NAME

QUICKTEST − Quick association testing, for quantitative traits, allowing genotype uncertainty

SYNOPSIS

```
quicktest −−pheno file −−geno file −−out file −−method−mean −−method−ML

quicktest −−pheno file −−npheno name −−ooops−line −−missing−code code −−geno file ...
```

EXAMPLE

```
./quicktest −−pheno ex.sample −−npheno hta −−geno ex.geno −−method−ML −−out ex.out
```

DESCRIPTION

QUICKTEST estimates genetic effect sizes and tests for significance, testing for association between a quantitative trait, and SNP genotypes that are uncertain. It allows (in principle) arbitrary quantitative trait distributions, by modelling the trait distribution as a mixture of normal distributions. It can compute a likelihood ratio test, using an iterative algorithm to find the exact MLE of the missing data likelihood [KJB+08]. QUICKTEST can also compute several other tests, including a score test, and simple regression onto mean genotypes.

OPTIONS

The complete list of available options is as follows.

−−call−thresh *value*

Sets the probability threshold used for calling genotypes to *value*. Default value is 0.9. Only used if the **−−copy−calls**, **−−method−call** or **−−test−HW** options are used.

−−compute−alphaHat

Compute a simple method-of-moments estimator for alpha. This assumes the probability vectors are i.i.d. Dirichlet over individuals and defines alpha to be the sum of the parameters of that Dirichlet distribution. This option is of very obscure interest.

−−compute−MAF

Compute the (expected) minor allele frequency for each SNP/locus. This option is provided simply for convenience, since the (expected) MAF can be computed easily from the (expected) genotype counts, this is technically redundant. Not using this option will slightly reduce output file size.

−−compute−rSqHat

Compute r−squared-hat for each SNP, which is the (estimated) fraction of variance in unobserved 0/1/2 genotype explained by the the individual mean genotypes. (We assume this is the same definition used by Abecasis et al.)

−−copy *file*

For SNPs/loci and individuals analysed, write a copy of the genotype probabilities to *file*. This can be used to gunzip and extract subsets from the genotype probability file, column-wise by using the **−−only** option, and/or row-wise by using a sample file with missing phenotypes,

−−copy−calls

This option alters the behaviour of the **−−copy** option. Instead of outputting genotype probabilities, QUICKTEST will output called (best guess) genotypes, using the probability threshold set using the **−−call−thresh** option. The output is identical to the .tped format used by PLINK. You should be able to reformat the .sample/.pheno file to make a matching .tfam file.

−−exclude *file*

An exclusion list is constructed by reading all strings in *file*. Strings are whitespace delimited, so multiple strings on a single line are treated identically to if they were on separate lines. For each individual, if either of the first two fields in the phenotype file match any of the strings in *file*, the individual is excluded. Matching must be stringwise exact. Note that this is very different to the linewise matching using by PLINK.

−−geno *file*

Specifies a name of an input genotype file. If the file has been compressed using gzip, QUICKTEST will detect this and uncompress it on the fly. If this option is used multiple times, the files will be read in the order they are specified, and the results concatenated into a single output and/or copy file.

−−ignore−ties
By default, QUICKTEST will print a warning message if multiple individuals have identical (tied) phenotypic values. The intended behaviour is to detect when a special value (like `−9`) is being used to code missing data, but that the **−−missing−code** option has not been used to specify the special value. Use this option only if you have genuinely tied values and you want to disable the warning message.

−−method−binary *value*
Carries out a logistic regression, by dichotomising the phenotype at split point *value*. Mean genotypes are used, and all covariates specified with **−−ncovar** are included in the analysis.

−−method−call
Will carry out an analysis by calling the genotypes, with a threshold that can be set using the **−−call−thresh** option.

−−method−interaction
Will carry out an analysis for SNPxCovariate interaction, using the first covariate specified by the first use of the **−−ncovar** option. Other covariates are included in the analysis, but not as interaction terms. The statistical test is analogous to the mean genotype method.

−−method−ML
Will carry out an analysis by computing the exact MLE, using the iterative algorithm, and then a LR test, as described by [KJB+08].

−−method−mean
Will carry out an analysis using mean genotypes, i.e. the so-called expected dosage method.

−−method−robust
Adjusts the behaviour of **−−method−interaction**. The beta coefficients and standard errors for both the SNP and SNPxCovariate regression terms will be output, along with the error covariance between the two. The Huber sandwich estimate is used to provide a robust estimate of the error variance-covariance matrix. Code implementing this method was written by Alisa Manning and Han Chen (see AUTHORS).

−−method−score
Will carry out an analysis using the missing data likelihood method of [MHM+07], equivalent to the **−proper** option in the SNPTEST software.

−−method−MCMC
Will carry out an analysis using Markov chain Monte Carlo simulations to average over the missing data distribution. Do not confuse this option with the options for Monte Carlo permutation tests!

−−method−ranks
This option is currently disabled. It will use mean genotypes, but will carry out an analysis using a linear model with an (approximate) marginal likelihood based on the ranks of the phenotypes, as described by [Pet82]. At least for large sample sizes, we find that **−−method−mean −−quantile−norm** gives very similar results and is faster.

−−missing−code *value*
Phenotype values exactly matching *value* will be treated as missing. The matching is done with strings and must be exact, so `−9` and `−9.` are different. The default is that this option is not set. Regardless of this option, QUICKTEST always assumes that NA codes for missing data.

−−mixture *number*
Attempt to model errors using a mixture of *number* Gaussians. This allows heavy tailed (or light tailed) error distributions, and makes analyses of imputed genotypes more reliable. Currently only implemented for **−−method−score**, **−−method−mean** and **−−method−ML**. This requires a pre-analysis to fit the mixture model to the phenotype distribution. This can be slow, because we find it necessary to restart the EM algorithm from multiple random initial points. If repeatedly analysing

the same phenotype, it is possible to save and then reuse an already fitted model, see **−−mix−hint**.

−−mix−centered
Constrains the mixture model such that all components have zero mean.

−−mix−hint *file*
Parse *file* to determine initial values of mixture parameters, and then run EM algorithm just once, using those initial values, to estimate the mixture distribution. The initial value specifications in *file* must take forms like `pi1 = 0.333` `mu2 = −0.12` `sigma3 = 1e−3` etc. WITH SPACES AROUND THE EQUALS SIGNS. Any content in *file* that cannot be parsed is ignored. The simplest way to create a suitable file is to cut the relevant lines from the screen output of a previous run.

−−no−normal
When either quantile-quantile normalisation or mixture model methods are used, do not run analyses for the simple normal model. This option saves computation time.

−−no−progress
Do not print a refreshing line reporting progress.

−−ncovar *value*
Specifies that a column in the phenotype file is to be used as a covariate in the analysis. The argument *value* is handled in the same way as for **−−npheno**. The **−−ncovar** option can be used multiple times to specify multiple covariates, but at present only the first covariate is used for interaction analysis.

−−npheno *value*
Specifies which column in the phenotype file to use for analysis. If *value* matches any token in the header, the corresponding column is used. Otherwise, *value* is interpreted as a number, and the (*value*+3)−th column will be used. The default *value* is 1, meaning to analyse the phenotype in the 4−th column of the phenotype file.

−−only *file*
Perform all operations only for SNPs/loci corresponding to those in *file*. Entries in *file* should specify the identifier in the second column of the genotype probability file, which is usually used for rs number.

−−ooops−line
Assume that the second line of the phenotype file is a `0 0 0 P P ...` line and should be ignored. Using this option means that SNPTEST phenotype files can be read.

−−out *file*
The name of the output file.

−−perm−Besag−Clifford *value1 value2*
Perform a Monte Carlo permutation test for each SNP, using the mean genotype method. The simple closed sequential scheme of [BC91] is used. Permutations are performed until either (i) there have been *value1* occurrences of a statistic more extreme than the observed value, or (ii) *value2* permutations have been performed. This permutation approach is recommended if you wish to control the false positive rate, when the assumptions of the normal linear model do not hold.

Under the null, about *value1*(1+log(*value2*/*value1*)) permutations per SNP are needed on average.

−−perm−verbose *value1*
Perform *value1* Monte Carlo permutations for each SNP, using the same set of permutations over all SNPs. The p−value, computed using the mean genotype method, is written to the output file for each permutation. The output may be subsequently analysed to estimate an empirical p−value, or using various methods for controlling or estimating the genome-wide false discovery rate (FDR).

Warning: The output file may be large.

−−pheno *file*
The name of the input phenotype file.

−−qqnormal
Quantile normalise data before analysis.

−−sim−num *number*
The number of genotype simulations to perform for each SNP, if the MCMC method is to be used.

−−sim−seed *int1 int2*
The seeds for the random number generator. QUICKTEST links against libRmath, and calls `set_seed(int1,int2)` to seed the generator. According to my documentation, the generator is the Marsaglia-multicarry.

−−snptest
Equivalent to **−−ooops−line −−missing−code −9**.

−−test−beta *value*
Allows evaluation of the performance of different methods, by simulating an effect of size *value*. Technically, a different synthetic phenotype is computed for each SNP, by simulate genotypes according to imputation probabilites and then multiplying by *value*. The errors are taken from a random permutation of the input phenotype. Note that a different permutation is used for each SNP. The synthetic phenotype is then analysed using all **−−method−XXX** options selected, as if the simulated genotypes had not been observed. If the **−−mixture** option is used, the mixture model is refitted to the synthetic phenotype before analysis. The magic estimates in the output file are the estimates that would have been obtained if the the simulated genotypes had been observed.

−−test−HW
Will perform the exact test of [WCA05], using the best guess genotypes determined using the probability threshold set by the **−−call−thresh** option.

RETURN VALUE

The program returns 0 upon successful completion. Other return values indicate a problem with the command line arguments or input files, and are accompanied by a diagnostic message.

FILE FORMATS

THE PHENOTYPE FILE

Each line of the phenotype file corresponds to one individual. There can either be a single header line (the default), or two header lines are allowed (as used by SNPTEST) if the **−−ooops−line** option is used. Missing values should be specified by writing `NA`, but an additional missing code can be specified using the **−−missing−code** option.

In summary, to read SNPTEST phenotype files, use:

−−ooops−line −−missing−code −9 or equivalently **−−SNPTEST**

It is absolutely essential that the order of individuals in this file is the same as the order of individuals in the genotype file (see below).

THE GENOTYPE FILE

Each line of the genotype file corresponds to one SNP. The first five fields contain information about the SNP. These are not used by QUICKTEST, but are copied verbatim into the output file. Subsequent fields contains the probabilities of each genotype for *n* individuals. For each individual, three consecutive fields give the probabilities of genotype 0, 1 and 2.

This file format is identical to the file format used by the **CHIAMO**, **IMPUTE** and **SNPTEST** programs that implement the methods of [MHM+07].

Genotypes of individuals with phenotypes written as missing (using `NA` or `−9`) must be included in the genotype file, and are ignored by the program.

DEFAULT CONVERGENCE CRITERIA

For uncertain genotypes and a mixture error distribution, we declare convergence of the EM algorithm if (i) the sum log ratio of parameters being estimated is less than 1e−6, or if (ii) the increase in log likelihood is less than 1e−4 per iteration, averaged over 100 iterations. If more than 500 iterations occur, we declare non-convergence.

DIAGNOSTICS

Email us if you do not understand the messages output by the program.

EXAMPLES

See the files *ex.sample*, *ex.gen.gz*, and *run.sh*.

REFERENCES

[BC91] Besag and Clifford (1991) Sequential Monte Carlo p−values. *Biometrika* **78**(2):301−4.

[KJB+09] Kutalik, Johnson, Bochud, Mooser, Vollenweider, Waeber, Waterworth, Beckmann and Bergmann (2009?) A comparison of methods for testing association between uncertain genotypes and quantitative triats. Manuscript submitted.

[JK09] QUICKTEST: Quick testing of genetic assocation for quantitative traits. In preparation.

[MHM+07] Marchini, Howie, Myers, McVean and Donnelly (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genetics* **39**:906−913.

[OF04] O'Hagan and Forster (2004) *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*. Arnold, London, 2nd edition.

[Pet82] Pettitt (1982) Inference for the Linear Model using a Likelihood based on Ranks. *J. R. Statist. Soc. B* **44**(2):234−243.

[SS07] Servin and Stephens (2007) Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genetics* **3**(7):e114. doi:10.1371/journal.pgen.0030114

[WCA05] Wigginton, Cutler and Abecasis (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**:887−93.

FILE SUBSETTING AND MERGING

QUICKTEST can be used to extract subsets from genotype probability files. To choose a subset of SNPs or loci, use the **−−only** option. To choose a subset of individuals, make a dummy phenotype file where the phenotypes are non-missing for the individuals you want, and missing otherwise. Combine the two to extract a subset of loci and individuals. Use the **−−copy** option to write the subset of genotype probabilities to a new file.

Use multiple instances of the **−−geno** option to merge multiple input files, assuming they describe the same individuals in the same order.

The **−−only** option builds a hash table of identifiers, and thus is reasonably fast even with very large lists.

BUGS

There are probably bugs. Although all functions have been tested, we have reorganised the internal data structures and the way that many functions are called, so some functions may have been broken.

We continue to develop the software. Please check that you are using the most up-to-date version.

It would be nice to implement more general coding schemes, 2 degree of freedom tests, Bayes factors, missing data likelihood for interactions, generalised error variance (estimating equation) methods, etc.

AUTHORS

Toby Johnson Toby.Johnson@unil.ch

Zoltan Kutalik Zoltan.Kutalik@unil.ch

Alisa Manning amanning@bu.edu

Han Chen hanchen@bu.edu

QUICKTEST uses gzstream, by Deepak Bandyopadhyay and Lutz Kettner.

QUICKTEST uses SNP-HWE, by Jan Wiggington.

COPYRIGHT

QUICKTEST uses gzstream, which is Copyright (C) 2001 Deepak Bandyopadhyay, Lutz Kettner.

The programs, source code, and documentation, can be copied *only* under the terms of the licences below.

LICENCE AND WARRANTY

With appropriate citation, the SNP-HWE routines are freely available for your use and can be incorporated into other programs.

The GNU General Public Licence can be viewed at http://www.gnu.org/licenses/gpl.html.

The GNU Lesser General Public Licence can be viewed at http://www.gnu.org/licenses/lgpl.html.