# ROADTRIPS 2.0 Software Documentation (Beta Version)

Version 2.0

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

Department of Biostatistics[1]
The 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 2.0 (Betat) November 13, 2013

==========



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 program, written in C, that performs single-SNP, case-control association testing in samples with partially or completely unknown population and pedigree structure. ROADTRIPS is suitable for applications such as

(1) correcting for possible population structure 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

ROADTRIPS release 2.0 is an enhanced version of previous release of the program (release 1.1), where some of the new features of the program include:

(1) supports PLINK's (Purcell et al., 2007) transposed PED file as the input SNP genotype data file

(2) user option to print an empirical covariance matrix calculated by the program to a file

(3) user option to read in an empirical covariance matrix from a file and bypass calculating the empirical covariance matrix to save computational time

(4) allows for males and females to have different prevalence values for a trait

(5) more accurate assessment of p-values of the test statistics in the extreme tail of the $\chi^2_1$ null distribution.

Additional features of the ROADTRIPS tests include:

(1) ROADTRIPS does not require known pedigree information for the sampled individual, but when pedigree information is available, ROADTRIPS improves power (by use of the *RM* test) by taking advantage of the principle that there is an enrichment for predisposing variants in affected individuals with affected relatives.

(2) ROADTRIPS has the the ability to incorporate both unaffected controls and controls of unknown phenotype (e.g., general population controls) in the analysis.

(3) ROADTRIPS appropriately handles missing genotype data to construct valid tests by taking into account both structure and the particular missing genotype pattern at each SNP.

(4) ROADTRIPS (by use of the *RM* test) can 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.

ROADTRIPS uses an empirical covariance matrix calculated from genome-screen SNP data from the autosomal chromosomes to correct for known and unknown structure in the data. Separate association tests are performed at each SNP. The program is applicable to association studies with completely general combinations of related and unrelated individuals. The main reference for this program is Thornton and McPeek (2010).

For each marker, the ROADTRIPS program computes 3 different test statistics for association: the $RM$ test, which is the ROADTRIPS extension of the $M_{QLS}$ test statistic of Thornton and McPeek (2007); the $R\chi$ test, which is the ROADTRIPS extension of the corrected $\chi^2$ test statistic of Bourgain et al. (2003); and the $RW$ test, which is the ROADTRIPS extension of the $W_{QLS}$ test statistic of Bourgain et al. (2003). The $RM$ test has also been generalized to allow sex-specific prevalences. See the next section for details on this change. For each test, a p-value is calculated based on a $\chi^2_1$ asymptotic null distribution.

When pedigree information on the sampled individuals is available, we recommend using the $RM$ test, since in the context of complex trait mapping in samples of related individuals with population structure, $RM$ has been shown to generally have more power than $R\chi$ and $RW$. For a more detailed comparison of the ROADTRIPS statistics, see Thornton and McPeek (2010).

# 2 Detailed Descriptions of Some of the New Features

In this section, we give more details on some of the new features in ROADTRIPS 2.0, specifically, the options for: (1) sex-specific prevalences, (2) saving the empirical genetic relatedness matrix calculated by the program, and (3) a user-specified genetic relatedness matrix

The ROADTRIPS software calculates three test statistics ($RM$, $RW$, and $R\chi$). Note that the $RM$ is the preferred statistic for most applications.

### Extension of $RM$ to Include Sex-Specific Prevalences

ROADTRIPS 2.0 includes a new extension of the $RM$ to also include sex-specific prevalences. This extension is closely related to the work of Thornton et al. (2012), but we spell it out in a little more detail here.

The $RM$ test is an extension of the $M_{QLS}$ test (Thornton and McPeek 2007) to samples from structured populations. The original $M_{QLS}$ tests (and $RM$ test) involves a population prevalence estimate, $k$, that is not sex-specific. To describe the extension of the $M_{QLS}$ (and the $RM$ test) to sex-specific prevalences, it is convenient to use the formulation of the $M_{QLS}$ statistic given by equations (2) and (3) on p. 440 of Thornton et al. (2012). This formulation is in terms of a phenotypic residual vector, $\mathbf{R}$, having $i$th entry $R_i = 0$ if individual $i$ has unknown phenotype and $R_i = 1_{\{i \text{ case}\}} - k$ if $i$ has known phenotype, where $1_{\{i \text{ case}\}}$ is the indicator function for the event that $i$ is affected. The extension of the $M_{QLS}$ to

sex-specific prevalences is obtained by replacing **R** by a different phenotypic residual vector, **A** (given on p. 443 of Thornton et al. 2012), where $A_i = 0$ for $i$ of unknown phenotype, $A_i = 1_{\{i \text{ case}\}} - k_f$ for $i$ female with known phenotype and $A_i = 1_{\{i \text{ case}\}} - k_m$ for $i$ male with known phenotype, where $k_f$ and $k_m$ are estimates of the population prevalence of the trait in, respectively, females and males, with $0 < k_f, k_m < 1$. Analogously, the extension of the $RM$ to sex-specific prevalences is obtained using the sex-specific phenotypic residual vector, **A** (given on p. 443 of Thornton et al. 2012).

**What should I plug in for the sex-specific prevalences used in $RM$?**

To calculate the $RM$ statistic, estimates of the prevalences of the trait among males and among females in a suitable reference population must be specified by the user. These sex-specific prevalences can be replaced by a common, pooled estimate of the population prevalence if sex-specific prevalence information is not available or if there is no evidence of a sex difference in prevalence. We recommend using prevalence estimates from previous studies or registry data from the population, when available. For studies with random ascertainment, the sex-specific prevalences could be estimated by the sample case frequencies in females and in males, or, if the male and female prevalences are assumed to be equal, then the overall sample case frequency could be used. However, when ascertainment is phenotype-based, prevalence estimates should ideally be obtained from external, population-based data, rather than from the sample case frequencies in the data set. We emphasize that the $RM$ test will be valid regardless of the prevalence values used, but accurate prevalence values would be expected to increase power (see Thornton and McPeek 2007 and Thornton et al. 2012 for details).

**Saving the Empirical Genetic Relatedness Matrix**

A new feature of the ROADTRIPS software is that there is a user option to save the empirical correlation matrix calculated by the software to a file. As this matrix will only need to be calculated once, saving the empirical matrix to a file and then reading the matrix in for any future analyses can significantly reduce the computational time of the program.

**User Specified Genetic Relatedness Matrix**

The default setting in the ROADTRIPS software is to calculate the three test statistics using an empirical matrix that is calculated from the input genotype data file. The ROADTRIPS software now includes a user option to read in a file specified by the user that contains a genetic relatedness matrix. If this option is used, the program will bypass the calculation of the empirical matrix and the three test statistics will be calculated using the genetic relatedness matrix in the file specified by the user.

# 3 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 $i$th 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 $\phi_{ij}$ 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 R$\chi$ test** The $R\chi$ statistic is a version of the standard Pearson $\chi^2$ 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 $i$th 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 $R\chi$ 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.

# 4 Running ROADTRIPS

Installation instructions:

1. Download the ROADTRIPS 2.0 package. This package contains documentation, source code, example input and output files, and a precompiled executable for Linux platforms.

2. Read the file ROADTRIPS2_Documentation.pdf 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". If the message "make: 'ROADTRIPS' is up to date" appears after typing "make", then to build the executable program you must first delete the precompiled binary ROADTRIPS program that comes with the software by typing "rm ROADTRIPS", and then type "make" to build the executable program ROADTRIPS.

5. ROADTRIPS is run from the command line via the command 'ROADTRIPS' with all information, including the type of anlaysis, specified by command line options. To run the executable program ROADTRIPS:

   First, prepare the input files, e.g., genofile, phenofile, kinfile, prevalence (see Section 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 -g genofile -p phenofile -k kinfile -r prevalence -e matrixfile -f -u -m

   We briefly summarize the meanings of the flags below. More details can be found in section 4:

   **-g genofile** Allows the user to specify the name of the SNP genotype data input file. Filename defaults to "genofile" if this flag is not used. To specify a different filename, replace "genofile" with the appropriate filename.

   **-p phenofile** Allows the user to specify the name of the phenotype information input file. (This file also includes family ID numbers, individual ID numbers, and sex, in addition to phenotype.) The filename defaults to "phenofile".

   **-k kinfile** Allows the user to specify the name of the pedigree information input file which contains inbreeding coefficients for the sampled individuals and kinship coefficients for pedigrees that are partially or completely known. Filename defaults to "kinfile".

   **-r prevalence** Allows the user to specify the name of the prevalence input file, which contains estimates of the prevalence of the binary trait in males and females from a suitable reference population. Filename defaults to "prevalence".

   **-u** Allows the user to exclude individuals with unknown phenotype from the analysis for the 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 specified by the known pedigree information who is genotyped at the marker.

**-f** Allows the user to print the empirical correlation matrix to a file named "ROADTRIPS_MATRIX.txt", which can then be read in for other analyses with the software with the user-option"-e" followed by the name of file containing the matrix, as discussed below.

**-e matrixfile** Allows the user to specify a file containing an empirical correlation matrix for calculating the association test statistics and bypass calculating the empirical correlation matrix from the input genotype data file. To specify a different filename, replace "matrixfile" with the appropriate filename. For example to read in an empirical correlation matrix that was saved to a file from a previous analysis using the user-option " -f " (discussed above), include the following in the command line: "-e ROADTRIPS_MATRIX.txt". If this option is not used, an empirical correlation matrix will be calculated from the input genotype data file.

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

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

# 5 Input

## Required Input Files:

1. **genotype data file**

   The genotype data file is a transposed genotype file containing the SNP names and locations and the genotypes of the sampled individuals. The genotype data file is in the PLINK tped file format, with some additional restrictions, namely:

   1. genotypes for individuals from the same family must be listed consecutively;
   2. the order of individuals must be the same in the genotype data file and in the phenotype information file described in the next subsection.
   3. The genotype data file should contain genotypes only for autosomal SNPs.
   4. The two alleles of a SNP must be coded as 1 and 2, and missing alleles must be coded as 0.

   To illustrate the format of the genotype data file, consider a study sample with a total of 8 individuals. The first few rows of the genotype data file for this sample could be as follows:

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs3094315 | 0 | 742429 | 1 | 2 | 2 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 |
| 1 | rs2286139 | 0 | 751595 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 1 | 1 | 2 |
| 1 | rs11240776 | 0 | 755132 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| 1 | rs2980300 | 0 | 775852 | 0 | 0 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) | (14) | (15) | (16) | (17) | (18) | (19) | (20) |

Column (1) contains the chromosome name (1-22). This information will be ignored by the program (but some string must be present in the column).
Column (2) contains the rs number or SNP identifier.
Column (3) contains the genetic distance in Morgans (0 if genetic distance is unknown). This information will be ignored by the program (but some string must be present in the column).
Column (4) contains the base-pair position in bp units (0 if base-pair position is unknown). This information will be ignored by the program (but some string must be present in the column).
Columns (5) and (6) contain the marker genotype (one allele in each column) for the 1st individual. (Note restrictions 1-4 listed above on how the individuals should be ordered and how the genotypes should be coded.)
Columns (7) and (8) contain the marker genotype for the 2nd individual.
⋮
Columns (17) and (18) contain the marker genotype for the 7th individual.
Columns (19) and (20) contain the marker genotype for the 8th individual.

For more details on the tped file format, you could consult the PLINK website (http://pngu.mgh.harvard.edu/ purcell/plink/data.shtml#tr). PLINK provides a convenient venue to convert from many different file formats. For example, supposed you had a PLINK ped file named "mydata.ped" which was coded as A/C/G/T or 1/2/3/4 for the four alleles. Then, assuming that restrictions 1 and 3 above were met by mydata.ped, i.e. individuals from the same family were listed consecutively and the file did not mix autosomal SNPs with X-chromosome SNPs, you could generate the desired ROADTRIPS genotype input file with the PLINK command:

./plink --file mydata --recode12 --output-missing-genotype 0 --transpose --out newfile

The PLINK software would then create the two files "newfile.tped" and "newfile.tfam". The file "newfile.tped" is a genotype data file that is in the appropriate format for the ROADTRIPS software. **If, in addition, the phenotype was coded as 2=affected, 1=unaffected, and 0=unknown** in the file mydata.ped, then the tfam file "newfile.tfam" could be converted to the required phenotype information file as discussed in item 2 below.

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 the PLINK SNP genotype data file named "newfile.tped", you could type the command

./ROADTRIPS -g newfile.tped

2. **phenotype information file**

This file contains the phenotype data as well as family ID and individual ID numbers for the study individuals. Important restrictions to keep in mind include:

1. Individuals must be listed in the same order as in the genotype data file;
2. Individuals from the same family must appear in a single cluster, and families must be numbered consecutively in the file from 1 to $F$, where $F$ is the total number of families in the sample. In other words, the first family listed in the file must have family ID number 1, the second family listed in the file must have family ID number 2, and so on.
3. Individual IDs must be positive integers.
4. The phenotype must be coded as 2=affected, 1=unaffected, 0=unknown.

To illustrate the format of the input phenotype information file, consider a study sample with a total of 8 individuals from 3 families. The columns in the phenotype information file should be organized as follows:

| | | | | | |
|---|---|---|---|---|---|
| 1 | 11 | 0 | 0 | 1 | 1 |
| 1 | 23 | 0 | 0 | 2 | 2 |
| 1 | 13 | 11 | 23 | 1 | 2 |
| 2 | 2 | 0 | 0 | 2 | 1 |
| 2 | 51 | 0 | 0 | 1 | 2 |
| 2 | 3 | 51 | 2 | 1 | 0 |
| 2 | 41 | 51 | 2 | 2 | 1 |
| 3 | 1 | 0 | 0 | 2 | 0 |
| (1) | (2) | (3) | (4) | (5) | (6) |

Column (1) contains family ID (consecutive positive integers from 1 to $F$).
Column (2) individual ID (positive integer)
Column (3) father's ID (0=founder)

Column (4) mother's ID (0=founder)

Column (5) sex (1=male, 2=female)

Column (6) affection status (0=unknown, 1=unaffected, 2=affected)

Sampled individuals who are unrelated to anyone else in the sample should be included in this file by giving each such person their own unique family ID. 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.

As mentioned in the previous sub-section, a PLINK tfam file that meets certain restrictions can easily be converted to a phenotype information file by using the FORMAT_PED_PHENO software, which can be found at
**http://galton.uchicago.edu/∼mcpeek/software/index.html**. The restrictions are:

1. phenotype values must be coded as 2=affected, 1=unaffected, and 0=unknown;
2. individuals from the same family must be listed consecutively;
3. the order of individuals must be the same as that in the genotype data file (previous subsection).

For example, to use the FORMAT_PED_PHENO software to convert the PLINK file "newfile.tfam" discussed in the previous subsection, the following command can be used:

./FORMAT -f newfile.tfam -o newfile

This command will generate the file "newfile.pedpheno", which is in the appropriate input format for the ROADTRIPS software.

The default filename for the phenotype information file used by the ROADTRIPS is "phenofile". To specify a different filename, use the command-line flag -p followed by the filename. For example, to use the file "newfile.pedpheno", you could type the command

./ROADTRIPS -p newfile.pedpheno

3. **kinship coefficient file**

This file contains kinship coefficients for any known or partially known pedigree information for individuals that are members of the same family (i.e., have the same family ID) in the phenotype information file. The file also contains inbreeding coefficients for all sampled individuals, where the inbreeding coefficient will be 0 for all outbred individuals.

Kinship coefficients are required for all pairs of individuals in the same family, regardless of phenotype.

A sampled individual who does not share a family ID with anyone else in the phenotype information file should be represented in this file by a single line that contains the the individuals inbreeding coefficient value. As previously mentioned, the inbreeding coefficient will be 0 for all outbred individuals, and we recommend an inbreeding coefficient value of 0 for any individual for whom there is no available pedigree information.

The pedigree information file has the following format:

Column (1) family ID

Column (2) individual 1 ID (Id1)

Column (3) individual 2 ID (Id2)

Column (4) kinship coefficient between Id1 and Id2 if Id1 is different from Id2, and inbreeding coefficient of Id1 if Id1 equals Id2

The family ID and individual ID should match exactly with the Id's in the phenotype information file.

| | | | |
|---|---|---|---|
| 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 | 1 | 1 | 0.01251 |
| 2 | 1 | 2 | 0.26124 |
| . | . | . | . |
| . | . | . | . |
| (1) | (2) | (3) | (4) |

An individual who does not share a family ID with anyone else in the phenotype information file should be represented in the pedigree information file by a single line that contains the the individuals inbreeding coefficient value. As previously mentioned, the inbreeding coefficient will be 0 for all outbred individuals, and we recommend using an inbreeding coefficient value of 0 for any individual for whom there is no available pedigree information. As an example, consider once again a study sample with a total of 11 individuals that are included in the phenotype information file, and assume that there is no pedigree information available for the individuals or all of the individuals are known to be unrelated and are outbred. In the phenotype information file, each of the study individuals should be in a separate family, and the family ID's should be numbered from 1 to 11. If every individual within each family has a study ID of "1" in the phenotype information file, then the pedigree information file for these 11 individuals should be as follows:

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 follow the order of the individuals given in the pedigree data file. Considering a family numbered 14 with 3 individuals with ID's of 7, 8 and 9, with the individuals listed in this order in the phenotype information data file. Then the order in the pedigree information file for these 3 individuals should be:

where $h_i$ is the inbreeding coefficient of individual $i$, and $\phi_{ij}$ is the kinship coefficient

| | | | |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 2 | 1 | 1 | 0 |
| 3 | 1 | 1 | 0 |
| 4 | 1 | 1 | 0 |
| 5 | 1 | 1 | 0 |
| 6 | 1 | 1 | 0 |
| 7 | 1 | 1 | 0 |
| 8 | 1 | 1 | 0 |
| 9 | 1 | 1 | 0 |
| 10 | 1 | 1 | 0 |
| 11 | 1 | 1 | 0 |
| 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 | $h_9$ |

between individuals $i$ and $j$.

We provide a software program that can be used to calculate the required coefficients and generate the kinship coefficient file. The output file of this program has the exact format required for the ROADTRIPS kinship coefficient input file:

1. The KinInbcoef software for calculating autosomal kinship and inbreeding coefficients.

This program can be found at
**http://galton.uchicago.edu/∼mcpeek/software/index.html**

The FORMAT_PED_PHENO software, discussed in the previous subsection, for converting a phenotype information file (e.g., a PLINK tfam file) also creates input files that are in the appropriate format for the KinInbcoef software program. For example, the ROADTRIPS input phenotype file "newfile.pedpheno" mentioned in the previous subsection will be created by the FORMAT_PED_PHENO software with the following command:

./FORMAT -f newfile.tfam -o newfile

This command also creates the output file "newfile.kinpedigree" and "newfile.kinlist", which can be used as input to the KinInbcoef software.

To obtain kinship coefficients using the KinInbcoef software, type the following command:

./KinInbcoef newfile.kinpedigree newfile.kinlist newfile.kinship

This command creates the output file "newfile.kinfile" which is in the exact format required for the ROADTRIPS software.

The default filename for the kinship coefficient file used by ROADTRIPS is "kinfile". To specify a different filename, use the command-line flag -k followed by the filename. For example, to use the file "newfile.kinfile", you could type the command

./ROADTRIPS -k newfile.kinfile

4. **prevalence file**

   This file contains estimates of the male and female prevalence values for the binary trait in an appropriate reference population. These values are used in the calculation of the *RM* statistic. The two estimates can either be in a single row or a single column, where the first estimate is for the male prevalence and the second estimate is for the female prevalence. If the prevalence estimates are in a single row, then the values must be separated by a blank space. If only one value is given in the prevalence file, then both the male and female prevalences will be set to this value in the analysis. Please read the subsection **What should I plug in for the sex-specific prevalences used in** *RM* in section 2 of this document.

   The default filename 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:

**-f** Allows the user to save the empirical correlation matrix to a file named "ROADTRIPS_MATRIX.txt", which can then be read in for future analysis with the software with the user-option discussed below.

**-e matrixfile** Allows the user to specify a file containing an empirical correlation matrix for calculating the association test statistics and bypass calculating the empirical correlation matrix from the input genotype data file. To specify a different filename, replace "matrixfile" with the appropriate filename. For example to read in an empirical correlation matrix that was saved to a file using the user-option " -f " (discussed above), include the following in the command line: "-e ROADTRIPS_MATRIX.txt". If this option is not used, an empirical correlation matrix will be calculated from the input genotype data file.

5. **Print empirical correlation matrix to a file**

   The command-line flag -f would be used to print the empirical correlation matrix to a file named "ROADTRIPS_MATRIX.txt". For example, to print the empirical correlation matrix to a file, you could type the command

   ./ROADTRIPS -f

5. **Read in an empirical correlation matrix from a file**

   The command-line flag -e followed by the name of a file containg an empirical correlation matrix would be used to bypass calculating the empircal correlation matrix from

the input genotype data file and to calculate the association test statistics with the user specified file containing an empirical matrix. For example to read in file named "ROADTRIPS_MATRIX.txt" containing an empirical correlation matrix name that was saved to a file using the user-option " -f " (discussed above) from a previous analysis, you could type the command

./ROADTRIPS -e ROADTRIPS_MATRIX.txt

5. **Exclude individuals with unknown phenotype**

   The command-line flag -u can be used to exclude all individuals with unknown phenotype from the analysis. For example, to exclude individuals with unknown phenotype, you could type the command

   /ROADTRIPS -u

   The $RM$ test explicitly allows for individuals of unknown phenotype and handles them appropriately in the analysis if this flag is not used. In contrast, the $RW$ and $R\chi$ tests classify individuals of unknown phenotype as controls (same as unaffected) if this flag is not used.

7. **Exclude phenotyped individuals with missing genotypes for 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. The command-line flag -m can be used to exclude phenotyped individuals with missing genotypes at a SNP from contributing to the statistics. For example, to exclude phenotyped individuals with missing genotype from the analysis, you could type the command

   /ROADTRIPS -m

   This flag has no effect on the $RW$ and $R\chi$ tests, which always exclude individuals with missing genotypes.

# 6 Output

1. **ROADTRIPStest.out** is the primary output file. The file 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 values used in the $RM$ calculations.
   - For each marker
     - SNP identifier/rs number
     - 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
- 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 $RM$ and $RW$ result.
- a warning message is printed when some allele counts are small, a situation in which the $\chi^2$ asymptotic null distribution might not provide accurate p-values
- allele frequencies and s.d.'s estimated using the quasi-likelihood score function proposed by McPeek, Wu and Ober (2003) 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 changed by changing the value of MAXTOP (currently set to 20) in the ROADTRIPS_SOURCE.c file.

3. **ROADTRIPStest.testvalues** lists, for every SNP, the values of each of the three test statistics.

4. **ROADTRIPStest.pvalues** lists the p-values for every SNP for each of the three statistics.

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

   - when a line has an incorrect number of fields in the genotype data file
   - when an individual from the kinship coefficient file is not listed in the pedigree data file

# 7 Tips

1. Input

   The program will stop if errors are detected in the formats of any of the input files. Please read Section 5 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 computation 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. If the number of SNPs were increased by a factor of 10, i.e., one million SNPs were used to calculate $\hat{\mathbf{\Psi}}$, then we expect the computation time to be around 130 minutes. 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 related individuals and population structure, we found that 100,000 SNPs across the genome is an adequate number of SNPs for $\hat{\mathbf{\Psi}}$ to capture hidden structure. The empirical covariance matrix will only need to be calculated once and can be saved to a file named "ROADTRIPS_MATRIX.txt" with the user-option "-f". The empirical correlation matrix can then be read in with the user-option"-e ROADTRIPS_MATRIX.txt" for any future analyses with the ROADTRIPS software.

# 8 Example

1. Consider a PLINK ped file named "mydata.ped". Suppose that the following two conditions are met: (1) individuals from the same family are listed consecutively in the file; and (2) the file contains only autosomal genotypes. Then the PLINK command below can be used to obtain tped and tfam output files:

   ./plink --file mydata --recode12 -- output-missing-genotype 0 --transpose --out newfile

   This command creates the two files "newfile.tped" and "newfile.tfam". The file "newfile.tped" is a genotype data file that is in the appropriate format for the ROADTRIPS software package.

2. If, in addition to the two requirements above, the phenotype values are coded as 2=affected, 1=unaffected, and 0=unknown in the "newfile.tfam" file, then the FORMAT_PED_PHENO software can be used to obtain a phenotype information file that is in the required format for the ROADTRIPS software. To convert the "newfile.tfam" file with the FORMAT_PED_PHENO software, the following command can be used:

   ./FORMAT -f newfile.tfam -o formattedfile

   This command will create a phenotype information file named "formattedfile.pedpheno" file that is in the appropriate format for the ROADTRIPS software.

3. As mentioned in the previous subsection, the "formattedfile.pedpheno" file will be created by the FORMAT_PED_PHENO software with the command

   ./FORMAT -f newfile.tfam -o formattedfile

This command also creates the files "formattedfile.kinpedigree" and "formattedfile.kinlist" which are in the appropriate format for the KinInbcoef software. To obtain autosomal kinship and inbreeding coefficients with the KinInbcoef software and the input files "formattedfile.kinpedigree" and "formattedfile.kinlist", the following command can be used:

./KinInbcoef formattedfile.kinpedigree formattedfile.kinlist final.kinship

This command creates the output file "final.kinship" which is in the exact format required by the ROADTRIPS software for the autosomal kinship coefficient file.

4. Now, to run the ROADTRIPS software for association testing of autosomal SNPs using the genotype input file from the PLINK software, the phenotype information file from that FORMAT software, the output autosomal kinship file from the KinInbcoef software, and a file called "myprev" that contains the male and female prevalences, the following command can be used:

   ./ROADTRIPS -g newfile.tped -p formattedfile.pedpheno -k final.kinship -r myprev

# 9 Acknowledgements

We gratefully acknowledge Jerry Halpern (funn@stanford.edu) for his contribution in implementing an algorithm for calculating p-values based on a $\chi^2_1$ asymptotic null distribution.

# 10 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. (2012). BLUP genotype imputation for case-control association testing with related individuals and missing data. J. Comp. Biol. 19, 756-765.

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

4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. Am. J. Hum. Genet. 81, 559-575. with Partially or Completely Unknown Population and Pedigree Structure. Am. J. Hum. Genet. 86, 172-184

5. 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.

6. 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

7. Thornton T, Zhang Q, Cai X, Ober C, and McPeek MS (2012) XM: Association Testing on the X-Chromosome in Case-Control Samples with Related Individuals. Genet Epidemiol 36, 438-450.