Q-ROADTRIPS Software Documentation (Beta Version)

Version 1.0

Timothy Thornton[1]

Department of Biostatistics[1]
The University of Washington

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

A C program for association testing with general quantitative traits that allows for partially or completely unknown population and pedigree structure

Homepage: http://faculty.washington.edu/tathornt/software/QROADTRIPS/index.html

Release 1.0 April 29, 2013

==========

The manuscript for the Q-ROADTRIPS method and software is currently in preparation.

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 Q-ROADTRIPS

Q-ROADTRIPS is a program, written in C, that performs single-SNP, quantitative trait association testing in samples with unknown population and/or pedigree structure. Q-ROADTRIPS is an extension of the ROADTRIPS case-control method to quantitative traits:

> 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.

The Q-ROADTRIPS program can be used for general quantitative traits/phenotypes and there are no restrictions on the types of distributions the traits can have. For example, a trait can follow a normal distribution, can be count values (e.g., from a poisson distribution), a mixture of different distributions, or the distribution of the trait can be completely unknown. Q-ROADTRIPS is suitable for applications such as correcting for partially or completely unknown population structure and/or relatedness in the context of GWAS. We recommended using phenotype values in the Q-ROADTRIPS software that have been adjusted for relevant covariates such as age, sex, etc., which can improve the power to detect phenotype/genotype association. For example, one can use as the input phenotype values for the Q-ROADTRIPS software the residuals from a regression analysis with phenotype as the response and any relevant explanatory variables included as covariates (or predictors) in the model.

Q-ROADTRIPS uses an empirical correlation matrix, calculated from genome-screen data, to capture structure in the sample. Q-ROADTRIPS does not require known pedigree information for the sampled individual, and the current release does not incorporate known relatedness in the analysis. Another software package will be made available for family-based GWAS that allow the user to input any known familial relations, which could potentially improve power. Please email the author for a version of the software that can incorporate known familial information for family-based GWAS. Some of the features of the Q-ROADTRIPS software include:

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

(1) supports an input phenotype information file that is in the format of a PLINK fam or tfam 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) accurate assessment of p-values of the test statistics in the extreme tail of the χ^2_1 null distribution.

Q-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.

For each marker, the Q-ROADTRIPS program computes a test statistics, and for each test, a p-value is calculated based on a χ^2_1 asymptotic null distribution.

2 Description of the Q-ROADTRIPS Statistic QR

The **QR** test statistic is an extension of the ROADTRIPS case-control association test $R\chi$ of Thornton and McPeek (2010) to quantitative traits. Similar to the $R\chi$ test, the QR test uses an empirical correlation matrix $\hat{\mathbf{\Psi}}$ so that the test is valid when there is cryptic structure.

3 User Specified Empirical Correlation Matrix

The default setting in the Q-ROADTRIPS software is to calculate the test statistic using an empirical correlation matrix that is calculated from the input genotype data file. The Q-ROADTRIPS software includes a user option to read in a file specified by the user that contains an empirical correlation matrix. If this option is used, the program will bypass the calculation of the empirical matrix and the QR test statistics will be calculated using the empirical correlation matrix specified by the user. Another feature of the Q-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.

4 Running Q-ROADTRIPS

Installation instructions:

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

2. Read the file Q-ROADTRIPS_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 "QROADTRIPS". If the message "make: 'QROADTRIPS' is up to date" appears after typing "make",

then to build the executable program you must first delete the precompiled binary QROADTRIPS program that comes with the software by typing "rm QROADTRIPS", and then type "make" to build the executable program QROADTRIPS.

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

   First, prepare the input files, e.g., example.tped and example.pheno.

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

   ./QROADTRIPS

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

   ./QROADTRIPS -g example.tped -p example.pheno -e matrixfile -f

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

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

   -p example.pheno 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 "example.pheno". To specify a different filename, replace "example.pheno" with the appropriate filename.

   -f Allows the user to print the empirical correlation matrix to a file named "QROADTRIPS_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 QROADTRIPS_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 QROADTRIPS by running it with the sample input files: example.tped and example.pheno. You can then compare the resulting output, which will be printed to the files QROADTRIPStest.out, QROADTRIPStest.top, QROADTRIPStest.pvalues, and QROADTRIPStest.testvalues, with the correct output provided in the sample output files QROADTRIPStest.out.ex, QROADTRIPStest.top.ex, QROADTRIPStest.pvalues.ex, and QROADTRIPStest.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 the following requirements:

   1. the order of individuals must be the same in the genotype data file and in the phenotype information file described in the next subsection.
   2. The genotype data file should contain genotypes only for autosomal SNPs.
   3. 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 1-5 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, suppose 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. the file did not mix autosomal SNPs with X-chromosome SNPs, you could generate the desired QROADTRIPS 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 QROADTRIPS software. **In addition, if a quantitative phenotype in the mydata.ped file has missing values coded as 'NA" , then the tfam file "newfile.tfam" would also be in the required format for phenotype information file as discussed in item 2 below.**

The default filename for the genotype data file is "example.tped". 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

./QROADTRIPS -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. The order of the individuals in the phenotype information file must be the same as the order of the individuals in the tfam generated by PLINK when the tped file was created, as discussed in item 1 above, i.e., the order of individuals must be the same in the genotype data file and in the phenotype information file. The columns in the phenotype information file should be organized as follows:

Column (1) contains family ID.
Column (2) contain individual ID.
Column (3) contain father's ID. This information will be ignored by the program (but some string must be present in the column).

F192	ID101	0	0	1	0.0735
G195	ID133	0	0	2	1.309
L197	ID201	ID101	ID102	2	-0.154
M013	205HG	0	0	1	0.23982
Q015	ID309	0	0	2	NA
R013	ID2999	205HG	ID309	1	0.1148
N019	ID9203	205HG	ID309	2	0.423
S230	LF6950	0	0	2	-1.123
(1)	(2)	(3)	(4)	(5)	(6)

Column (4) contains mother's ID. This information will be ignored by the program (but some string must be present in the column).
Column (5) contains sex. This information will be ignored by the program (but some string must be present in the column).
Column (6) phenotype value (phenotype values should be a numeric, and missing phenotype values are coded as "NA".)

Each family ID is assumed to be a character string that does not contain a space, newline or tab. The same assumption holds for each individual ID.

The number of individuals in the study is set to be smaller than 10,000. To increase this limit, just change the value of MAXFAM in the QROADTRIPS_SOURCE.c source file and recompile the program.

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

./QROADTRIPS -p newfile.pheno

Optional Input:

-f Allows the user to save the empirical correlation matrix to a file named "QROADTRIPS_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 QROADTRIPS_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 "QROADTRIPS_MATRIX.txt". For example, to print the empirical correlation matrix to a file, you could type the command

   ./QROADTRIPS -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 empirical 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 "QROADTRIPS_MATRIX.txt" containing an empirical correlation matrix name that was saved to a file using the user-option " -f " (discussed above), you could type the command

   ./QROADTRIPS -e QROADTRIPS_MATRIX.txt

6 Output

1. **QROADTRIPStest.out** is a detailed output file that contains the following:
   - Summary of the phenotype file information: total number of individuals in the phenotype file, number of individuals with non-missing phenotype values, and number of individuals with no available phenotype information.
   - For each marker
     - SNP identifier/rs number
     - the numbers of genotyped individuals at the marker who have non-missing phenotype values.
     - value of the QR statistic and corresponding p-value using the chi-squared null distribution.
     - 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 estimated based on allele counts in the sample
2. **QROADTRIPStest.top** lists the top 20 SNPs with the smallest p-values for QR test. The number of markers output to this file can be changed by changing the value of MAXTOP (currently set to 20) in the QROADTRIPS_SOURCE.c file.
3. **QROADTRIPStest.testvalues** lists, for every SNP, the values of the QR test statistic.
4. **QROADTRIPStest.pvalues** lists the p-values for every SNP for the QR test statistic.

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

   - when a line has an incorrect number of fields in the genotype data file
   - when the number of genotypes read in for a SNP does not match the number of individuals in the phenotype information 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 QROADTRIPS 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 QROADTRIPS 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 "QROADTRIPS_MATRIX.txt" with the user-option "-f". The empirical correlation matrix can then be read in with the user-option"-e QROADTRIPS_MATRIX.txt" for any future analyses with the QROADTRIPS 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. Note that if phenotype of interest is in the mydata.ped file with missing values coded as "NA", then the "newfile.tfam" file would also be in the appropriate format for the phenotype information file that is required by the QROADTRIPS software.

3. Now, to run the QROADTRIPS software for association testing of autosomal SNPs using the genotype input file from the PLINK software and a phenotype information file named "mystudy.pheno" , for example, the following command can be used:

   ./QROADTRIPS -g newfile.tped -p mystudy.pheno

9 Acknowledgements

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

10 References

1. 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

2. 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