ROADTRIPS Software Documentation

Version 1.2

Timothy Thornton[1] and Mary Sara McPeek[2,3]

Department of Biostatistics[1]
University of Washington

Departments of Statistics[2] and Human Genetics[3]
The University of Chicago

ROADTRIPS (RObust Association-Detection Test for Related Individuals with Population Substructure)

A C program for case-control association testing that allows for partially or completely unknown population and pedigree structure

Homepage: http://galton.uchicago.edu/∼mcpeek/software/index.html

Release 1.2 June 24, 2010

==========

License

We request that use of this software be cited in publications as follows:

> Thornton T., McPeek M. S. (2010) "ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure" American Journal of Human Genetics, vol 86, pp. 172-184.

To contact the first author:

Timothy A. Thornton
Department of Biostatistics
University of Washington
Health Sciences Building F-600
Box 357232
Seattle, WA 98195-7232

email: tathornt@u.washington.edu

Contents

1 Overview of ROADTRIPS

ROADTRIPS is a C program that performs single-SNP, case-control association testing in samples with partially or completely unknown population and pedigree structure. The main reference for this program is **Thornton and McPeek (2010) AJHG 86: 172-184**. ROADTRIPS uses an empirical covariance matrix calculated from genomewide SNP data to correct for unknown population and pedigree structure, while maintaining high power by taking advantage of known pedigree information when it is available. The program is applicable to association studies with completely general combinations of related and unrelated individuals. Analysis can be performed genomewide (currently just for autosomes).

ROADTRIPS is suitable for applications such as

(1) correcting for possible population structure and/or misspecified relationships in the context of case-control association testing in samples of unrelated individuals and/or related individuals with well-characterized pedigrees

(2) case-control association testing in samples from isolated populations for which pedigree information is limited or unavailable

For each marker, the ROADTRIPS program computes test statistics and p-values for 3 different tests for association: the RM test, which is the ROADTRIPS extension of the M_{QLS} test of Thornton and McPeek (2007), and the $R\chi$ and RW tests, which are the ROADTRIPS extensions of the corrected χ^2 and W_{QLS} tests, respectively, of Bourgain et al. (2003). For each test, the p-value is calculated based on a χ^2_1 asymptotic null distribution. When sampled individuals are related and partial or complete pedigree information is available, we recommend using the RM test, because our simulations indicate that it has higher power than the others in that context (Thornton and McPeek 2010). For more information on the 3 test statistics and recommendations for their use, see **Chapter 2**.

Additional features of the ROADTRIPS tests include:

(1) Appropriate handling of missing genotype data, so that the tests are valid at each SNP.

(2) Pedigree information is not required, but when partial or complete pedigree information is available, the RM and RW tests can improve power by using this information.

(3) With the RM test, two different types of controls, unaffected controls and controls of unknown phenotype (e.g., general population controls) can be incorporated into the same analysis analysis (see **Chapter 2** for details).

(4) When a phenotyped individual has missing genotype at the SNP being tested, the individual's phenotype is still informative to the analysis if he or she has a sampled relative who is genotyped at the SNP. The RM test can incorporate this information.

2 Description of the ROADTRIPS Statistics

The ROADTRIPS program computes test statistics and p-values for 3 different tests of association: the RM test, the $R\chi$ test, and the RW test. All three test statistics are of the form

$$\left(\mathbf{V}^T\mathbf{Y}\right)^2 \Big/ (\hat{\sigma}_1^2 \mathbf{V}^T \hat{\mathbf{\Psi}} \mathbf{V}).$$

Remarks on this formula:

- $\mathbf{Y}$ is the vector of genotype data, with ith element $Y_i = .5\times$(the number of alleles of type 1 in individual i)
- $\mathbf{V}$ is a weight vector that is different for each statistic
- $\hat{\mathbf{\Psi}}$ is a structure matrix that ROADTRIPS estimates from genome-screen data on autosomes, where this genome-screen data must be provided as input. The formula for $\hat{\mathbf{\Psi}}$ is given in Equation (12) of Thornton and McPeek (2010). In simulations we obtained excellent results for ROADTRIPS with $\hat{\mathbf{\Psi}}$ estimated from 10^5 SNPs genomewide. However if SNPs are available from only a handful of regions, instead of genomewide, then $\hat{\mathbf{\Psi}}$ may not be an accurate estimate, and it may not be advisable to analyze the data with ROADTRIPS in that case. (In that case, try MQLS instead, if you have related individuals with known pedigrees.)
- $\hat{\sigma}_1^2 = .5\bar{Y}(1-\bar{Y})$ is used in the software.

Following is a description of the three tests:

(1) **The RM test** The RM test is the one that was most powerful in our simulation studies. It has a particular advantage over the other statistics when sampled individuals are related and partial or complete pedigree information is available. It is also the only one of the three statistics that allows a distinction to be made between unaffected controls and unknown controls (but note that this distinction is important only if you plan to incorporate both types of controls into the same analysis). If there is no pedigree information available and only one type of control is available, then the RM, $R\chi$ and RW statistics should all be equal.

The weight vector $\mathbf{V}$ for the RM test is given in Table 1 of Thornton and McPeek (2010). It depends on (i) the working kinship matrix $\mathbf{\Phi}$, (ii) phenotype data coded as affected, unaffected or unknown, and (iii) prevalence k. Following are specific suggestions about how to specify these when you run the program:

- **working kinship matrix $\mathbf{\Phi}$**: Pedigree information is not required for ROADTRIPS, but when partial or complete pedigree information is available, the RM and RW tests can improve power by using this information. Regardless of whether

or not pedigree information is available, ROADTRIPS requires the information of a working kinship matrix. When there is no pedigree information available, or when individuals are assumed to be outbred and unrelated, the working kinship matrix $\mathbf{\Phi}$ should be specified as the identity matrix. (See **Chapter 4** for specific input instructions.) In contrast, with well-characterized pedigrees, $\mathbf{\Phi}$ would have (i, j)th element equal to $2\phi_{ij}$ for $i \neq j$ and $1 + h_i$ for $i = j$, where ϕ_{ij} is the kinship coefficient between individuals i and j and h_i is the inbreeding coefficient for individual i. When there is partial pedigree information available, but it is believed to be incomplete, then the matrix $\mathbf{\Phi}$ corresponding to a putative set of relationships could be used. In this case, the $\mathbf{\Phi}$ used must be consistent with some possible pedigree. (Otherwise $\mathbf{\Phi}$ may not be positive definite, which could cause numerical problems for the algorithm.) The pedigree from which the working kinship matrix $\mathbf{\Phi}$ is derived does not have to be the true one, but the power is expected to be higher if it is true or close to true. Note that the estimated structure matrix $\hat{\mathbf{\Psi}}$ is used in the variance calculation to account for structure that may not be captured by the working kinship matrix $\mathbf{\Phi}$.

- **phenotype data**: The *RM* test allows three possible values for an individual's phenotype: "affected," "unaffected," and "unknown," where the label "unknown" is used to represent unphenotyped individuals, e.g. general population controls, or individuals who are deemed too young to have developed an age-related trait such as Alzheimer's, whereas the label "unaffected" is reserved for true unaffecteds. As they have different expected frequencies of predisposing alleles, the two types of controls are treated differently in the *RM* analysis. This is the default setting of ROADTRIPS. Alternatively, if one uses the flag **-u**, then individuals of unknown phenotype will be dropped from the analysis.
- **prevalence** k: To calculate the *RM* statistic, an estimate, k, of the prevalence of the trait in a suitable reference population must be specified by the user. This value has no effect on the calculation unless the two different types of controls (unaffected and unknown phenotype) are both used in the same analysis. In that case, k is used as the basis for assigning different weights to these two types of controls. The value k should not be the prevalence in the case-control sample (assuming there is phenotype-based ascertainment), but rather should be the "general population" prevalence for an appropriate reference population. We emphasize that the test will be valid regardless of the value of k used. However, an accurate prevalence value should provide better power. We recommend using an estimate from previous studies or registry data from a suitable reference population.

When a phenotyped individual i has missing genotype at the SNP being tested, the individual's phenotype is still informative to the analysis if he or she has a sampled relative j who is genotyped at the SNP. The *RM* test can incorporate this information provided that (i) the working kinship matrix $\mathbf{\Phi}$ specifies a nonzero kinship coefficient between individuals i and j and (ii) i's phenotype is either "affected" or "unaffected." (If i's phenotype is "unknown" and i's genotype is missing at a given SNP, then i

will not make a contribution to the analysis for that SNP.) The default setting in ROADTRIPS is for the RM test to make use of the information of i's phenotype. Alternatively, when the flag **-m** is used, individuals with missing genotype at a SNP will be excluded from making any contribution to the analysis at that SNP.

(2) **The Rχ test** The $R\chi$ statistic is a version of the standard Pearson χ^2 test statistic that is corrected (by means of $\hat{\mathbf{\Psi}}$) for the presence of population and pedigree structure. $R\chi$ is similar in spirit to genomic control, but $R\chi$ performs better than genomic control in the case when different SNPs have different rates of genotyping error. In that case, genomic control may not be correctly calibrated, while $R\chi$ maintains correct type 1 error, because $R\chi$ applies a different correction to each SNP depending on its pattern of missing genotypes, whereas genomic control applies the same correction to all SNPs.

The $R\chi$ test ignores the information of the working kinship matrix $\mathbf{\Phi}$, and when partial or complete pedigree information is available, RM generally has higher power than $R\chi$ because it is able to use the information in $\mathbf{\Phi}$. $R\chi$ also allows only one type of control, so if one wants to incorporate both unaffected controls and controls of unknown phenotype in the same analysis and make an appropriate distinction between them in the analysis, then one should use RM. If there is no pedigree information available and only one type of control is available, then the RM, $R\chi$ and RW statistics should all be the same.

The weight vector $\mathbf{V}$ for $R\chi$ is given in Table 1 of Thornton and McPeek (2010). It is a function of the case-control indicator vector $\mathbf{1}_c$, having ith element equal to 1 if individual i is a case and 0 if individual i is a control. In contrast, the ROADTRIPS software specifies that individuals' phenotypes can be coded as "affected", "unaffected" or "unknown." To define the vector $\mathbf{1}_c$, the default option in ROADTRIPS is to combine the two categories "unaffected" and "unknown" into a single "control" category. Alternatively, when the flag **-u** is used, the individuals of unknown phenotype will be dropped from the analysis (for all 3 test statistics) and only the "unaffected" individuals will be used as controls when constructing the vector $\mathbf{1}_c$. If there is only one type of control in the data (i.e. either all controls are unaffected or all controls are of unknown phenotype), then the $R\chi$ test will handle the controls appropriately. However, if you want to include both types of controls in the same analysis, the $R\chi$ will not make a distinction between the two types, so RM would be preferable in that case.

(3) **The RW test** The RW test has not been previously introduced in the published literature (to our knowledge), but it is similar to the other tests introduced in Thornton and McPeek (2010). It is an extension of the W_{QLS} test statistic of Bourgain et al. (2003) to allow for possible population structure and/or misspecified relationships. The weight vector $\mathbf{V}$ for the RW test is given by

$$\mathbf{V} = \mathbf{\Phi}^{-1}\mathbf{1}_c - \mathbf{1}_c^T\mathbf{\Phi}^{-1}\mathbf{1}(\mathbf{1}^T\mathbf{\Phi}^{-1}\mathbf{1})^{-1}\mathbf{\Phi}^{-1}\mathbf{1},$$

where $\mathbf{\Phi}$ is the working kinship matrix described above in **The RM test** subsection, and $\mathbf{1}_c$ is the case-control indicator described above in **The Rχ test** subsection. Here $\mathbf{1}$ is a vector whose entries are all equal to 1.

Similarly to $R\chi$, the RW test allows only one type of control, as specified by the vector $\mathbf{1}_c$. Similarly to RM, the RW test can make use of partial or complete pedigree data that is specified by the working kinship matrix $\mathbf{\Phi}$. In our simulation studies, when there is pedigree information available, the RM test is generally more powerful than the RW test, but for some genetic models (e.g. a rare, fully-penetrant, dominant model), the RW might be expected to have more power. We have chosen to include the RW in the ROADTRIPS software because potential users have expressed interest in having it. If there is no pedigree information available, then the RW and $R\chi$ statistics will be equal, and RM will also be equal to RW and $R\chi$ in this setting if there is only one type of control in the data.

Summary of Recommendations Regarding the 3 Tests: When there is available pedigree information for the sampled individuals, we recommend using the RM test. We expect that RM will generally have more power than both RW and $R\chi$ in the setting of complex trait mapping in samples with related individuals. If there is no available pedigree information, RW and $R\chi$ will be equal, and RM will be equal to RW and $R\chi$ in this setting whenever there is only one type of control in the study, e.g., all of the controls are unaffected or all of the controls have unknown phenotype. If there are two types of controls in the study and there is no known pedigree information, we recommend using the RM test over the $R\chi$ and RW tests. $R\chi$ and RW do not make a distinction between the two control types, while RM treats the two types of controls differently, and as a result, power may be improved by use of the RM test in this setting.

3 Running ROADTRIPS

Installation instructions:

1. Download the ROADTRIPS package. This package contains documentation, source code, example input and output files, and a precompiled executable for Linux platforms.

2. Read the entire documentation (this document) carefully to understand the purpose of this program and how it works.

3. Edit the Makefile as necessary according to the instructions in the Makefile. You should only need to make sure that the correct compiler and compiler options for your machine are chosen.

4. Type "make". This will build an executable program called "ROADTRIPS".

5. To run the executable program ROADTRIPS:

   First, prepare the input files, e.g., phenofile, genofile, pedinfo, prevalence (see **Chapter 4** for more details).

   Then, to run ROADTRIPS with the default input filenames and settings, one need only type

./ROADTRIPS

Alternatively, to change input filenames or settings, use flags in the command line. The following flags are available:

./ROADTRIPS -p phenofile -g genofile -k pedinfo -r prevalence -n snpnames -u -m

-p phenofile Allows the user to specify the name of the phenotype data input file, which also includes family ID numbers and individual ID numbers. The filename defaults to "phenofile" if this flag is not used. To specify a different filename, replace "phenofile" with the appropriate filename.

-g genofile Allows the user to specify the name of the SNP genotype data input file. Filename defaults to "genofile".

-k pedinfo Allows the user to specify the name of the pedigree information input file which contains the working kinship and inbreeding coefficients for the sampled individuals (where these can all be set to zero if there is no pedigree information available). Filename defaults to "pedinfo".

-r prevalence Allows the user to specify the name of the prevalence input file, which contains an estimate of the prevalence of the binary trait from a suitable reference population. Filename defaults to "prevalence".

-n snpnames Allows the user to specify a file that contains the names (or rs numbers) for all of the SNPs. If this flag is not used, a default name will be given to each SNP. The default name for the ith SNP is "SNP_i", e.g., the 20th SNP in the genotype input file will be given a default name of "SNP_20."

-u Allows the user to exclude individuals with unknown phenotype from the analysis for all three test statistics. All individuals will be included in the analysis if this option is not used.

-m Allows the user to specify that only individuals who have non-missing genotypes at a marker will be included in calculating the RM statistic at that SNP, i.e., phenotype information for individuals with missing genotype data at a SNP will not be used. If this option is not used, the RM statistic will incorporate phenotype information for individuals with missing genotype data at a SNP being tested, provided that those individuals have a sampled relative who is genotyped at the marker.

6. You can test the executable program ROADTRIPS by running it with the sample input files: phenofile, genofile, pedinfo, and prevalence. You can then compare the resulting output, which will be printed to the files ROADTRIPStest.out, ROADTRIPStest.top and ROADTRIPStest.pvalues, with the correct output provided in the sample output files ROADTRIPStest.out.ex, ROADTRIPStest.top.ex, and ROADTRIPStest.pvalues.ex, respectively.

7. The program stops if any errors are detected in the format of the input files.

4 Input

Preface: How to Assign Family ID Numbers When Pedigree Information is Not Available

ROADTRIPS does not require that pedigree information be available, but when partial or complete pedigree information is available, the *RM* and *RW* tests can improve power by using this information. To accommodate this in a computationally efficient way, ROADTRIPS requires that the individuals in the sample be organized into families. Furthermore, ROADTRIPS requires working inbreeding coefficients for all individuals and working kinship coefficients for every pair of individuals in the same family, where these can all be set to zero if no pedigree information is available.

The idea behind working kinship and inbreeding coefficients is that if you have pedigree information that is at least fairly accurate, you can obtain working kinship and inbreeding coefficients based on this pedigree information (e.g. using the program **KinInbcoef**), and this information can be used to improve power of the tests. Then the empirical covariance matrix that ROADTRIPS estimates from genome-screen data is used to correct for additional population structure or misspecified relationships not accounted for in the working kinship and inbreeding coefficients.

If you have pedigrees for your sample, then individuals in the same pedigree should be assigned the same family ID number, and completely separate pedigrees should have different family ID numbers. (Note: assigning the same family ID number to separate pedigrees can cause the program to run much more slowly than it would if you assigned them different family ID numbers.) Kinship and inbreeding coefficients calculated from these pedigrees should be input as the working kinship and inbreeding coefficients.

If you have unrelated individuals or if you have no pedigree information at all, then the most computationally efficient approach is to assign each individual their own unique family ID number (note that the family ID numbers must be consecutive positive integers starting from 1). Then you can input zero for all the working inbreeding coefficients, and there is no need to input any working kinship coefficients. (A less computationally efficient approach would be to include some of these individuals in the same family, in which case you would also have to specify that all their working kinship coefficients are zero.)

If you want to enter some nonzero working kinship coefficients, but are uncertain of the true pedigree, then you can do that provided:

1. The working kinship and inbreeding coefficients you input must all be consistent with some possible pedigree. (Otherwise the resulting kinship matrix might not be positive definite which could cause numerical problems for the algorithm).

2. If individuals i and j have nonzero working kinship coefficient $\phi_{i,j}$, then they must have the same family ID.

Required Input Files:

1. **phenotype data file**

   This file contains the phenotype data as well as family ID and individual ID numbers for the study individuals. The columns in the file should be organized as follows:

1	1	1
1	2	2
1	3	2
1	4	1
1	5	0
2	86	2
2	20	1
2	30	0
2	14	1
3	110	0
4	51	0
(1)	(2)	(3)

   Column (1) family ID (positive integer)
   Column (2) individual ID (positive integer)
   Column (3) affection status (0=unknown, 1=unaffected, 2=affected)

   Family ID's should be numbered from 1 to F, without gaps, where F is the total number of families. Individuals from the same family must appear in a single cluster, and the clusters must be in consecutive order from 1 to F, though there is no requirement on the order of the individuals within a family. Each individual should be entered only once. The individual ID can be any positive integer. Individual ID must be unique within the family, but individuals in different families are permitted to have the same individual ID number. There is no limit on the number of individuals, but the number of families is set to be smaller than 10,000. To increase this limit, just change the value of MAXFAM in the ROADTRIPS_SOURCE.c source file and recompile the program. Sampled individuals who are unrelated to anyone else in the sample (i.e. all working kinship coefficients are zero) should be given their own unique family ID. See the previous subsection for more details about specifying family IDs when pedigree information is incomplete or unavailable.

   The default filename for the phenotype data file is "phenofile". To specify a different filename, use the command-line flag -p followed by the filename. For example, to use a phenotype data file called "myphenofile", you could type the command

   ./ROADTRIPS -p myphenofile

2. **genotype data file**

   This file contains the genotype data for the individuals listed in the phenotype data file. Exactly the same individuals must be in both files. Each row in this file corresponds to a SNP, and each column corresponds to a study individual, where the individuals must be in the same order in which they are listed in the phenotype data file. The ordering of SNPs is arbitrary, although if a SNP names file is specified, then the ordering in the genotype data file and the SNP names file must be the same.

   Each entry in the genotype data file should be one of the following integers: 0, 1, 2 or -9. The entry in the ith row and jth column represents the genotype, at the ith SNP, of the individual listed on the jth row of the phenotype file. A value of 0, 1, or 2 corresponds to the number of reference alleles the individual has at the SNP, while the value -9 is reserved for missing data. If any other value besides, 0, 1, 2, or -9 appears, then the corresponding SNP genotype for that individual is set to missing.

   To illustrate the format of the input genotype file, consider a study sample with a total of 11 individuals in the phenotype data file. The first few rows of the genotype data file for this sample could be as follows:

1	2	2	2	1	1	0	0	1	1	1
1	0	1	1	1	1	-9	0	1	2	0
2	1	2	1	1	2	1	2	1	1	0
0	-9	2	1	1	1	2	1	0	1	2
.	.	.	.	.	.	.	.	.	.	.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)

   Column (1) corresponds to the marker genotypes for the 1st individual listed in the phenotype data file
   Column (2) corresponds to the marker genotypes for the 2nd individual listed in the phenotype data file
   ⋮
   Column (10) corresponds to the marker genotypes for the 10th individual listed in the phenotype data file
   Column (11) corresponds to the marker genotypes for the 11th individual listed in the phenotype data file

   The default filename for the genotype data file is "genofile". To specify a different filename, use the command-line flag -g followed by the filename. For example, to use a SNP genotype data file called "mygenofile", you could type the command

   ./ROADTRIPS -g mygenofile

3. **pedigree information file**

   The pedigree information file contains working inbreeding coefficients for all the individuals in the phenotype data file and working kinship coefficients for all pairs of distinct individuals with the same family ID in the phenotype data file. If you have no pedigree information, then we recommend that you set all the working kinship and

inbreeding coefficients to zero. (See the preface to **Chapter 4** for more details on how to specify the family IDs and the working kinship and inbreeding coefficients.)

The pedigree information file has the following format (where the family and individual ID numbers listed below correspond to those in the example given in the phenotype data file section above):

1	1	1	0
1	1	2	0.25
1	1	3	0.25
1	1	4	0.25
1	1	5	0.25
1	2	2	0
1	2	3	0.25
1	2	4	0.25
1	2	5	0.25
1	3	3	0
1	3	4	0.25
1	3	5	0.25
1	4	4	0
1	4	5	0
1	5	5	0
2	86	86	0.01251
2	86	20	0.26124
.	.	.	.
.	.	.	.
(1)	(2)	(3)	(4)

Column (1) family ID

Column (2) individual 1 ID (Id1)

Column (3) individual 2 ID (Id2)

Column (4) kinship coefficient between individuals 1 and 2 if Id1 $\neq$ Id2, and inbreeding coefficient of individual 1 if Id1 = Id2

The family and individual ID's used here must be the same as the ones in the phenotype data file.

The default filename for the pedigree information file is "pedinfo". To specify a different filename, use the command-line flag -k followed by the filename. For example, to use a pedigree information file called "mypedinfo", you could type the command

```
./ROADTRIPS -k mypedinfo
```

Tips for inputting the pedigree information file

Case 1: When pedigrees are available:

When there is known pedigree information, the program runs faster when the coefficients are ordered in the following way :

In each family, the order of the pairs follows the order of individuals given in the pedigree data file. For instance, suppose that in family 14 there are exactly 3 individuals, their individual ID numbers are 7, 8 and 9, and the individuals are listed in this order (7,8,9) in the phenotype data file. Then, to maximize computational speed, the order in the pedigree information file for these 3 individuals should be:

$$\begin{array}{cccc} 14 & 7 & 7 & h_7 \\ 14 & 7 & 8 & \phi_{78} \\ 14 & 7 & 9 & \phi_{79} \\ 14 & 8 & 8 & h_8 \\ 14 & 8 & 9 & \phi_{89} \\ 14 & 9 & 9 & h_9 \end{array}$$

where h_i is the inbreeding coefficient of individual i, and ϕ_{ij} is the kinship coefficient between individuals i and j.

Two software programs that can be used to obtain kinship and inbreeding coefficients from pedigree data are:

1. The KinInbcoef software. The output file of the KinInbcoef program has the exact format required for the pedigree information input file to the ROADTRIPS software. The KinInbcoef program can be found at
   http://galton.uchicago.edu/∼mcpeek/software/index.html
2. The **IdCoefs** software by Mark Abney, which can be found at
   http://home.uchicago.edu/∼abney/Software.html
   The IdCoefs software computes identity coefficients for pairs of individuals. Kinship and inbreeding coefficients can then easily be computed from the identity coefficients (the output from this software).

Case 2: Incorporating unrelated individuals:

An individual who does not share a family ID with anyone else in the phenotype data file should be represented in the pedigree information file by a single line that specifies the individual's working inbreeding coefficient, which will almost always be zero in that case (unless a specific degree of inbreeding is known for that person).

In the example given in the phenotype data file subsection, there is an individual who has family ID 3 and individual ID 110, and there are no other individuals with family ID 3. Therefore, this person would be represented by the single line:

$$3 \quad 110 \quad 110 \quad 0$$

(with 0 possibly replaced by some other inbreeding coefficient if a specific degree of inbreeding is known for that person).

Case 3: When no pedigree information is available:

When no pedigree information is available, we recommend assigning a unique family ID number to each individual and setting all the working inbreeding coefficients to zero. (When each individual has a unique family ID number, the working kinship coefficients will automatically be assumed to be zero.)

For example, suppose the sample consists of 11 individuals with individual ID numbers 100, 12, 53, 41, 5, 86, 20, 30, 14, 110, and 51. If there is no pedigree information (or if the individuals are known to be unrelated), then these individuals would each be assigned a unique family ID, where these are numbered from 1 to 11, so that the phenotype data file could read, for example

1	100	1
2	12	2
3	53	2
4	41	1
5	5	0
6	86	2
7	20	1
8	30	1
9	14	0
10	110	2
11	51	2

(where the 3 columns are family ID, individual ID and phenotype), and the pedigree information file for these 11 individuals should be as follows:

1	100	100	0
2	12	12	0
3	53	53	0
4	41	41	0
5	5	5	0
6	86	86	0
7	20	20	0
8	30	30	0
9	14	14	0
10	110	110	0
11	51	51	0

4. **prevalence file**

   This file contains an estimate of the prevalence of the binary trait in an appropriate reference population. This prevalence value is used in the calculation of the *RM* statistic. This should not be the prevalence in the case-control sample, but rather the

"general population" prevalence for an appropriate reference population. See **Chapter 2** for further details about how to specify this and how it is used by the software.

The default filename for the prevalence file is "prevalence". To specify a different filename, use the command-line flag -r followed by the filename. For example, to use a prevalence file called "myprevalence", you could type the command

./ROADTRIPS -r myprevalence

Optional Input:

5. **SNP names file**

   This optional input file contains the names (or rs numbers) of the SNPs. The SNP names should be in the same order as the rows of the genotype data file, where each row corresponds to genotype data for a SNP. There should be not blank spaces within a SNP name. For example, "rs1000000" is a valid name for a SNP, while "rs 1000000" is not a valid name for a SNP because there is a blank space in the name. The SNP names can either be in a single row or a single column. If the SNP names are in a single row, then the names must be separated by a blank space.

   To specify a file that contains the names of the SNPs, use the command-line flag -n followed by the filename. For example, to use a file called "mysnpnames" that contains the names of the SNPs, you could type the command

   ./ROADTRIPS -n mysnpnames

6. **Exclude individuals with unknown phenotype**

   The command-line flag -u can be used to exclude all individuals with unknown phenotype from the analysis. (See **Chapter 2** for details.) To do this, you could type:

   ./ROADTRIPS -u

7. **Exclude phenotyped individuals with missing genotypes from the RM test**

   The RM test statistic can allow phenotyped individuals with missing genotypes at a SNP to contribute to the statistic, provided that those individuals have a sampled relative who is genotyped at the SNP. (See **Chapter 2** for details.) The command-line flag -m can be used to exclude phenotyped individuals with missing genotypes at a SNP from contributing to the RM statistic. To do this, you could type:

   ./ROADTRIPS -m

5 Output

1. **ROADTRIPStest.out** is the primary output file. It contains
   - Summary of the phenotype file information: total number of individuals in the phenotype file, number of independent families, number of individuals in each phenotype class (affected/unaffected/unknown)

- Prevalence value used in the RM calculations.
- For each marker
  - among those genotyped at the marker, the numbers who are affected, unaffected, and of unknown phenotype, respectively.
  - value of the RM statistic and corresponding p-value using the chi-squared null distribution.
  - value of the $R\chi$ statistic and corresponding p-value using the chi-squared null distribution. (Note: In the output of the ROADTRIPS software, "$R\chi$" will appear as "RCHI.")
  - value of the RW statistic and corresponding p-value using the chi-squared null distribution.
  - the signs of the RM and RW quasi-scores associated to each allele when the p-value is smaller than 0.05, in order to know the direction of the change in allele frequency associated with the RM or RW result.
  - a warning message is printed when some allele counts are small, a situation in which the χ^2 asymptotic null distribution might not provide accurate p-values
  - allele frequencies and s.d.'s estimated using the BLUE proposed by McPeek, Wu and Ober (2004) in the case sample, the unaffected control sample, the unknown phenotype control sample, and the entire sample (cases, unaffected controls, and unknown phenotype controls).
  - allele frequencies estimated by naive counting in the case sample, the unaffected control sample, the unknown phenotype control sample, and the entire sample (cases, unaffected controls, and unknown phenotype controls).

2. **ROADTRIPStest.top** lists the top 20 SNPs with the smallest p-values for each of the 3 tests. The number of markers output to this file can be decreased or increased by the user by changing MAXTOP (currently set to 20) in the ROADTRIPS_SOURCE.c source file.

3. **ROADTRIPStest.pvalues** lists the p-values for every SNP for RM, $R\chi$, and RW.

4. **ROADTRIPStest.err** is an error file that may contain warnings

   - when lines have incorrect number of fields in the genotype data file
   - when individuals from the pedigree information file are not listed in the phenotype data file
   - Whenever the -n option is used and the number of SNPs in the genotype data file is different from the number of SNP names that are given in the SNP name file.

6 Tips

1. Input

The program will stop if errors are detected in the formats of any of the input files. Please read **Chapter 4** carefully and make sure the input files are in the correct format and have concordant information.

2. Computation Time

   The computation time for calculating the empirical covariance matrix $\hat{\mathbf{\Psi}}$ used in the ROADTRIPS statistics will depend on the sample size, the number of SNPs, and the type of machine being used. For example, the time to compute $\hat{\mathbf{\Psi}}$ for a sample of 1,020 individuals and 100,000 SNPs for the ROADTRIPS software was approximately 13 minutes using a single processor on a shared machine with eight quad-core AMD Opteron 8384 25 GHz processors with 64 GB RAM. The computation time for the matrix scales linearly with the number of SNPs. We allow the user to specify the maximum number of SNPs that will be used to calculate $\hat{\mathbf{\Psi}}$ by changing MAX_SNPS_FOR_MATRIX (currently set to 500,000) in the ROADTRIPS_SOURCE.c source file. From simulation studies with 1,020 related individuals and population structure, we found that 100,000 SNPs across the genome was an adequate number of SNPs for $\hat{\mathbf{\Psi}}$ to correct for hidden structure, though this number may increase as the number of individuals in the study increases.

3. Small P-values

   ROADTRIPS outputs test statistics for each of the 3 tests. Then a χ^2 routine is used to convert the test statistics to p-values, which are also output. While the calculated test statistics are highly accurate, the χ^2 algorithm used in ROADTRIPS to calculate p-values for the 3 statistics is only accurate for p-values larger than 2.0e-09. For any test statistic that has a p-value less than 2.0e-09, the algorithm will report a p-value of 0. To get a more accurate p-value, one could simply plug the test statistic value into the χ^2 routine of another software program, such as the R software package. For instance, if the test statistic value were 44.56, then the following command in R

   ```
   1-pchisq(44.56,1)
   ```

   yields the p-value

   ```
   [1] 2.466805e-11
   ```

   In the next release of ROADTRIPS we hope to implement a different χ^2 algorithm that allows for an accurate assessment of much smaller p-values than the current release.

7 References

1. Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., McPeek, M.S. (2003). Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am. J. Hum. Genet. 73, 612-626.

2. McPeek, M.S., Wu, X., Ober, C. (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60, 359-367.

3. Thornton, T., McPeek, M.S. (2007). Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am. J. Hum. Genet. 81, 321-337.

4. Thornton T., McPeek M. S. (2010) ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. Am. J. Hum. Genet. 86, 172-184