QUICKTEST user guide

Toby Johnson* Zoltán Kutalik†

December 11, 2008
for QUICKTEST version 0.94

The most up-to-date version of the QUICKTEST program, source code and documentation, including this user guide, are all available from:
`http://toby.freeshell.org/software/quicktest.shtml`

*<toby.johnson@unil.ch>
†<zoltan.kutalik@unil.ch>

Contents

1 Installing or compiling QUICKTEST **3**
1.1 Precompiled versions 3
1.2 Compiling using the source code 3

2 Quick start for SNPTEST users **4**

3 The main features of QUICKTEST **5**

4 Accounting for covariates **6**

5 Exclusions and analyses of subsets of your data **7**
5.1 Analysing a subset of SNPs 7
5.2 Analysing a subset of individuals 7

6 Preparation and format of input files **8**
6.1 Genotype files 8
6.2 Phenotype files 8

7 Association testing methods **9**
7.1 Differences between QUICKTEST and SNPTEST output 10

8 File manipulation utilities **10**

9 Testing interactions **10**

10 Binary phenotype analysis **11**

11 Problems, error messages, FAQ **11**

1 Installing or compiling QUICKTEST

1.1 Precompiled versions

We provide precompiled versions of QUICKTEST for some computer systems. As of December 11, 2008, we provide:

- `quicktest-0.94-i686`
- `quicktest-0.94-x86_64`

These are intended for GNU/Linux operating systems, running on i686 (32 bit) and x86_64 (64 bit) hardware architectures respectively. To find out your operating system, kernel version number, and hardware architecture, type "`uname -a`" at the shell prompt.

Throughout the rest of this guide, we will assume that you have renamed the program you want to use, so that it is called simply "`quicktest`", and that you have moved it to somewhere on your path.

The precompiled versions are statically linked, which means that you can run them without having any special libraries installed. Besides the fact that we only provide precompiled versions for a limited range of computer systems, there are at least two further disadvantages to running the precompiled versions. Firstly, they probably will not work with major kernel versions different from 2.6. Secondly, they will not be able to take advantage of any high performance linear algebra libraries on your system (such as ATLAS, or Intel MKL).

If you have difficulty using the precompiled versions of QUICKTEST, we suggest that you either contact us for help, or try compiling your own version using the source code provided.

1.2 Compiling using the source code

You should be able to compile your own version of QUICKTEST on any GNU/Linux system, and perhaps on other UNIX-like systems that have a C^{++} compiler. In the best case scenario, it will be sufficient to type "`make clean`" followed by "`make`" in the directory containing the QUICKTEST sourcecode and Makefile. This should generate a program file (binary) called simply "`quicktest`".

To compile QUICKTEST, you will need some standard libraries, and also a few less standard ones. If you get error messages during the compilation process, it is likely that one or more of the necessary libraries are missing. Specifically, QUICKTEST requires standalone libRmath, and liblapack. It also makes use of the GNU-specific hash_set extension to the C^{++} STL.

If you run a reasonably up-to-date Debian, Ubuntu or related distribution, it should be sufficient to type:

```
sudo apt-get install r-mathlib lapack3-dev
```

to install the necessary libraries.

If you experience difficulties compiling QUICKTEST, please contact us.

2 Quick start for SNPTEST users

If you already have SNPTEST format input files, then it should be easy for you to try out our QUICKTEST software. Suppose you would run SNPTEST on your chromosome 1 data (`chr1.gen.gz`), to analyse a continuous trait detailed in a sample file (`bp.sample`), using the command:

```
snptest -gen_gz -controls chr1.gen.gz bp.sample -qt -proper -o sres.txt
```

Then, to run the same analysis using QUICKTEST, use the following command (all typed on one long line, and all in lower case):

```
quicktest --geno chr1.gen.gz --snptest --pheno bp.sample
                            --method-mean --method-score --out qres.txt
```

Note that the above commands will cause output to be written to two different files, so that you can compare the results.

Both output files begin with a single line of column names, and then have one line of results per SNP. Compared with the SNPTEST output file (`sres.txt`), the QUICKTEST output file (`qres.txt`) has fewer columns, and the column names are different. The columns labelled `normal.score.beta`, `normal.score.se`, and `normal.score.p` contain the effect size estimate, its standard error, and the association p-value, calculated using the same score test that is invoked by SNPTEST's `-proper` option.

We suggest that you also use the `--method-mean` option, because we have found that, for quantitative traits, a test based on mean genotype values has equal power, but a lower false positive rate, than the score test. The columns labelled labelled `normal.mean.beta`, `normal.mean.se`, and `normal.mean.p` are calculated by regressing the phenotype onto the mean genotypes (which are also known as the expected allele dosages). Because the results from the two tests are written side by side in the same output file, it should be easy for you to compare the results.

QUICKTEST has no analogues of SNPTEST's `-qt` and `-gen_gz` options. This is because, for QUICKTEST, quantitative trait analysis is implicit, and because there is no need to specify whether the genotype files are gzipped or not. For QUICKTEST, the `--snptest` option is necessary to tell it that the phenotype file is in the format used by SNPTEST. Specifically, this means QUICKTEST expects the second line of the phenotype file to be something like "`0 0 0 P P P`", and that missing data are coded using "`-9`".

QUICKTEST also implements a number of alternative association testing methods that can perform better than either the score test or the mean genotype test, and which are described in detail in the remainder of this manual. For example, you could try (again, all typed on one long line):

```
quicktest --geno chr1.gen.gz --snptest --pheno bp.sample
                --compute-rSqHat --mixture 2 --method-ML --out qres.txt
```

3 The main features of QUICKTEST

QUICKTEST is dedicated to association analysis between quantitative traits and uncertain genotypes, for which it offers a rich suite of analysis methods. However, it has rather limited facilities for analysis of binary or case/control phenotypes. The main features are that QUICKTEST:

- Reads genotypes in the common file format as used by IMPUTE, SNPTEST, and related software. For convenience, QUICKTEST can successively read multiple genotype files (e.g. for separate chromosomes) and write all results to a single output file, and can also restrict analyses to SNPs from a prespecified list.
- Reads phenotypes in either of two different formats:
  - the file format used by SNPTEST
  - the file format generated by the R function `write.table()`
- Offers a range of statistical methods for dealing with genotype uncertainty:
  - best guess genotypes (a.k.a. the calling method)
  - expected genotype dosages (a.k.a. the mean method)
  - a score test (the "-proper" method of SNPTEST)
  - exact maximum likelihood
  - a Markov chain Monte Carlo method
- Offers two optional methods for dealing with phenotypes that do not have an exactly normal distribution:
  - Quantile–quantile transformation, which uses only the ranks of observations
  - A Gaussian mixture model based method that uses the full data
- Can run permutation tests, either using an adaptively chosen number of permutations, or using the same ensemble of phenotype permutations for all SNPs
- Can test for interaction between each SNP and a covariate. This can be used to test for GxE, or for GxG between all SNPs and a single SNP
- Runs much faster than SNPTEST, for an equivalent analysis
- Is freely available, including source code, under the terms of the GNU GPL

The current limitations are that QUICKTEST:

- Has poor facilities for analysis of binary phenotypes
- Does not allow non-additive genetic models

- With the exceptions of interaction and binary trait analyses, other quantitative trait analyses do not take covariates into account on a SNP by SNP basis. Instead, a univariate "residual phenotype" is constructed once, and then used for all association tests.

4 Accounting for covariates

QUICKTEST requires that the phenotype to be analysed, and any covariates, are supplied as columns in a phenotype file. Nothing in the phenotype file itself specifies which column is the phenotype, or which columns are covariates. (If SNPTEST format files are used, the type-indicating line of the phenotype file is ignored.) The QUICKTEST command line options `--npheno` and `--ncovar` are used to indicate the names of columns to be used for an analysis. This arrangement means that you can have many phenotypes in a single phenotype file, and that you can analyse both association for LDL cholesterol (with BMI as a covariate), and association for BMI (with LDL cholesterol as a covariate), using a single unvarying phenotype file. For example:

```
quicktest --geno example1.gen.gz --snptest --pheno multivar1.sample
                --npheno ldl --ncovar bmi --ncovar sex
                --method-mean --out ldl.txt
```

and

```
quicktest --geno example1.gen.gz --snptest --pheno multivar1.sample
                --npheno bmi --ncovar ldl --ncovar sex
                --method-mean --out bmi.txt
```

As shown above, you can account for multiple covariates by using the `--ncovar` option multiple times. Note that the `--ncovar` option itself is used multiple times, and *not* that a single `--ncovar` option is used with multiple arguments.

All covariates must be numeric. If you need to take into account more complex types of covariates (such as factors), you must use your own statistical software to construct columns of an appropriate and full rank design matrix. The R function `model.matrix()` can be used for this purpose. Note that you should not include the intercept column, because this is always added by QUICKTEST.

For all analyses implemented in QUICKTEST, all covariates specified using `--ncovar` options will be used. However, the exact manner in which they are used varies across the different analysis methods, and is a little bit idiosyncratic.

For all the standard quantitative trait analyses, a residual phenotype is constructed before any assocation tests are carried out. This residual phenotype is then used for all association tests. See manuscript by Kutalik and Johnson for details about what difference this makes.

For the interaction test, and also for binary phenotype analysis, all covariates are included in a multiple regression analysis for each SNP.

To do: implement multiple regression for normal mean method, so that effect of two-step method can be explored directly using QUICKTEST.

5 Exclusions and analyses of subsets of your data

5.1 Analysing a subset of SNPs

We do not provide a facility for excluding particular SNPs from analysis. If there are relatively few SNPs to exlude, we suggest you analyse all SNPs and then remove the ones you don't want from the output file. We do provide a facility to restrict analysis to a list of SNPs, since we generally find it this useful. Use the `--only` option to indicate the file containing ids of SNPs to be analysed. These will be matched against the string in the *second* column of the genotype file, which is conventially used to store an rs number or similar identifier.

5.2 Analysing a subset of individuals

We think that the best way to determine the individuals to be included or excluded from analysis is to exclude individuals by making their phenotypes "`NA`" in the phenotype file. However, we do provide a simple command line facility for excluding individuals from analysis. Use the `--exclude` option to indicate a file containing ids of individuals to be excluded. An individual will be excluded if either of the first two columns in the phenotype file is an exact string match to any of the strings in the exclusion file. This mechanism is very different to the "both must match" mechanism used by PLINK, and it may cause you to inadvertently exclude more individuals than you expected. Consider the following phenotype file:

```
FID IID missing phenotype
  1   1       0      12.6
  1   2       0      13.2
  1   3       0      11.7
  2   1       0      13.4
```

Here, the two ID fields have been used in the traditional pedigree way, to indicate families (FID) and individuals within families (IID). If you have an exclude file with a single line

```
2   1
```

then *all* individuals will be excluded, because all of them match at least one of the exclude strings.

We provide our mechanism because we find that, in association studies, it is more common to use the two ID fields to store two different IDs, each unique. E.g. one is for patient ID, and the second is a sample handling barcode. We wanted to be able to exclude individuals using either labelling system since it is tedious to have to cross reference.

If this causes problems, go with the simple method of making the phenotypes NA.

6 Preparation and format of input files

6.1 Genotype files

QUICKTEST requires information about genotypes to be input in the same format as for the widely used IMPUTE and SNPTEST software. Therefore, if you already have data in this format, you don't need to do anything. It does not matter whether the files are gzipped, or not: QUICKTEST will read either.

The genotype file format, as implemented in QUICKTEST, is the following: The file consists of one line per SNP. On each line, the fields are whitespace-delimited. The first five fields can be arbitrary text, traditionally used to store information about the SNP. These fields are copied verbatim into the output file. These are followed by $(3 \times n)$ fields, arranged as n triples, one for each individual. Each triple contains the probabilities of the genotypes 0, 1, and 2.

Because of the large filesizes, it seems common practice to have separate files for imputed genotypes on different chromosomes. As long as all the files are for the same set of individuals (in the same order), multiple genotype files can be read by QUICKTEST by using the `--geno` option multiple times. All the results will be written to a single file, with a single header line at the beginning. For example:

```
quicktest --geno chr1.gen.gz --geno chr2.gen.gz --geno chr3.gen.gz
                      --pheno bp.dat --method-mean --out results123.txt
```

Note that the `--geno` option itself is used multiple times, and *not* that a single `--geno` option is used with multiple arguments.

If you want to analyse only certain SNPs from the genotype file(s), you can use the `--only` option to tell QUICKTEST the name of a file containing the names of SNPs to analyse. The contents of this file are compared, string-wise, against the *second* field oon each line of the genotype file. Internally, the SNP names are stored in a hash table, allowing fast yes/no decisions and therefore fast program execution, even when the file of SNP names is very large.

6.2 Phenotype files

For compatibility, we have tried to make it easy for QUICKTEST to read phenotype files as used by SNPTEST. Such files have two interesting features. Firstly, there is a line after the header line that is like "`0 0 0 P P P`" with one or more "`P`"s. We call this the ooops line. You have to use the option `--ooops-line` to tell QUICKTEST to expect an ooops line. Secondly, SNPTEST uses "`-9`" to code for missing phenotypes. You have to use the option `--missing-code -9` to tell QUICKTEST this information. For convenience, the single option `--snptest` is equivalent to supplying both these SNPTEST-specific options.

If you are generating phenotype files specifically for QUICKTEST, you might find it tedious to have to insert the ooops line, and to recode your missing data as "`-9`". Therefore, the default behaviour of QUICKTEST is *not* to expect the ooops line, and

that missing data are coded only using "NA". Such files can be written from R using the `write.table()` function, for example.

To reduce the possibility of errors when reading the phenotype file, QUICKTEST detects when multiple individuals have exactly identical phenotype values, and prints warning messages. This should alert you to the possibility that you are coding missing data in a way that QUICKTEST has not recognised. Using the `--missing-code` option, you can specify any single value for your missing data code, but note that matching is done using strings, so that "`-9.0`" is different to "`-9`". The string "`NA`" is always understood to be missing data. If you genuinely have multiple individuals with identical phenotypic values, you can turn off the warning messages using `--ignore-ties`.

You can store multiple phenotypes in a single phenotype file, but QUICKTEST will only analyse one of them at a time. You can select which phenotype using the `--npheno` option, followed by an argument that is the name or number of the column containing the phenotype to be analysed. The default is equivalent to using the option `--npheno 1`.

When interpreting the argument of the `--npheno` option, QUICKTEST first attempts to match the argument against all column names in the phenotype file. If no match is found, QUICKTEST tries to interpret the argument as the number of a column, *after the first three "special" columns of the file*. This means that the default option `--npheno 1` will analyse the phenotype in the leftmost column with name "`1`", if one exists, and otherwise will analyse the phenotype in the fourth column. To avoid possible confusion, it is best to give all your columns distinct names, and to identify them by name and not by number.

QUICKTEST only analyses univariate phenotypes. If you want to allow for covariates, we suggest that you use your own statistical software to regress the phenotype onto the covariates, and then use the residuals from that regression as a univariate "residual phenotype" in QUICKTEST. We call this approach a "two step" approach, in contrast to the "one step" approach of accounting for the covariates simultaneously to performing the association analysis. There are several reasons why we prefer the two step approach:

- Some of the association tests used by QUICKTEST would be hard to program, and very slow to run, if we simultaneously took account of covariates.
- For simple association testing methods (like known genotypes, or mean genotypes), the one step and two step approaches give numerically very similar results.
- You may want to perform complex covariate adjustments (e.g. involving interactions between factors and continuous covariates), and we cannot hope to implement all the functionality of R's `lm()` function into QUICKTEST.

7 Association testing methods

[**To be written**] All the association testing methods compared in [**our paper**] are implemented in QUICKTEST. We allow all methods to be applied to the same data, and all the results from different tests to be written as multiple sets of columns in a single output file.

7.1 Differences between QUICKTEST and SNPTEST output

There are several possible reasons why you might get different output from QUICKTEST and SNPTEST. These include:

- Only the output of the normal score method of QUICKTEST is supposed to be equivalent to SNPTESToutput.
- For QUICKTEST, quantitative trait analysis is the default. For SNPTEST, case/control analysis is the default.
- SNPTEST will correct for covariates in quantitative trait analysis, whereas QUICKTEST will not. Do the differences persist when you analyse a univariate phenotype?
- Check that both programs are reading the phenotype data correctly. Check the reported range of the phenotype, and the number of individuals with non-missing phenotypes. By default, QUICKTEST does *not* treat "`-9`" as missing data.
- SNPTEST sometimes treats uncertain genotypes as certain, if the certainty exceeds some (unspecified) threshold. SNPTEST also filters out results when the Fisher observed information matrix is close to singular. The filtering threshold is undocumented, and the behaviour is not invariant under linear transformation of the phenotype.
- Other possible rounding errors and differences in implementation!

8 File manipulation utilities

[**To be written**] By constructing phenotype files with missing phenotypes for certain individuals, and by ingenious combination with the `--only` and `--copy` options, along with multiple `--geno` options, you can perform various file subsetting and merging operations. There is no way to use QUICKTEST to merge files containing data from different sets of individuals.

9 Testing interactions

QUICKTEST currently offers one method for testing for interaction effects, which is invoked using the `--method-interaction` option. You can test for an interaction effect involving a single fixed covariate, with each SNP in turn in the genotype file(s). The test is only applied using the mean method. The covariate must be in the phenotype file, along with the response phenotype. You use the `--ncovar` option to indicate the name or column number of the covariate, in an exactly analogous way to using the `--npheno` option.

The mechanism just described can be used to test for gene×environment (G×E) interactions, by using an environmental variable such as smoking as the covariate. It

can also be used to test for gene×gene (G×G) interactions, by using a coded genotype as the covariate.

Note that all covariates specified using the `--ncovar` option are taken into account, using a multiple regression approach. However, only the *first* specified covariate is tested for interaction.

Example invocation:

```
quicktest --geno example2.gen.gz --pheno multivar1.pheno
                --npheno ldl --ncovar smoke
                --method-mean --method-interaction
                --out interex.txt
```

10 Binary phenotype analysis

We have implemented a simple logistic regression approach, allowing arbitrary numbers of numerical covariates. The mean genotypes are used for each SNP. This is probably not an optimal approach.

Unlike SNPTEST, we assume that genotypes of all individiuals to be analysed are contained in a single file. We define the binary phenotype by allowing an arbitrary cut-point to be defined for a continuous phenotype. (Thus, if you have a real binary phenotype, code it as 0/1 and use a cut-point of 0.5.)

Example invocation:

```
quicktest --geno example2.gen.gz --pheno multivar1.pheno
                --npheno ldl --ncovar smoke --ncovar age --ncovar sex
                --method-mean --method-binary 4
                --out logistex.txt
```

If you want to analyse using factor covariates, you will have to construct the appropriate columns of indicator variables yourself. Note that QUICKTEST automatically includes an intercept column, so make sure that you do not specify a rank deficient design matrix.

11 Problems, error messages, FAQ

If the QUICKTEST program produces an error or warning message, or if it produces unexpected output, or fails to produce the expected output, please read *all* the messages it writes. Often the problem is more easily diagnosed by reading the first warning or error message, rather than the last.