frequency of allele A in controls (minor allele)
- **B_Co** Allele frequency of allele B in controls (major allele)
- **OR_A** Odds ratio of allele A (minor allele)
- **LCL_A** lower limit of confidence interval of allel A (minor allele)
- **RCL_A** upper limit of confidence interval of allel A (minor allele)
- **OR_B** Odds ratio of allele B (major allele)
- **LCL_B** lower limit of confidence interval of allel B (major allele)
- **RCL_B** upper limit of confidence interval of allel B (major allele)
- **beta1** Beta of linear/logistic regression (allelic effect)
- **se1** Standard error of linear/logistic regression (allelic effect)
- **OR1** Odds ratio of linear/logistic regression (allelic effect)
- **LCL1** lower limit of confidence interval of linear/logistic regression (allelic effect)
- **RLC1** upper limit of confidence interval of linear/logistic regression (allelic effect)
- **beta1D** Beta of linear/logistic regression (allelic effect)
- **se1D** Standard error of linear/logistic regression (dominance effect)
- **OR1D** Odds ratio of linear/logistic regression (dominance effect)
- **LCL1D** lower limit of confidence interval of linear/logistic regression (dominance effect)
- **RLC1D** upper limit of confidence interval of linear/logistic regression (dominance effect)
- **sigma1[·][·]** upper triangle elements of the covariance matrix of model parameters

2.5 Produce SNP-Proxy-Files for Meta-Analysis with YAMAS

With INTERSNP it is possible to produce a file with r^2-values for all SNPs that lie within a pre-specified distance, for instance from HAPMAP and/or 1000 Genomes data (.tped/.tfam files). In addition to the basic keywords (TPED/TFAM, QC options,...), the following keywords have to be specified in the selectionfile:

```
TWO_MARKER 1
TEST 13
QT 1
HAPLO 1
HAPLO_DIST 10000
DOHAPFILE 1
```

In the above example, r^2 will be computed for all SNPs with a distance smaller than 10000 bp. The outputfile (*hapfile.txt) will contain all SNP pairs for which $r^2 \geq 0.20$.

Example:

```
16 rs1211375 A C rs3918352 A G 0.338 1000 0
16 rs3918352 A G rs11642609 C T 0.659 2000 1
16 rs8051485 T C rs11248914 C T 0.212 15722 1
...
```

Column 1) chromosome

Column 2) marker id of the first marker, e. g. rs-id of the first SNP

Column 3) minor allele of the first marker or SNP

Column 4) major allele of the first marker or SNP

Column 5) marker id of the second marker, e. g. rs-id of the second SNP

Column 6) minor allele of the second marker or SNP

Column 7) major allele of the second marker or SNP

Column 8) the difference of the markers base pair positions

Column 9) r^2

Column 10) identifies proxy alleles. In line 1, `0` indicates that the haplotype A-A (composed of columns 3 and 6, for each SNP the first listed allele) is more frequent than expected under linkage disequilibrium. It immediately follows that also the C-G haplotype is more frequent than expected under linkage disequilibrium. Therefore, allele A from SNP 1 and allele A from SNP 2 are mutual proxy alleles, and likewise allele C from SNP 1 and allele G from SNP 2. In line 2, the situation is reverse, as indicated by a `1` in the last column. Here, allele A (column 3) from SNP1 and allele T (second(!) allele of SNP 2, column 7) are the mutual proxy alleles.

A more detailed output is obtained with the choice DOHAPFILE 2.

```
16 rs1211375 A C rs3918352 A G 0.378 0.622 0.425 0.575 0.300 0.078
0.125 0.497 0.641 0.338 0
16 rs1203957 T G rs3918352 A G 0.110 0.890 0.425 0.575 0.110 0.000
0.315 0.575 1.000 0.167 0
16 rs3918352 A G rs11642609 C T 0.412 0.588 0.485 0.515 0.000 0.412
0.485 0.103 1.000 0.659 1
```

The first seven columns are the same as before. Columns 8-9 are allele frequencies for SNP 1, columns 10-11 are allele frequencies for SNP 2. Columns 12-15 are the four haplotype frequencies, here in lexicographic order, i.e. ignoring the proxy definition. Column 16 is D', column 17 is r^2, and column 18 is the proxy indicator. The computation of the indicator is based on the allele and haplotype frequencies listed.

For more details visit the YAMAS web-page `http://yamas.meb.uni/bonn.de/`

Chapter 3

Examples

Here we present how to conduct a full Genome-wide Interaction Analysis with a pre-test, six GWIA strategies to demonstrate the usage of the software and how to do a Genome-wide Haplotype Analysis. Notice that only the relevant options are presented in the selection file.

3.1 Genome-wide Interaction Analysis with Pre-test

```
TPED    ./test.tped   // path to the .tped file
TFAM    ./test.tfam   // path to the .tfam file
SINGLE_MARKER  1  // Armitage's trend test
PRETEST 1 // pre-test: 1=yes
PRETEST_CUTOFF 0.01 // pre-test threshold. Here, threshold for
                       pre-test is 0.01
TWO_MARKER  1  // two-marker analysis: 1=yes
TEST  2  // log-linear model, test for genotypic interaction
OUTPUTNAME  ./myproject/test  // specify the output name
END
```

3.2 Strategy I

All pairs of SNPs with at least one SNP among the top 10 single-marker results

```
TPED    ./test.tped  // path to the .tped file
TFAM    ./test.tfam  // path to the .tfam file
SINGLE_MARKER  1  // Armitage's trend test
TWO_MARKER  1  // two-marker analysis: 1=yes
TEST  2  // log-linear model, test for genotypic interaction
SINGLETOP   10  // top 10 single-marker results
M_WITH_SINGLETOP  1  // at least one SNP among the top 10
                        single-marker results
OUTPUTNAME  ./myproject/test   // specify the output name
END
```

3.3 Strategy II

All pairs of SNPs which are among the top 1000 single-marker results

```
TPED    ./test.tped  // path to the .tped file
TFAM    ./test.tfam  // path to the .tfam file
SINGLE_MARKER  1  // Armitage's trend test
```

```
TWO_MARKER  1  // two-marker analysis: 1=yes
TEST  5  // logistic regression model:
            test for additive interaction
SINGLETOP  1000  // top 1000 single-marker results
M_WITH_SINGLETOP    2   // all pairs of SNP which are among the top 1000
                           single-marker results
OUTPUTNAME  ./myproject/test   // specify the output name
END
```

3.4 Strategy III

All pairs of SNPs which are among the top 50000 single-marker results and also lie in a coding region of a gene

```
TPED    ./test.tped  // path to the .tped file
TFAM    ./test.tfam  // path to the .tfam file
ANNOTATIONFILE  ./annotationfile.txt    // path to the annotation file
SINGLE_MARKER  1  // Armitage's trend test
TWO_MARKER  1  // two-marker analysis: 1=yes
TEST  5  // logistic regression model:
            test for additive interaction
SINGLETOP       50000  // top 50000 single-marker results
M_WITH_SINGLETOP    2   // all pairs of SNPs which are among the
                           top 50000 single-marker results
GENETIC_IMPACT  3  // genetic impact: within coding region
M_WITH_GENETIC_IMPACT   2  // all pairs of SNPs which lie in a
                              coding region of a gene
OUTPUTNAME  ./myproject/test  // specify the output name
END
```

3.5 Strategy IV

All pairs of non-synonymous SNPs

```
TPED    ./test.tped // path to the .tped file
TFAM    ./test.tfam     // path to the .tfam file
ANNOTATIONFILE  ./annotationfile.txt // path to the annotation file
SINGLE_MARKER  1  // Armitage's trend test
TWO_MARKER  1  // two-marker analysis: 1=yes
TEST  5  // logistic regression model:
            test for additive interaction
GENETIC_IMPACT  4  // genetic impact: non-synonymous SNPs
M_WITH_GENETIC_IMPACT  2  // all pairs of non-synonymous SNPs
OUTPUTNAME  ./myproject/test   // specify the output name
END
```

3.6 Strategy V

All pairs of SNPs which are among the top 50000 single-marker results and also lie in the same pathway

```
TPED    ./test.tped // path to the .tped file
TFAM    ./test.tfam     // path to the .tfam file
PATHWAYFILE ./KEGG_2_SNP_229.txt // path to the pathway file
```

```
SINGLE_MARKER  1  // Armitage's trend test
TWO_MARKER  1  // two-marker analysis: 1=yes
TEST  2  // log-linear model: test for genotypic interaction
SINGLETOP  50000  // top 50000 single-marker results
M_WITH_SINGLETOP  2  // all pairs of SNPs which are among the
                        top 50000 single-marker results
PATHWAY  1  // all pairs of SNPs which lie in the same pathway
OUTPUTNAME  ./myproject/test   // specify the output name
END
```

3.7 Genome-wide Haplotype Analysis

```
TPED    ./test.tped // path to the .tped file
TFAM    ./test.tfam // path to the .tfam  file
SINGLE_MARKER  1  // Armitage's trend test
TWO_MARKER  1  // two-marker analysis: 1=yes
HAPLO 1 // haplotype analysis
HAPLO_DIST  50000  // pathway analysis with Fisher score
OUTPUTNAME  ./myproject/test  // specify the output name
END
```

Chapter 4

Tests in INTERSNP

4.1 Single-marker tests

By default, a single-marker p-value is computed for all QC-SNPs. For the autosomes and the pseudoautosomal region of the Y chromosome the user can either choose Armitage's trend test (Armitage, 1955) or the genotype test with 2 degrees of freedom (d.f.). It is also possible to include covariates in the logistic/linear regression, the allelic test with 1 d.f. and the genotypic test with 2 d.f. Another advantage of the logistic/linear regression is that all chromosomes, including X and Y, can be analyzed. Without logistic/linear regression Y-chromosomal markers are evaluated with the χ^2-test for the 2×2 table of allele counts in male individuals. For X-chromosomal SNPs, we use the allele-based test with 1 d.f. suggested by D. Clayton (Clayton, 2008). Claytons's approach guarantees that hemizygote males and homozygote females contribute equally to the test statistic, reflecting the fact that only one of the X-chromosomes is active in females.

4.2 Two- and three-marker analysis

4.2.1 χ^2-test for full genotype contingency table (`TEST 1`)

For two SNPs, there are $3 \times 3 = 9$ two-marker genotypes. We consider the respective $3 \times 3 \times 2$ contingency table of counts in cases and controls. Let T be the standard test statistic for contingency tables, i.e., $T = \sum_{i,j,k} \frac{(O_{i,j,k} - E_{i,j,k})^2}{E_{i,j,k}}$, where $i, j \in \{1, 2, 3\}$, $k \in \{1, 2\}$, $O_{i,j,k}$ is the observed number of counts in cell (i, j, k) and $E_{i,j,k}$ is the number of counts in cell (i, j, k) expected under the null hypothesis. T is χ^2-distributed with 8 d.f. By construction, `TEST 1` does not yield a test for interaction, but a test that has power if both SNPs interact. The test can become significant through marginal SNP effects alone.

4.2.2 Test for interaction using a log-linear model (`TEST 2`)

Following (Bishop et al., 2007) the observations x_{ijk} of the $3 \times 3 \times 2$ contingency table are modeled by fitting a log-linear model with expected cell counts m_{ijk} to the cell entries $\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$ without the term $u_{123(ijk)}$. Testing the null hypothesis $H_0 : u_{123(ijk)} = 0$ yields an explicit test for genotypic interaction. With maximum likelihood estimates m_{ijk} for the cell counts we obtain the test statistic $T = 2 \sum_{i,j,k} x_{ijk} \log \frac{x_{ijk}}{m_{ijk}}$ that is χ^2-distributed with 4 d.f. The test is equivalent to the respective 4 d.f. genotypic test in the logistic regression framework, but is much faster to compute.

4.2.3 Logistic regression (`TESTS 3-12`)

`TEST 1` and `TEST 2` are quick screening tests.

With logistic regression (Cordell and Clayton, 2002) a more flexible modeling is possible: Test for and under interaction, allelic und genotypic tests and the use of covariates.

We define $\mathrm{logit}(p) := \log(\frac{p}{1-p}) = \beta^T x$, where β is the vector of coefficients to be estimated and x is a vector representing genotype information.

4.2.3.1 Parameter coding

We follow the coding suggested by (Cordell and Clayton, 2002). For each SNP i, $i \in \{1, 2, 3\}$, we model its allelic effect x_i by coding the genotypes $(1, 1)$, $(1, 2)$, and $(2, 2)$ as $x_1 = -1$, $x_2 = 0$, $x_3 = 1$. Next, we model dominance effects $x_{i,D}$, $i \in \{1, 2, 3\}$, as $x_{1,D} = -0.5$, $x_{2,D} = 0.5$, $x_{3,D} = -0.5$ for the genotypes $(1, 1)$, $(1, 2)$, and $(2, 2)$, respectively. By multiplication, we obtain interaction terms, for instance, $x_1 x_2$ represents allelic interaction between SNPs 1 and 2, while $x_{1,D} x_{2,D}$ represents interaction between the dominance terms of SNPs 1 and 2. Note that these interaction terms code interaction on an additive logit scale and, hence, on a multiplicative odds ratio scale.

4.2.3.2 Likelihoods

Let β_0 be the intercept parameter that defines the baseline likelihood $L_0 := \mathrm{logit}(p) = \beta_0$. Next, the likelihood $L_1^A := \beta_0 + \beta_1 x_1$ models the allelic effect of SNP 1 and comparison to L_0 yields a likelihood-ratio test with 1 degree of freedom. Analogously, comparison of $L_1^G := \beta_0 + \beta_1 x_1 + \beta_{1,D} x_{1,D}$ to L_0 yields a genotypic test for SNP 1 with 2 d.f. In general, we let $L_{1,2}^A$ and $L_{i,j}^G$ denote likelihoods containing allelic terms, or, respectively, allelic and genotypic terms, for SNPs 1 and 2. In addition let, $L_{1,2}^{A,I}$ and $L_{1,2}^{G,I}$ be likelihoods that also contain interaction terms, for instance, $L_{1,2}^{A,I} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2$ and $L_{1,2}^{G,I} = \beta_0 + \beta_1 x_1 + \beta_{1,D} x_{1,D} + \beta_2 x_2 + \beta_{2,D} x_{2,D} + \beta_{1,2} x_1 x_2 + \beta_{1,2D} x_1 x_{2,D} + \beta_{1D,2} x_{1,D} x_2 + \beta_{1D,2D} x_{1,D} x_{2,D}$

Test No	Test	Formula	d.f.	Comment
1	chi-square-test		8	Full genotype test (contingency table)
2	log-linear model	$l_{1,2}^{G,I}$ vs $l_{1,2}^{G}$	4	Test for genotypic interaction
3	logistic regression	$L_{1,2}^{A,I}$ vs L_0	3	Full additive test
4	logistic regression	$L_{1,2}^{G,I}$ vs L_0	8	Full genotype test
5	logistic regression	$L_{1,2}^{A,I}$ vs $L_{1,2}^{A}$	1	Test for additive interaction
6	logistic regression	$L_{1,2}^{G,I}$ vs $L_{1,2}^{G}$	4	Test for genotypic interaction
7	logistic regression	$L_{1,2}^{A,I}$ vs L_1^{A}	2	Additional allelic effect of SNP2
8	logistic regression	$L_{1,2}^{G,I}$ vs L_1^{G}	6	Additional genotypic effect of SNP2
9	logistic regression	$L_{1,2,3}^{A,I}$ vs L_0	7	Full additive test
10	logistic regression	$L_{1,2,3}^{A,I}$ vs $L_{1,2,3}^{A}$	4	Test for allelic interaction
11	logistic regression	$L_{1,2,3}^{A,I}$ vs $L_{1,2,3}^{A,I_2}$	1	Test for 3-fold allelic interaction
12	logistic regression	$L_{1,2,3}^{A,I}$ vs $L_{1,2}^{A,I}$	4	Additional allelic effect of third locus
15	chi-square-test		1	case-only allelic interaction test
16	chi-square-test		4	case-only SNP-SNP interaction test
M	logistic regression	-	-	User defined logistic regression (modelfile)

Table 4.1: Model table for case-control data

4.2.4 Linear regression

To analyze quantitative traits we use linear regression. The parameter coding is equal to logistic regression. For linear regression, we define $y = \beta^T x$, where y is a quantitative trait variable, and the sum of squared errors (SSE) is computed. The errors measure, per person i, the deviation between the observed value y_i and its estimate $\hat{y}_i$ predicted by the model.

Test No	Test	Formula	d.f.	Comment
3	linear regression	$SSE_{1,2}^{A,I}$ vs SSE_0	3	Full additive test
4	linear regression	$SSE_{1,2}^{G,I}$ vs SSE_0	8	Full genotype test
5	linear regression	$SSE_{1,2}^{A,I}$ vs $SSE_{1,2}^{A}$	1	Test for additive interaction
6	linear regression	$SSE_{1,2}^{G,I}$ vs $SSE_{1,2}^{G}$	4	Test for genotypic interaction
7	linear regression	$SSE_{1,2}^{A,I}$ vs SSE_{1}^{A}	2	Additional allelic effect of SNP2
8	linear regression	$SSE_{1,2}^{G,I}$ vs SSE_{1}^{G}	6	Additional genotypic effect of SNP2
9	linear regression	$SSE_{1,2,3}^{A,I}$ vs SSE_0	7	Full additive test
10	linear regression	$SSE_{1,2,3}^{A,I}$ vs $SSE_{1,2,3}^{A}$	4	Test for allelic interaction
11	linear regression	$SSE_{1,2,3}^{A,I}$ vs $SSE_{1,2,3}^{A,I_2}$	1	Test for 3-fold allelic interaction
12	linear regression	$SSE_{1,2,3}^{A,I}$ vs $SSE_{1,2}^{A,I}$	4	Additional allelic effect of third locus
M	linear regression	-	-	User defined linear regression (modelfile)

Table 4.2: Model table for quantitative traits

4.2.4.1 Tests (2 SNPs)

For two SNPs, comparison of $L_{1,2}^{A}$ against L_0 yields a full allelic test with 3 d.f. (`TEST 3`), while comparison of $L_{1,2}^{G}$ against L_0 yields a full genotypic test with 8 d.f. (`TEST 4`). In order to test for allelic interaction, one compares $L_{1,2}^{A,I}$ against $L_{1,2}^{A}$ (1 d.f., `TEST 5`) and in order to test for genotypic interaction one compares $L_{1,2}^{G,I}$ against $L_{1,2}^{G}$ (4 d.f., `TEST 6`). Furthermore, it is possible to test for the additional effect of SNP 2 while controlling for the effect of SNP 1 by comparing $L_{1,2}^{A,I}$ against L_{1}^{A} (2 d.f., `TEST 7`), or, by comparing $L_{1,2}^{G,I}$ against L_{1}^{G} (6 d.f., `TEST 8`). These tests are summarized in the model Table Model table for case-control data4.1 and serve as INTERSNP standard tests that can called via their test number.

4.2.4.2 Tests (3 SNPs)

For three SNPs, likelihoods and tests can be formalized analogously. Here, we describe only allelic tests, but note that INTERSNP allows also explicit model specification, in particular genotypic 3-SNP tests. Let $L_{1,2,3}^{A} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ be the 3-SNP allelic likelihood, let $L_{1,2,3}^{A,I_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 + \beta_3 x_3$ be the 3-SNP allelic likelihood including the interaction term for SNPs 1 and 2 and let $L_{1,2,3}^{A,I} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 + \beta_3 x_3 + \beta_{1,3} x_1 x_3 + \beta_{2,3} x_2 x_3 + \beta_{1,2,3} x_1 x_2 x_3$ be the allelic likelihood including all pairwise and the three-fold interaction term. Then, testing $L_{1,2,3}^{A,I}$ against L_0 yields a full 3-SNP allelic test with 7 d.f. (`TEST 9`), while testing $L_{1,2,3}^{A,I}$ against $L_{1,2,3}^{A}$ yields a test for two- and three-fold allelic interaction (4 d.f., `TEST 10`). Testing $L_{1,2,3}^{A,I}$ against $L_{1,2,3}^{A,I_2}$ yields a test for three-fold allelic interaction (1 d.f., `TEST 11`). Finally, testing $L_{1,2,3}^{A,I}$ against $L_{1,2}^{A,I}$ yields a test for the additional allelic effect of SNP 3 while controlling for SNPs 1 and 2 (4 d.f., `TEST12`).

4.2.4.3 X-chromosome

For X-chromosomal markers, all dominance and dominance interaction terms are ignored.

4.2.4.4 Covariates

All regression tests can be combined with up to 10 covariates (Options: `COVARIATES`, `COVARIATEFILE`). In particular, adjustment for population strata is possible.

4.2.4.5 User-defined regression models

Further tests using logistic regression can be specified via a modelfile. Choose `M` for the option `TEST` and specify the path to the modelfile (`MODELFILE` option). Include/exclude likelihood parameters with 1/0, respectively, as shown in the modelfile (see :`TP` 2.2.5) provided in the download area. Likelihood $L1$ is tested against likelihood $L2$.

4.2.5 Pre-tests

Linear and logistic regression allow flexible modeling and inclusion of covariates, but their computation can be time-consuming, mostly because of the operations on the relatively large design matrices involved. Hence, when a Genome-wide Interaction Analysis shall be conducted, it makes sense to, first, apply a simple, easy-to-compute, pre-test to all SNP pairs and, second, to apply the more sophisticated test only to those pairs for which the pre-test yields a p-value smaller than `PRETEST_CUTOFF`. A cutoff of 0.01 is a reasonable choice, since it will reduce the number of computations of the "proper" test by a factor of at least 100 on average, but guarantees that no "real" p-values smaller than 10^{-5} are missed. Since p-values in the range of 10^{-12} are required for a complete GWIA (Becker, 2011), the loss of pairs with p-values less significant than 10^{-5} will in most cases be irrelevant. Even a cutoff of 0.001 or 0.0001 for the pre-test would be acceptable. The only situation in which relevant SNP pairs may be missed occurs when there are covariates with an extreme impact on the p-value, simply because the pre-tests ignore covariates.

4.2.5.1 Pre-test allelic interaction (logistic regression)

When `PRETEST 1` is selected in combination with `TEST 3` (full allelic test, 3 d.f.) or `TEST 5` (test for allelic interaction, 1 d.f.), a log-linear model with 1 d.f. (test for allelic interaction) is applied as a pre-test. We define the log-linear model in analogy to the log-linear model in the genotypic case (our `TEST 2`, section :`TP` 4.2.2). There, the $3 \times 3 \times 2$ genotype contingency table is considered. In the allelic case, we consider the $2 \times 2 \times 2$ two-SNP-allele table, where the first dimension reflects the two alleles of SNP 1, the second dimension reflects the two alleles of SNP 2, and the last dimension reflects cases/control-status. Double heterozygote cases (controls) contribute 0.5 units to each cell of the 2×2 table of cases (controls). This practice is a simple work-around to treat double heterozygotes but renders the accompanying log-linear test conservative when its test statistic is evaluated with the χ^2-distribution with 1 d.f. In this sense, the 1 d.f. log-linear test is not a proper test. However, its p-values are strongly correlated with the p-values of the test for allelic interaction in the logistic model, and, therefore, are well suited to serve as a pre-test, when the cutoff is chosen as suggested above.

4.2.5.2 Pre-test genotypic interaction (logistic regression)

In the genotypic case, the log-linear test (`TEST 2`, section :`TP` 4.2.2) would be the obvious pre-test. Since this is a proper test in its own right, a useful strategy is to conduct a complete GWIA with `TEST 2`, produce a `COMBILIST` (section :`TP` 2.2.6) with the best results, and run another test on the `COMBILIST`.

4.2.5.3 Pre-test genotypic interaction (linear regression)

When `PRETEST 1` is selected in combination with `TEST 4` (full genotype test, 8 d.f.) or `TEST 6` (test for genotypic interaction, 4 d.f.), an ANOVA with 4 d.f. (test for genotypic interaction) is applied as a pre-test. Note that when the pre-test is combined with `TEST 4`, pairs with marginal effects in both SNPs but with just modest interaction may be missed.

4.2.5.4 Pre-test allelic interaction (linear regression)

When `PRETEST 1` is selected in combination with `TEST 3` (full allelic test, 3 d.f.) or `TEST 5` (test for allelic interaction, 1 d.f.), an ANOVA with 1 d.f. (test for allelic interaction) is applied as a pre-test. Double heterozygotes are treated as described for the logistic regression, with the same consequences. Note that when the pre-test is combined with `TEST 3`, pairs with marginal effects in both SNPs but with just modest interaction may be missed.

Chapter 5

Case-only analysis

Case-only (or case-series, case-control without controls) design is based on a sample consisting of cases only. It is an efficient approach to test for gene-environment or gene-gene interaction in disease etiology, when there is a strong reason to believe that particular assumptions are fulfilled. Concerning a test for gene-environment interaction it is necessary to assume that the genetic type and the exposure occur independently in the population. A test for gene-gene interaction is reasonable to perform assuming that the frequencies of genetic variants are independent in the population or, equivalently, that the genes/**SNPs under study are in linkage equilibrium**.

Here, we accept that the genetic interaction between two genetic factors (e.g. genes, SNPs, alleles) is present if the joint effect of two factors on a disease differs from the additive combination of independent effects from both of them. The joint effect of two genetic factors is the effect due to the presence of both of them. The independent effect of one of the factors is its effect on a disease in the absence of the other factor. Sometimes the term "epistasis" is used instead of genetic interaction in the literature. The same definition is valid for the gene-environment interaction, where one of the genetic factors is replaced by the exposure.

From now on only the gene-gene interaction is addressed.

If the data satisfy the assumptions discussed above, then the so called case-only test for genetic interaction can be performed by using a χ^2- test of independence between genotypes (with four degrees of freedom) or between alleles (with one degree of freedom) as well as logistic regression models. Notably, the case-only analysis leads sometimes to a more valid estimation of interaction than the corresponding case-control analysis, i.e. case-only based tests possess a higher power, see e.g. (Cordell, 2009), (Gatto et al., 2004), (Piegorsch et al., 1994), (Yang et al., 1999).

INTERSNP suggests several tests to check for the gene-gene interaction, where the genetic factors are chosen to be a pair of SNPs (SNP-SNP interaction). Starting from a pair of SNPs, the allelic interaction test can be performed as well.

IMPORTANT: In our study it was observed that the LD distance between two SNPs has to be at least 20 Mb $(20000000$ bp) in order to satisfy the assumption of linkage equilibrium. The recommended LD distance is even bigger and equals **50 Mb**.

5.1 Case-only allelic interaction test (`TEST 15`)

`TEST 15` in INTERSNP is asymptotically equivalent to the case-only allelic interaction test in whole-genome analysis package PLINK and based on a χ^2-test of independence between two alleles with one degree of freedom, see e.g. (Cordell, 2009), (Purcell et al., 2007). For this test the given SNPs data are organized in a 3×3 table of genotypic counts, which is in turn reduced to the 2×2 table of allelic counts.

5.2 Case-only SNP-SNP interaction test (TEST 16)

TEST 16 in INTERSNP suggests to check for SNP-SNP interaction in the case-only mode. TEST 16 is based on a χ^2-test of independence between two SNPs with four degrees of freedom, where the 3×3 table of genotypic counts is used. Let x_{ij}, $i,\ j \in \{1,2,3\}$ be the observed counts. Denote $x_{+j} = \sum_{i=1}^{3} x_{ij}$, $j \in \{1,2,3\}$, $x_{i+} = \sum_{j=1}^{3} x_{ij}, i \in \{1,2,3\}$, and $x_{++} = \sum_{i,j=1}^{3} x_{ij}$. Expected counts are determined by

$$m_{ij} = \begin{cases} \dfrac{x_{i+}x_{+j}}{x_{++}}, & \text{if } x_{++} \neq 0, \\ 0, & \text{if } x_{++} = 0, \end{cases}$$

for all $i,\ j \in \{1,2,3\}$. The test statistic T is defined by

$$T = 2 \sum_{i,j=1}^{3} x_{ij} \log \frac{x_{ij}}{m_{ij}},$$

where only positive counts $x_{ij},\ m_{ij}$ are included. The situation, when some of the observed counts are zero, may course a decrease of the degrees of freedom in the corresponding χ^2-test. The same is true for $X-$chromosomal SNPs in males. TEST 16 uses all cases for analysis.

5.3 Selection file for case-only analysis

An example of a selection file to perform the case-only TEST 15 with INTERSNP is given below. Analogously, a selection file for the TEST 16 can be composed.

```
TPED     ./test.tped // path to the tped-File
TFAM     ./test.tfam // path to the tfam-File
LD_DISTANCE 50000000 // distance between two SNPs, in base pairs (bp)
SINGLE_MARKER   1        // Armitage's trend test
TWO_MARKER      1        // two-marker analysis: 1=yes
TEST  15  // case-only allelic interaction test
SINGLETOP       1000     // top 1000 single-marker results
PRINTTOP        100      // best 100 multimarker p-values are printed
OUTPUTNAME  ./myproject/test  // specify the output name
END
```

Chapter 6

Pathway association analysis

Pathway association analysis (PAA) tests for an excess of moderately significant SNPs in genes from a common pathway. Here, we present a Monte-Carlo simulation framework that allows to formulate the main ideas of existing PAA approaches:

- SNP ratio (SNP ratio test (O'Dushlaine et al., 2009)) (`PATHWAYTEST 1`)
- Fisher score (all significant SNPs) (`PATHWAYTEST 2`)
- Gene ratio (motivated by ALIGATOR (Holmans et al., 2009)) (`PATHWAYTEST 3`)
- Fisher Max (motivated by GenGen (Wang et al., 2007)) (`PATHWAYTEST 4`)
- Fisher MaxPlus(`PATHWAYTEST 5`)

Our paper "Integrated Genome-Wide Pathway Association Analysis with INTERSNP" (Herold et al., 2012) describing the test in more details.

In order to test for association of a pathway with a phenotype the following parameters have to be specified. See also the example selectionfile "selectionfile_S6.txt".

- `PATHWAYANALYSIS` 1
- `PATHWAYFILE` <filename>
- `SINGLETOP` <singletop>
- `PATHWAYTEST` <testnumber>
- `SIMULATION` <n>

`PATHWAYANALYSIS 1` indicates that pathway association analysis shall be carried out. The keyword `PATHWAYFILE` has to be followed by the path to the pathwayfile that shall be analyzed, for instance the KEGG pathwayfile "KEGG_2_snp_b129.txt.ann" that can be found in the download section of the homepage.

The various pathway scores are constructed from all SNP with a p-value smaller than a prespecified p-value p^*. <singletop> has to be chosen accordingly. For instance, if $p^* = 0.05$ is the desired threshold for SNPs to be included into the pathway score and if 500000 QC-SNPs are in the .tped file, then $500000 * 0.05 = 25000$ is the appropriate value for <singletop> (`SINGLETOP 25000`). Of course, the data set may contain more SNPs with a p-value smaller then 0.05, but fixing the number 25000 instead protects against inflation of type I error due to residual inflation of the single-marker test statistics. Note that all single-marker tests can be combined with the pathway analysis. In particular, quantitative traits and covariates can be handled.

6.1 Pathway scores

Currently 5 different pathway scores are available, identified by the numbers 1 to 5.

- `PATHWAYTEST 1`
  This is the "SNP Ratio" (O'Dushlaine et al., 2009), i.e. the ratio s/t, where s is a number of SNPs per pathway with a p-value among the top <singletop> p-values and t the number of SNPs in a pathway.

- `PATHWAYTEST 2`
  This is the Fisher-Score $-2\sum \log p$ that runs over all p-values that are among the <singletop> p-values. In contrast to `PATHWAYTEST 1`, the strength of association is used. An additional adjustment for residual inflation is performed and is described in (Herold et al., 2012).

- `PATHWAYTEST 3`
  This is the "Gene Ratio", i.e. s/t, where s is the number of genes with at least one p-value from <singletop> and where t is the total number of genes of the pathway. This score is motivated by the ALIGATOR software (Holmans et al., 2009).

- `PATHWAYTEST 4`
  This "Fisher-Max-Score" $-2\sum \log p$ runs over the most significant SNP from each gene. The top-SNP from a gene is only included if it is among the <singletop> p-values. This score is motivated by that of the GenGen software (Wang et al., 2007).

- `PATHWAYTEST 5`
  This score extends the score from `TEST 4` with SNPs from a gene that show independent association after conditioning on the top SNP of that gene.

- `PATHWAYTEST 6`
  This pathway score is similar to the "SNP ratio" but instead of using the results of the single-marker analysis the results of the interaction analysis are used to calculate the ratio. So the "interaction ratio" is defined as the number of significant SNP pairs per pathway divided by the number of SNP pairs per pathway. The p-value threshold for the interaction pairs can be defined with the keyword `P_INTERRATIO`. To use this test the following parameter are needed:

  - `TWO_MARKER 1`
  - `TEST 2` or another interaction test
  - `PATHWAYANALYSIS 1`
  - `PATHWAYTEST 6`
  - `SIMULATION 999`
  - `P_INTERRATIO 0.05`

6.2 Simulation

Significance of the individuals' pathways as well as after adjustment for testing multiple pathways is determined via Monte-Carlo simulations based on genome-wide permutation of case-control status or quantitative trait value. At least 3500 simulations are the minimum requirement to get experiment-wise significant results with the KEGG pathways. Rather then choosing 1000 or 10000 as the number of replicates we recommend 999 or 9999, respectively.

6.3 Selection file PAA

Here is an example of a selection file to perform the PAA with `PATHWAYTEST 1`. The selection files for `PATHWAYTEST 2,3,4,5` are analogous.

```
TPED    ./test.tped // path to the .tped file
TFAM    ./test.tfam     // path to the .tfam file
PATHWAYFILE     ./KEGG_2_snp_b129.txt   // path to the pathway file
SINGLE_MARKER   1 // Armitage's trend test
SINGLETOP   25000 // top 25000 single-marker results
PATHWAYANALYSIS  1  // pathway association analysis
PATHWAYTEST  1  // pathway analysis with Fisher score
SIMULATION  999  // number of permutations
OUTPUTNAME  ./myproject/test  // specify the output name
END
```

Selection file for the interaction-Ratio (`PATHWAYTEST 6`)

```
TPED    ./test.tped // path to the .tped file
TFAM    ./test.tfam     // path to the .tfam file
PATHWAYFILE     ./KEGG_2_snp_b129.txt   // path to the pathway file
SINGLE_MARKER   1 // Armitage's trend test
TWO_MARKER      1       // two-marker analysis: 1=yes
TEST  2  // log-linear model, test for genotypic interaction
PATHWAYANALYSIS  1  // pathway association analysis
PATHWAYTEST  6  // pathway analysis with Fisher score
P_INTERRATIO 0.05  // p-value threshold for the interaction pairs
SIMULATION  999  // number of permutations
OUTPUTNAME      ./myproject/test  // specify the output name
END
```

Chapter 7

Test for Supra-Multiplicativity of SNP Sets

This test can be used to test for deviation from multiplicativity for SNP sets of size up to 500. It is powerful when risk alleles combine in a risk fortifying way (rather than in the presence of interaction defined by alteration of risk directions in SNP combinations). The test is particular powerful in the presence of models as suggested by Zuk et al. (Zuk et al., 2012). The test is described in detail by Herold et al. (Herold et al., 2013).

These are the keywords needed to test for supra-multiplicativity of SNP sets.

```
  LIABILITY              // 1=SMT, 2=NSMT
  SNP_COVARIATES         // rs-numbers, e.g. rs11248850;
  SNP_RISK               // risk allele, e.g. A;T;T;G;
```

`LIABILITY 1` conducts a supra-multiplicativity test (SMT). Choose `LIABILITY 2` for the novel supra-multiplicativity test (NSMT), an improved version of the original SMT. By implementation of a generalized regression equation, the NSMT avoids multiple testing and provides more sophisticated modeling of amplifying interaction. The keyword `SNP_COVARIATES` specifies a list of rs numbers. These are the SNPs whose supra-multiplicativity shall be investigated. At least one SNP must be specified, the list can be of arbitrary length, but there are computational constraints for $n >> 200$.

The keyword `SNP_RISK` is optional. By default, the risk alleles are derived from the data. When there is knowledge from outside, risk alleles can directly be specified as *A*,*C*,*G*,*T*. The letter *R* indicates that the risk allele shall be derived from the data. When the keyword is used, the number of risk alleles specified has to match the number of snp covariates (`SNP_COVARIATES`) and the order of the alleles has to be consistent with the order of the SNPs. In the example `SNP_RISK A;C;R;` the risk allele for rs123 is “A”, the risk allele for rs456 is “C” and the risk allele for rs34567 is derived from the data.

7.1 Selection file and output

```
  TPED ./test.tped // path to the tped-File
  TFAM ./test.tfam// path to the tfam-File
  SINGLE_MARKER 3 // regression test
  LIABILITY 1// Test for supra-multiplicativity selected
  SNP_COVARIATES rs11248850;rs9929621;rs11645697;rs4786664;// reported SNPs
  SNP_RISK A;T;T;G;// risk allele
  OUTPUTNAME test// Specify the output name
END
```

The output is written to the file *_Liability.txt. If `LIABILITY 1` is chosen, then the decisive p-value is the total p-value which is adjusted for multiple testing. Beyond, the p-values (column p_Value) p_k refer to allele load thresholds k and are not corrected for multiple testing.

Chapter 8

Stratified analysis

To perform a stratified analysis we offer clustering and matching methods. Details on the methods can be found in (Lacour et al., 2015). The keywords to employ are

```
STRATIFY_STRUCT            // PAIR, GROUP, CLUSTER
STRATIFY_VALID             // VICINITY, CLUSTER
GROUP_TEST                 // SQUARE, LINEAR
MATCHINGCHOICE             // example chr4;chr12,12300-16000;
SIMULATION                 // integer value e.g. 9999
SINGLETOP                  // integer value e.g. 30143
PRINTTOP                   // integer value e.g. 30143
CLUSTER_COVARS             // 0=off,1=on
```

If activated by the keyword `STRATIFY_STRUCT`, the stratified analysis is be performed on any test, that involves resampling simulations of the phenotype. For resampling simulations the keyword `SIMULATION` has to be set greater than 0. Also, `SINGLETOP` and `PRINTTOP` should be set sufficiently large. The keyword `MATCHINGCHOICE` restricts population structuring to the stated genetic regions. The options for the keyword `STRATIFY_STRUCT` are

- empty (default) no matching methods are employed
- `PAIR` creates an optimal case-control matching with equal number of cases and controls. Ignores unmatched individuals. Use this only, when you know what you are doing.
- `GROUP` creates groups with at least one case and one control
- `CLUSTER` creates clusters (applicable to both CC and QT).

The keyword `STRATIFY_VALID` defines the validation method for pairwise and groupwise matching.

- `CLUSTER` perform clustering first, allows pairs/groups only within-cluster. Use this keyword if you expect large-scale population structure.
- `VICINITY` (default) checks for abundance of possible better matchings. Use this keyword if you expect small-scale population structure.

`STRATIFY_VALID` is ignored if the `CLUSTER` attribute is used for the keyword `STRATIFY_STRUCT`.

The keyword `GROUP_TEST` calls a modified version of the Cochran-Armitage-Trend (CAT) test, which is generalized for the usage of population structure. It overrides the `SINGLEMARKER` keyword.

- empty (default) `SINGLEMARKER` is used
- `SQUARE` calls the *squared Matching Cochran-Armitage-Trend*, MCAT$^{(2)}$, test. This test is based on the squared version of the CAT test. The usage of resampling simulations (keyword SIMULATION) is mandatory.

- `LINEAR` calls the *linear Matching Cochran-Armitage-Trend*, MCAT$^{(1)}$, test. This test is based on the linear version of the CAT test corrected for finite populations. The test is asymptotic: there is no need for resampling simulations, but may be performed optionally.

`CLUSTER_COVARS` produced n=#clusters covariates, and denotes peoples' cluster affiliation by 0/1. Usable only with tests that consider covariates (e.g. `SINGLEMARKER` 3).

Chapter 9

INTERSNP-RARE: Genome-Wide Rare Variant Analysis

9.1 Introduction

Important note: INTERSNP-RARE requires the preprocessor directive `#RARE` to be set to 1 in intersnp.cpp before compilation:

```
#RARE 1
```

INTERSNP-RARE is an extension build upon INTERSNP's framework for genome-wide rare variant analysis. Currently, three permutation-based rare variant testing procedures are implemented: Fisher-Rare - FR, based on the Fisher combination test (Fisher, 1925), COLLapsing test - COLL, (Li and Leal, 2008) and the Cumulative Minor Allele Test - CMAT, (Zawistowski et al., 2010). In addition, three regression tests are implemented, REG, which is the "full" regression test, and FRACREG and COLLREG, which are based on ideas laid out in (Morris and Zeggini, 2010). All tests are extended to their variable-threshold counterparts (keyword `VT`). Whether a SNP is rare is determined by the parameter to the keyword `MAFT`, which is interpreted as the threshold MAF_T for defining "rareness". MAF is calculated from the whole sample, irregardless of individuals' affection status. In the case of a zero MAF, the SNP in question is homozygous and is not considered to be rare. In the literature, common values for defining rare SNPs are often fixed at 0.01 or 0.05, though this may depend on the particular dataset.

Statistical testing of disease association is conducted on rare variants in physical proximity ("bins"). Thus, a physical position is required for each marker to be analyzed. The choice of a particular binning is not straightforward and has to be chosen taking various properties of the dataset into account. In INTERSNP-RARE, extensive support for bin assignment is provided. Prior to elaborating on the details, it is necessary to fix our terminology.

Bin: a subset of discrete SNPs in physical proximity, taken from the set of all SNPs in a study. The most simple way to define bins is to use a simple counting algorithm from genotype data. Automatic binning can be performed by number of SNPs (`BINSIZE_SNP`), by number of rare SNPs (`BINSIZE_RARE`) or by size of the region in base pairs (`BINSIZE_DIST`). Bins can be also be provided in SetID files.

Interval: a continuous region of a chromosome. Intervals are uniquely defined by a chromosome number, a start and an end position. Intervals may represent features like LD blocks or genes. Intervals are independent of any particular study data and are oblivious to any peculiarities therein. INTERSNP-RARE accepts generic interval files with one interval per row. If an interval file is provided, bins are created by projection of intervals on the genotype data.

Intervals are supplied by an external file (keyword `INTERVALFILE`). The maximal size for an interval is a whole chromosome. Pseudoautosomal regions and sex chromosomes are ideally treated

as independent chromosomes. Support for overlapping and non-overlapping bins and intervals is provided.
An overview of keywords for controlling rare variant testing are listed in the next section.

9.2 Options

Keyword	Example	Description
MAFT	0.05	Definition of the threshold MAF used to define ”rareness”
VT	1	1 for variable threshold analysis (VT)
MAF_ADJUST	1	1 (default): MAF is calculated from expected, (not observed) number of alleles. No effect if SNP missing rate=0.
BINSIZE_SNP	4000	Number SNPs per bin (rare and common)
BINSIZE_RARE	50	Number rare SNPs per bin
BINSIZE_DIST	800000	Physical distance in bp per bin
RARE_TESTS	REG;FRACREG;COLLREG	List of tests to be conducted. Default: COLL;CMAT;FR
LAMBDA_ADJUST	1	Correction for genome-wide FR testing
PRESCREEN	0.0001	Pretest for permutation-based analysis. Eliminates bins below (conservative) threshold before the first permutation
ADAPTIVE	0.001	Pretest for rare variant analysis using the Wilson confidence interval.
GENERATE_SETID	1	Skip association analysis, write modified intervals and SetID files.
SETID	Bins.SetID	File containing intervals in SetID format
INTERVALFILE	Ensembl_Genes75_hg19.txt	File containing intervals
INTERVAL_FORMAT	[Semicolon-separated list]	Formatting of the interval file
	HEAD=”1”;	Number of header lines
	SEPARATOR=”\t”;	Field separator
	COLUMNS=”5”;	Number of columns
	CHR=”1”;	Column with chromosome number
	START=”2”;	Column with interval start positions
	END=”3”;	Column of interval end positions
	FEATURE=”4”;	Column containing ”feature”

MERGE_INTERVALS	1	1 for merging of overlapping input intervals
FLANKING	1000	Flanking of intervals in bp
CONCATENATE_INTERVALS	0	Number of consecutive intervals to be concatenated
CLOSE_GAPS	1	For a non-overlapping set of intervals: assign uncovered space to intervals
MIN_RARE_IN_BIN	5	Filter bins: minimum required number of rare SNPs
MAX_RARE_IN_BIN	100	Filter bins: maximum allowed number of rare SNPs; bins will be split
VERBOSE	1 | 2	Verbosity level of output. Default: 1
WEIGHTS	1/SD | BETA;1;25; | LOGISTIC;0.07;150;	Weights of variants. Default: all weights are equal to 1.

9.3 Statistical Tests

The keyword `RARE_TESTS` allows to choose tests to be conducted from one or more of the following: FR;CMAT;COLL;REG;FRACREG;COLLREG. The list has to be separated by semicolons. By default, the three permutation tests FR;CMAT;COLL; are selected. Although it is possible to conduct all six tests in one run, the requirements for computing regression tests with permutations are usually to high to be practical, at least on genome-wide scale. We therefore recommend to conduct genome-wide regression testing analytically with a fixed MAF threshold.

9.3.1 FR

The implemented Fisher-Rare test is based on Fisher's combined probability test. The test statistic is determined by combining the N single-marker p-values of N rare variants in a particular bin

$$T_{FR} = -2\sum_{i=1}^{N} w_i \text{log}(p_i).$$

In our implementation, all weights are $w_j = 1$ by default (see :`TP` 9.3.2.1 on custom weights).

For fixed threshold analysis, the p-value is determined from the statistic by permutation of the case-control status (or quantitative trait), while keeping genotype data unchanged.

For independent p-values, the overall p-value may be determined via χ^2-distribution with $2N$ degrees of freedom. Due to linkage disequilibrium (LD), the p_i are not independent, which leads to the χ^2-distribution being an anticonservative approximation. However, the χ^2-distribution is useful for finding the optimal rare MAF by raising the threshold frequency and determining the overall p-value with the new set of rare SNPs, and taking the MAF for which the p-value is minimal. The "true" p-value is calculated by permutation (see :`TP` 9.4).

Setting the parameter of the keyword `LAMBDA_ADJUST` to 1 instructs INTERSNP to normalize singlemarker-p_i to be normalized by dividing corresponding test statistics by the inflation factor. Use this option only with a large number of SNPs, i.e. in genome-wide association studies.

9.3.2 CMAT

CMAT, the cumulative minor allele test (Zawistowski et al., 2010) utilizes the weighted sum (with weights w_j) over F minor allele frequencies of rare variants in N_A cases and N_U controls

$$T_{CMAT} = \frac{N_A + N_U}{2N_A N_U \sum_j w_j} \times \frac{(m_A M_U - m_U M_A)^2}{(m_A + m_U)(M_A + M_U)},$$

where m_A, m_U, M_A and M_U are weighted minor and major allele counts in affected and unaffected individuals

$$m_A = \sum_{i=1}^{N_A}\sum_{j=1}^{F} w_j X_{ij}$$

$$m_U = \sum_{i=1}^{N_U}\sum_{j=1}^{F} w_j Y_{ij}$$

$$M_A = \sum_{i=1}^{N_A}\sum_{j=1}^{F} w_j (2 - X_{ij})$$

$$M_U = \sum_{i=1}^{N_U}\sum_{j=1}^{F} w_j (2 - Y_{ij}).$$

Here, X_{ij} and Y_{ij} are the number of minor alleles at the jth variant position for the ith case and control, respectively.

The p-values for the CMAT test are determined by permutation. It is possible to conduct a modified CMAT test with one categorical covariate (Zawistowski et al., 2010).

9.3.2.1 Weights

The weights (default: $w_i = 1$) can be used in non-collapsing permutation tests to weight alleles by a pre-determined scheme that mirrors our estimation of their risk.

Currently, three weighting schemes are supported (see also Wu et al. (2011) and references therein).

1/SD. The weights are reciprocal to the standard deviation of allele counts

$$w_i = \frac{1}{\sqrt{MAF_i(1 - MAF_i)}}.$$

BETA;a;b;. The weights are proportional to the Beta distribution (default: $a = 1$, $b = 25$)

$$w_i = MAF_i^{a-1} \left(1 - MAF_i\right)^{b-1}.$$

LOGISTIC;a;b;. Instead of strongly upweighting rare variants, logistic weights are rather a "soft threshold" alternative to a fixed MAF threshold (default: $a = 0.07$, $b = 150$)

$$w_i = \frac{1}{1 + e^{(MAF_i - a)b}}.$$

The weights of rare variants are approximately 1 and for common variants 0.

9.3.3 COLL

The collapsing test COLL (Li and Leal, 2008) dichotomizes individuals by the presence of *any* rare variant in a given bin

$$T_{COLL} = \frac{(N_A + N_U) \times (XN_U - YN_A)^2}{N_A N_U (X + Y)(N_A + N_U - X - Y)},$$

where X is the number of cases carrying a rare variant:

$$X = \sum_{i=1}^{N_A} X_i.$$

X_i are indicator variables for the ith case:

$$X_i = \begin{cases} 1 \text{ if } w_j X_{ij} > 0 \text{ for any } 1 \leq j \leq F \\ 0 \text{ otherwise,} \end{cases}$$

and analogous definitions of Y, Y_i for controls.
Alternatively, the test statistic can be expressed as

$$T_{COLL} = \frac{(N_A + N_U)\left(N_A^C N_U - N_U^C N_A\right)^2}{N_A N_U \left(N_A^C + N_U^C\right)\left(N_A + N_U - N_A^C - N_U^C\right)}$$

The p-values of the COLL test are determined by the χ^2-distribution with one degree of freedom in case of a fixed threshold and by permutation in case of a variable threshold.

9.3.3.1 Odds Ratios

Both variable and fixed threshold analyses of COLL and CMAT return ”odds ratios” based on their respective contingency tables. They are useful for determining the direction of effect.

9.3.4 REG

REG is the full regression test conducted on the set of rare variants. Logistic and linear regression are used for quantitative and binary phenotypes, respectively. In the logistic regression, the probability to be a case is modelled as:

$$\text{logit}(p) = \beta^T b + \gamma^T g, \tag{9.1}$$

for every individual, with $\text{logit}(p) = \log\left(p(1-p)^{-1}\right)$, and β and γ vectors of regression coefficients to be estimated. The covariate vector $b = \{1, b_1, ..., b_k\}$ encodes covariate quantities of the individual to be included in the model, while g is the vector of genotypes of the individual. In the log-additive model, the genotypes are coded as -1, 0 and 1 for the three cases when the major allele is homozygous, the genotype is heterozygous or minor allele is homozygous. The vector of covariate coefficients β consists of the intercept parameter β_0 and k covariate coefficients $\{\beta_1, ..., \beta_k\}$, while γ estimates the risk contribution of each genotype. Maximum likelihood estimation of the null model with $H_0 : \gamma = 0$ together with the model where β is unconstrained yields a likelihood ratio test with m degrees of freedom.

REG and FR are both oblivious to the direction of effect of individual variants. This property sets them apart from other tests which are commonly classified as ”burden” tests. While burden tests are in general expected to be more powerful when most of rare alleles in a bin are associated with either affection status, they lose power when both protective and damaging effects are present. Two further burden regression tests, FRACREG and COLLREG, were described in Morris and Zeggini (2010) for the linear regression case. The main idea is to combine rare genotypes from a bin into one variable and to fit a single regression coefficient in a regression model. Therefore, FRACREG and COLLREG incorporate to an extent the idea of CMAT and COLL into a regression framework.

9.3.5 FRACREG

FRACREG estimates the coefficient λ with the trait modelled as:

$$\text{logit}(p) = \beta^T b + \lambda \frac{r}{n} \tag{9.2}$$

Here, r is the count of genotypes of an individual containing at least one minor allele. The ratio $\frac{r}{n}$ is the accumulated genetic burden with equal contribution from each genotype containing one or two minor alleles. The statistical test is conducted by comparing the null model $H_0 : \lambda = 0$ to $H_1 : \lambda \neq 0$.

9.3.6 COLLREG

The COLLREG is identical to FRACREG, up to the burden function. COLLREG is a regression model over collapsed variants. As in FRACREG, one coefficient λ is estimated. An individual is assumed to carry the full genetic burden when at least one rare allele is present:

$$\text{logit}(p) = \beta^T b + \lambda I(g) \tag{9.3}$$

The indicator function $I(g)$ equals 1 if an individual has at least one rare allele in the region of interest and 0 otherwise. The COLL and the COLLREG tests are members of the family of collapsing methods.

9.4 Variable Threshold

The Variable Threshold (VT) analysis is activated using the keyword `VT`. The extension to the variable threshold model for COLL and CMAT tests is motivated by the principle described in (Price et al., 2010).

COLL, CMAT: with N_{SIM} permutations, calculate T^j_{max} for each value of $MAF^j_i \leq MAF_T$ contained for each permutation j. The proportion of those T^j_{max} that are larger than the non-permuted T^0_{max} is used to determine the corresponding p-value:

$$p_{COLL,CMAT} = \frac{\#\{T'^j_{max} > T^0_{max}\}}{N_{SIM}}. \tag{9.4}$$

FR: The VT approach requires a modification in case of FR test. Since the test statistic depends on a different number of p-values at different MAF_i levels, they can not be directly compared. $T_{FR,MAFi}$ is also monotonically increasing with the value of MAF. The test statistic is thus not comparable in variable threshold mode. A multiparametric function is required to map different T_{FR} on an interval taking the number of variants into account, that allows a comparison operator.
One of the suitable functions that maps T_{FR} and allows comparison is the χ^2 distribution with $2m$ degrees of freedom. Although due to LD the distribution is, in general, anticonservative, it can be used since the absolute p_{Est} enter the calculation for being used for comparison ("<" and ">"), at different MAF levels only.

T'_{FR}=p (the χ^2-distribution with $2N$df) instead of T_{FR} is used to obtain optimal threshold. The p-value is determined by calculating the fraction of smaller minimal test statistics T'_{FR} across permutations and MAF levels:

$$p_{FR} = \frac{\#\{T'^j_{min} < T^0_{min}\}}{N_{SIM}} \tag{9.5}$$

The FR method is therefore similar to COLL and CMAT VT-methods, except that maximal test statistics are compared at different MAF levels. Small test statistics are considered to be more improbable, which leads to a change of sign in Eq. (equation9.5) .

In a straightforward extension, the VT method for regression tests utilizes the regression p-values at each MAF level:

$$p_{REG,FRACREG,COLLREG} = \frac{\#\{p^j_{min} < p^0_{min}\}}{N_{SIM}} \tag{9.6}$$

For rare regression tests, the VT method is, in general, not recommended for genome-wide analysis due to relatively long running times.

9.5 Rare Pretests

In a permutation framework, computational time and hardware requirements can be significantly reduced by conducting the full number of permutations only for bins that have a chance to come close to the significance level. This can achieved by two implemented methods.

`PRESCREEN` takes place before any permutations are conducted. We use the fact that the p-values obtained from test statistics are either asymptotically exact or anticonservative due to LD. Choosing a suitable passing threshold eliminates the majority of tests to be conducted.

To reduce computational time even further, we offer an implementation of an adaptive pretest. The keyword `ADAPTIVE` takes a floating point number equal to the threshold p-value for passing the pretest. Confidence intervals for permutation p-values are calculated using the Wilson score method (Wilson, 1927):

$$\text{CI}^{\pm} = \left(1 + \frac{z^2_{\alpha/2}}{n}\right)^{-1} \left(p + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{1}{n}p\,(1-p) + \frac{z^2_{\alpha/2}}{4n^2}}\right) \tag{9.7}$$

where $z_{\alpha/2} = 1.96$ for the 95% confidence interval. The Wilson method performs well even for small values and is known to be more accurate for binomial proportions than the commonly used normal approximation interval (Wald interval) (Agresti and Coull, 1998).

The passing condition satisfied if the lower bound of the confidence interval is below the value passed to the `ADAPTIVE` keyword. After each simulation in which the permuted p-value is smaller (or test statistic larger) than the original p-value (test statistic), each window and each individual test is checked for the passing condition.

If the passing conditions are satisfied after the whole set of permutations, the p-value will be marked with a plus sign in the output file.

9.6 Binning Methods

9.6.1 Count-based Binning

The most simple binning methods determine bins directly from the data, by number of contained variants (`BINSIZE_SNP`), rare variants (`BINSIZE_RARE`) or by size of the region in base pairs (`BINSIZE_DIST`).

9.6.2 User-Supplied Intervals

In addition to algorithmic binning procedures, extensive functionality for generating and modifying bins created from user-supplied intervals is provided. Bins based on functional classes or statistical characteristics can be generated from a multitude of sources: LD intervals, conserved elements, gene annotation files and more. This set of features is also helpful for validating bins in a different dataset.

Intervals are supplied by the user in a file, one interval per line, by keyword `INTERVALFILE`. Obviously, `INTERVALFILE` and the three algorithmic binning methods are mutually exclusive.

We provide an interval file (Ensembl_Genes75_hg19.txt) based on a list of coding genes imported from Ensembl using BioMart (http://www.ensembl.org/biomart/martview). The file was created using attributes ”Chromosome Name”, ”Gene Start (bp)”, ”Gene End (bp)”, ”Associated Gene Name”, ”Description”, and exported in the *.tsv format. The resulting file was sorted with respect to chromosome name and gene start.

The formatting of the interval file is defined in a string `INTERVALFILE_FORMAT`. Possible arguments are `HEAD`, `COLUMNS`, `CHR`, `START` and `END`, which give the overall number of lines in the header, number of columns, and the number columns containing start and end positions, chromosome names and possibly a column containing a descriptor of the feature in the file. Columns are counted starting from one. INTERSNP accepts different chromosome naming conventions, for instance ”CHR26”, ”chrMT”, ”M” will be recognized as chromosome 26.

Two additional fields from an interval file can be specified by the user, first of which is `FEATURE`. A feature is handled as a character string and serves the purpose of tracking intervals more easily. A feature might be a gene name or name of an LD block. The feature column is carried all the way through the calculation and will be appended to the output file. In case that intervals undergo modification, the feature column is modified as well. For instance, when consecutive fields are merged or concatenated, their respective features are put in one field. In addition, `SEP` defines the field separator. Currently, tab (`"\t"`), space (`"SPACE"`) and comma (`","`) are supported as field separators.

The default interval file format string is compatible with the provided gene interval file Ensembl_Genes75_hg19.txt:

```
HEAD=”0”;SEPARATOR=”\t”;COLUMNS=”4”;CHR=”1”;START=”2”;END=”3”
```

9.6.3 Interval Modification

9.6.3.1 Merging of Input Intervals

In case that some of the intervals from the input file are overlapping, setting `MERGE_INTERVALS` to 1 will instruct INTERSNP to merge them. In case that intervals have identical coordinates, the default behaviour is to dismiss all but the first of these intervals.

9.6.3.2 Flanking

Providing a parameter to `FLANKING` will instruct INTERSNP to extend intervals in both directions by that many base pairs.

9.6.3.3 Close Gaps

If the intervals do not cover the genome in a continuous way, some of the markers in a dataset may be found in the space between intervals and will be dismissed from rare variant analysis. If the provided intervals are non-overlapping or have been merged, this problem can be countered by setting `CLOSE_GAPS` to 1. The space in between intervals will be divided in half and each part will be assigned to the nearest interval.

9.6.3.4 Concatenating Intervals

In case that the provided intervals are found to be too small, one can set `CONCATENATE_INTERVALS` to the number of consecutive intervals to be concatenated into one interval.

9.6.4 Size of Bins: Quality Control

9.6.4.1 Filter Small Bins

Some bins may cover only few rare variants, especially when intervals are small and data sparse. In that case, INTERSNP will unnecessarily process useless bins. In addition, the number of bins will reduce power due to correction for multiple testing. This can be countered by setting `MIN_RARE_IN_BIN` to the minimum number of rare variants per bin. Bins not satisfying this condition will be simply dismissed.

9.6.4.2 Maximum Number of Rare Variants

Some bins may cover hundreds or thousands of rare variants. It is advised to set a maximal number of rare variants per bin using the keyword `MAX_RARE_IN_BIN` to prevent disproportionately high computational burden, in particular when using variable-threshold models.

Bins that are found to contain more rare SNPs than the given limit will be split in bins containing the same number of rare variants, if possible.

9.7 Imputation of Minor Allele Frequencies for Missing Genotypes

`MAF_ADJUST` is a keyword that controls a simple recalculation of MAFs in SNPs with a nonzero missing rate. The default value of the parameter is 1. The number of minor allele thresholds under consideration may be large if missing data is present. With stringent missing rate quality control, we use the approximation for the minor allele frequency:

$$MAF = \frac{\text{\#(Minor alleles)}}{2(N-M)} \approx \frac{\text{\#(Minor alleles)}}{2N} \tag{9.8}$$

where N is the number of individuals and M the number of missing genotypes. Note that this assignment is made only to reduce the number of MAF thresholds that are investigated, but that it does not affect the computation of the test statistics we use, for which we always use the actual genotype information.

9.8 Output

9.8.1 Association Testing Results: *Rare.txt

Test results and additional information is written to the *Rare.txt file, which is a fully compatible interval file.

- BinNr: bins are counted according to their start positions (with `VERBOSE 2`)
- SNPs: the number of variants in bin, rare or common (with `VERBOSE 2`)
- chr: chromosome number
- start, end: bin boundaries in base pair positions
- from_SNP, to_SNP: bin boundaries in variant counts (with `VERBOSE 2`)
- from_rare_SNP_Nr, to_rare_SNP_Nr: bin boundaries in rare variant counts (with `VERBOSE 2`)
- nRV: the number of rare variants with MAF$\leq$MAF$_T$
- rarefreq[FR|CMAT|COLL|REG|FRACREG|COLLREG]: the optimal frequency from VT analysis
- nRare[FR|CMAT|COLL|REG|FRACREG|COLLREG]: the number of variants with $MAF \leq MAF_T$
- p_[FR|CMAT|COLL|REG|FRACREG|COLLREG]: the p-value
- p_[FR|CMAT|COLL|REG|FRACREG|COLLREG]_Corr: the Bonferroni-corrected p-value (with `VERBOSE 2`)
- CI_[FR|CMAT|COLL|REG|FRACREG|COLLREG]: Wilson confidence interval
- nPerm_[FR|CMAT|COLL|REG|FRACREG|COLLREG] if `ADAPTIVE` is set to 1, the number of permutation conducted for the respective test.
- OR_[COLL|CMAT]: ”odds ratios” based on contingency tables
- FEATURE: optional feature column from interval file

9.8.2 Modified Interval File: *_modified.txt

The quality-controlled, modified, sorted intervalfile (`INTERVALFILE`) is written to a new interval file in the working directory. If `GENERATE_SETID` was selected, no analysis will not be performed. Instead, a SetID file will be generated that can be used for subsequent analysis with INTERSNP. The SetID file has the advantage that only SNPs relevant for the analysis will be loaded into memory, resulting in an overall speedup of the analysis.

Bibliography

Agresti, A., and Coull, B.A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. The American Statistician, 52(2), 119-126.

Armitage P. (1955) Tests for linear trends in proportions and frequencies. Biometrics 11(3): 375–386.

Becker T., Herold C., Meesters C., Mattheisen M., Baur M.P. (2011) Significance levels in genome-wide interaction analysis (GWIA). Ann. Hum. Genet. 75(1): 29–35.

Bishop Y.M., Fienberg S.E., Holland P.W. (2007) Discrete Multivariate Analysis. Theory and Application. Springer, Berlin.

Clayton D. (2008) Testing for association on the X chromosome. Biostatistics 9(4): 593–600.

Cordell H. and Clayton D. (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(1): 124–141.

Cordell H.J. (2009) Detecting gene-gene interactions that underlie human diseases. Nat. Rev. Genet. 10(6): 392–404.

Fisher, R.A. (1925). Statistical Methods for Research Workers. Oliver and Boyd (Edinburgh). ISBN 0-05-002170-2.

Gatto N.M., Campbell U.B., Rundle A.G., Ahsan H. (2004) Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. Int. J. Epidemiol. 33(5): 1014–1024.

Ge Y., Dudoit S., Speed T.P. (2003) Resampling-based multiple testing for microarray data analysis. Test 12(1): 1–77.

Herold C., Steffens M., Brockschmidt F.F., Baur M.P., Becker T. (2009) INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics 25(24): 3275–3281.

Herold C., Mattheisen M., Lacour A., Vaitsiakhovich T., Angisch M., Drichel D., Becker T. (2012) Integrated genome-wide pathway association analysis with INTERSNP. Hum. Hered. 73(2): 63–72.

Herold, C., Ramirez, A., Drichel, D., Lacour, A., Vaitsiakhovich, T., Nöthen, M.M., Jessen, F., Maier, W. and Becker, T. (2013) A one-degree-of-freedom test for supra-multiplicativity of SNP effects. PLos ONE. 2013 Oct 30; 8(10):e78038.

Holmans P., Green E.K., Pahwa J.S., Ferreira M.A., Purcell S.M., Sklar P., Wellcome Trust Case-Control Consortium, Owen M.J., O’Donovan M.C., Craddock N. (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am. J. Hum. Genet. 85(1): 13–24.

Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M., Hirakawa M. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34: 354–357.

Lacour A., Schüller, V., Drichel, D., Herold, C., Jessen, F., Leber, M., Nöthen, M.M., Ramirez, A., Vaotsiakhovich, T., Becker, T. (2015) Novel genetic matching methods for handling population stratification in genome-wide association studies. BMC Bioinformatics. 2015 Mar 14;16:84.

Li B., Leal S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 83(3): 311–321.

Morris, A.P. and Zeggini, E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. 34(2): 188–193.

O'Dushlaine C., Kenny E., Heron E.A., Segurado R., Gill M., Morris D.W., Corvin A. (2009) The SNP ratio test: pathway analysis of genome-wide association datasets. Bioinformatics 25(24): 2762–2763.

Piegorsch W.W., Weinberg C.R., Taylor J.A. (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat. Med. 13(2): 153–162.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J., Sham P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81(3): 559–575.

Price, AL., Kryukov, GV., de Bakker, PIW., Purcell, SM., Staples, J., Wei, LJ., Sunyaev, SR. (2010) Pooled Association Tests for Rare Variants in Exon-Resequencing Studies Am. J. Hum. Genet. 86(6) 832-838

Wang K., Li M., Bucan M. (2007) Pathway-based approaches for analysis of genomewide association studies. Am. J. Hum. Genet. 81(6): 1278–1283.

Wilson, E B. (1927) Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association 22.158:209-212.

Wu M.C., Lee S., Cai T., Li Y., Boehnke M., Lin X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89:82–93.

Yang Q., Khoury M.J., Sun F., Flanders W.D. (1999) Case-only design to measure gene-gene interaction. Epidemiology 10(2): 167–170.

Zawistowski M., Gopalakrishnan S., Ding J., Li Y., Grimm S., Zöllner S. (2010) Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am. J. Hum. Genet. 87(5): 604–617.

Zuk O., Hechter E., Sunyaev S.R., Lander E.S. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci USA 109: 1193-1198.