IQLS Software Documentation

Version 1.2

Zuoheng Wang[1*] and Mary Sara McPeek[1,2]

Departments of [1]Statistics and [2]Human Genetics
The University of Chicago

[*]Present address: Division of Biostatistics,
Yale School of Public Health

This work was supported by National Institutes of Health grant R01HG001645.

IQLS (Incomplete-Data Quasi-likelihood Score Test)

A C++ program for haplotype frequency estimation and case-control association testing between a binary trait and haplotypes of multiple tightly-linked SNPs, where the individuals in the sample can be related.

Homepage: http://galton.uchicago.edu/∼mcpeek/software/IQLS

Release 1.0 November 12, 2009
1.1 December 18, 2009
1.2 March 9, 2011

==========

License

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

- For full-df association tests and haplotype frequency estimation: Wang, Z., and McPeek, M.S. (2009). An incomplete-data quasi-likelihood approach to haplotype-based genetic association studies on related individuals. J. Am. Stat. Assoc. 104, 1251-1260.
- For 1-df association tests: Wang, Z., and McPeek, M.S. (2009). ATRIUM: Testing untyped SNPs in case-control association studies with related individuals. Am. J. Hum. Genet. 85, 667-678.

To contact the first author:

Zuoheng Wang
Division of Biostatistics
Yale School of Public Health
60 College St, Rm 200
New Haven, CT 06510

email: zuoheng.wang@yale.edu

Contents

1 Overview of IQLS

IQLS is a program, written in C++, for haplotype frequency estimation and case-control association testing between a binary trait and haplotypes of multiple tightly-linked SNPs, in samples that contain related individuals.

The program performs haplotype frequency estimation and both 1-degree-of-freedom and full-degree-of-freedom haplotype association tests for a sliding haplotype window of *nmark* SNPs, where *nmark* is an integer specified by the user. The sliding haplotype window can be applied over a whole chromosome, or the user can specify the beginning and end of the region to be analyzed. The program is applicable to association studies with completely general combinations of related and unrelated outbred individuals, where the relationships among the sampled individuals are assumed to be known.

The reference for the IQLS haplotype estimation methods and full-df IQLS haplotype association tests is Wang and McPeek (2009a), while the reference for the 1-df IQLS haplotype association tests is Wang and McPeek (2009b) [specifically, the 1-df IQLS haplotype association test is test (3) in subsection "Comparison to other approaches for related individuals" of Wang and McPeek (2009b)]. The IQLS haplotype estimation and testing methods can be considered extensions of the single-marker estimation and testing methods of McPeek, Wu, and Ober (2004) and Thornton and McPeek (2007), respectively.

For each set of consecutive SNPs in a sliding haplotype window, the IQLS program computes haplotype frequency estimates and both 1-df and full-df tests for haplotype association with the case-control phenotype. IQLS uses quasi-likelihood score testing and estimation methods that allow for partially missing-data: missing phase information and/or missing marker genotype data for some markers in the haplotype. By default, the MIQLS_b test (which was found to be the most powerful by Wang and McPeek 2009a) is used. Alternatively, the program can be run with the faster option, which uses MIQLS_a instead of MIQLS_b. The difference between MIQLS_a and MIQLS_b is that in MIQLS_a, the haplotype information for an individual is based only on that individual's own genotype data, while in MIQLS_b, more haplotype information is extracted by also considering parental genotype data when available. For a more detailed comparison of the 2 statistics, see Wang and McPeek (2009a). In either case, the p-value is calculated based on a χ^2 asymptotic null distribution.

To calculate the MIQLS statistic, an estimate of the prevalence of the trait in a suitable reference population must be specified by the user. We emphasize that the test will be valid regardless of the input value (see Thornton and McPeek 2007 for details). We recommend using an estimate from previous studies or registry data from a suitable reference population.

Additional features of the IQLS tests include:

(1) The IQLS tests do not require phased genotypes nor do they infer haplotypes. Instead, genotype information is directly incorporated into the haplotype analysis, and uncertainty in the phase information is accounted for.

(2) The IQLS tests improve power by taking advantage of the principle that there is

enrichment for predisposing variants in affected individuals with affected relatives.

(3) The IQLS tests allow for clustering of haplotypes to reduce the degrees of freedom. Rare haplotypes can be grouped with more common haplotypes that differ at a single site until every haplotype cluster has expected cell count at or above a user-specified threshold (see the Input section).

2 Running IQLS

Installation instructions:

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

2. Read the file IQLS_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 "IQLS".

5. To run the executable program IQLS:

   First, prepare the input files, e.g., pedigree, markid, ibdcoef, and parameter (see Section 3 for more details).

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

   ./IQLS

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

   ./IQLS -pheno pedigree -geno markid -ibd ibdcoef -r parameter -b 101955 -e 230788 -op b -thresh 15

   -pheno pedigree Allows the user to specify the name of the phenotype data input file, which also includes pedigree information. The filename defaults to "pedigree" if this flag is not used. To specify a different filename, replace "pedigree" with the appropriate filename.

   -geno markid Allows the user to specify the name of the marker data input file. Filename defaults to "markid".

   -ibd ibdcoef Allows the user to specify the name of the IBD coefficient input file. Filename defaults to "ibdcoef".

-r parameter Allows the user to specify the name of the parameter input file, which contains: (1) an estimate of the prevalence of the binary trait in a suitable reference population, and (2) *nmark*, which is the number of markers in each haplotype. (A sliding haplotype window of *nmark* markers will be applied over the specified region.) Filename defaults to "parameter".

-b 101955 Allows the user to specify the start position (in nucleotides) of the region over which the sliding haplotype window is used. Defaults to the position of the first marker in the marker data file. The value of the start position must be less than the value of the end position.

-e 230788 Allows the user to specify the end position (in nucleotides) of the region over which a sliding haplotype window is used. Defaults to the position of the last marker in the marker data file. The value of the start position must be less than the value of the end position.

-op b Allows the user to choose option (a) instead of option (b). The default, option (b), provides more powerful tests, but can be slower to run, while option (a) is generally faster but less powerful. In option (a), the haplotype information for an individual is based only on that individual's own genotype data, while in option (b), more haplotype information is incorporated by also considering parental genotype data when available. The default value is b. To choose option (a), use the flag: -op a

-thresh 15 Allows the user to specify a threshold (which must be a nonnegative integer) for haplotype clustering to reduce the degrees of freedom in the full-df IQLS haplotype test. Rare haplotypes are grouped with more common ones that differ at a single site until every haplotype cluster has an expected haplotype count that is no less than this threshold. In addition, the 1-df IQLS haplotype association tests are performed only for haplotypes for which the expected haplotype counts are no less than this threshold. The default value is 15.

6. Results of all haplotype tests, based on a sliding haplotype window of *nmark* markers over the specified region, are output to a file called "IQLStest.out". The top 20 haplotype windows with the smallest p-values based on the full-df IQLS association test are output to a file called "IQLStest.fdf.top". Additionally, the top 20 haplotype windows with the smallest minimum p-values, where the minimum is taken over all 1-df haplotype association tests in the window, are also output to a file called "IQLStest.1df.top". The number of haplotype windows output to IQLStest.fdf.top and IQLStest.1df.top can be decreased or increased by the user by changing MAXTOP (currently set to 20) in the peddata.cc source file.

7. You can test the executable program IQLS by running it with the sample input files: pedigree, markid, ibdcoef and parameter. You can then compare the resulting output, which will be printed to the files IQLStest.out, IQLStest.fdf.top and IQLStest.1df.top, with the correct output provided in the sample output files IQLStest.out.ex, IQLStest.fdf.top.ex and IQLStest.1df.top.ex.

8. The programs stop if any errors are detected in the format of the input files.

3 Input

Required Input Files:

1. **phenotype data file** (default filename is "pedigree")

   This file contains the pedigree and phenotype information. Individuals who are not listed in this file will not be included in the analysis. The columns in the file should be organized as follows:

1	1	7	6	1	1
1	2	7	6	2	2
1	3	7	6	1	2
1	7	0	0	1	1
1	6	0	0	2	0
2	11	18	19	2	2
2	12	18	19	1	1
2	18	0	0	1	0
2	19	15	16	2	1
2	15	0	0	1	0
2	16	0	0	2	0
(1)	(2)	(3)	(4)	(5)	(6)

   Column (1) family ID (positive integer)
   Column (2) individual ID (positive integer; must be unique)
   Column (3) father's ID (0 if the individual is a founder)
   Column (4) mother's ID (0 if the individual is a 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 by giving each such person their own unique family ID (as well as unique individual ID) and setting both parents' IDs to 0. There is no limit on the number of individuals nor on the number of families. Each individual should be entered only once. The individual ID is required to be unique (e.g. it cannot be reused in a different family). Individuals from the same family should appear in a single cluster, though there is no requirement on the order of individuals within a family nor on the order in which different families are listed.

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

   ./IQLS -pheno myphenofile

2. **marker data file** (default filename is "markid")

   This file contains the marker data. All markers should be on the same chromosome. (To analyze more than one chromosome, a separate run must be performed for each chromosome, with each chromosome having its own marker data file.) The columns in the file should be organized as follows

marker	chrom.	pos'n	orient.	allele0	allele1	1	2	3	7	6	11	12	18	19	15	16
rs7909677	10	101955	+	A	G	AG	AA	AA	AA	AG	AG	GG	GG	AG	AG	AG
rs9419560	10	142201	+	A	G	AA	GG	AG	AG	AG	AG	NN	GG	AG	AA	GG
rs9419419	10	153707	-	T	C	TC	TC	CC	TC	TC	TT	TC	TT	TC	TT	CC
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)

   Column (1) marker rs number
   Column (2) chromosome
   Column (3) physical position (in nucleotides)
   Column (4) strand orientation ('+'=same strand as HapMap, '-'=opposite strand from HapMap)
   Column (5) nucleotide for allele 0
   Column (6) nucleotide for allele 1
   Columns (7)... marker genotypes (NN for missing genotype)

   The first row of the file must contain the column headings. The headings for the first 6 columns can be arbitrary, but should not contain any space characters. Columns 7 and beyond contain marker genotype data for the sampled individuals, and each of these columns must have the corresponding individual's ID number as the heading. The order of the individuals is not required to be the same as the order in the pedigree file. The column headings must specify the order. There is no limit on the number of markers. However, all markers should be on the same chromosome and should be listed in increasing order of their position on the chromosome. All individuals in the marker data file should also appear in the phenotype data file, otherwise, they will not be included in the analysis.

   The number of columns should be the same for every marker: Use NN for missing genotype.

   The default filename is "markid". To specify a different filename, use the command-line flag -geno followed by the filename. For example, to use a marker data file called "mymarkfile" you could type the command

   ./IQLS -geno mymarkfile

3. **IBD coefficient file** (default filename is "ibdcoef")

   This file contains condensed identity coefficients for every pair of eligible individuals within each family (including an individual with himself/herself), where an individual is eligible if he or she has either (1) known affection status or (2) non-missing genotype for at least one marker. (E.g. an individual with unknown phenotype is still eligible if he or she has any non-missing genotype information.)

IBD coefficients should be included for every pair of eligible individuals who have the same family ID (including each individual with himself/herself). A sampled individual who does not share a family ID with anyone else in the sample, would be represented in the markid file by a single line that gives the IBD coefficients for the person with himself/herself.

The IBD coefficient file has the following format:

1	1	0	0	0	0	0	0	1	0	0
1	2	0	0	0	0	0	0	0.25	0.5	0.25
1	3	0	0	0	0	0	0	0.25	0.5	0.25
1	7	0	0	0	0	0	0	0	1	0
1	6	0	0	0	0	0	0	0	1	0
2	2	0	0	0	0	0	0	1	0	0
2	3	0	0	0	0	0	0	0.25	0.5	0.25
2	7	0	0	0	0	0	0	0	1	0
2	6	0	0	0	0	0	0	0	1	0
3	3	0	0	0	0	0	0	1	0	0
3	7	0	0	0	0	0	0	0	1	0
3	6	0	0	0	0	0	0	0	1	0
7	7	0	0	0	0	0	0	1	0	0
7	6	0	0	0	0	0	0	0	0	1
6	6	0	0	0	0	0	0	1	0	0
11	11	0	0	0	0	0	0	1	0	0
11	12	0	0	0	0	0	0	0.25	0.5	0.25
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)

Column (1) individual 1 ID
Column (2) individual 2 ID
Columns (3)...(11) condensed identity coefficients 1 through 9 between individuals 1 and 2 (Ken Lange's book, Mathematical and Statistical Methods for Genetic Analysis, has a good description of condensed identity coefficients)

Note that IQLS currently permits only outbred individuals in the analysis. For a pair of outbred individuals, columns (3)-(8) will always be 0, and columns (9), (10) and (11) represent the probabilities of sharing 2, 1 or 0 alleles IBD, respectively. If, for any pair of individuals, at least one of the values in columns (3)-(8) is not zero, these individuals will be excluded from the analysis. The individual IDs in this file should correspond to the individual ID's used in the phenotype data file.

The software program that can be used to obtain IBD coefficients is the **IdCoefs** software by Mark Abney, which can be found at
http://home.uchicago.edu/~abney/abney_web/Software.html

The IdCoefs software computes condensed identity coefficients for pairs of individuals within each family. The output of IdCoefs can be directly used as input to IQLS.

The default filename is "ibdcoef". To specify a different filename, use the command-line flag -ibd followed by the filename. For example, to use an IBD coefficient file called "myibdfile" you could type the command

./IQLS -ibd myibdfile

4. **parameter file** (default filename is "parameter")

   This file contains two numbers: (1) an estimate of the prevalence of the binary trait in an appropriate reference population. This prevalence value is used in the calculation of the MIQLS statistic. This should not be prevalence in the case-control sample, but rather the "general population" prevalence for an appropriate reference population. (2) *nmark*, which is the number of markers in each haplotype. (A sliding haplotype window of *nmark* markers will be applied over the specified region.)

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

   ./IQLS -r myprev

Optional Input:

5. **start position of the region in nucleotides** (default value is the position of the first marker in the marker data input file)

   This value is the start position (in nucleotides) of the region over which the sliding haplotype window is used. The default start position is the position of the first marker in the marker data file. The value of the start position must be less than the value of the end position. To change the start position, use the command-line flag -b followed by the position. For example, to start from the second marker, first look in the marker data input file to find the position of the second marker in nucleotides — say it's 142201 — then you could type the command

   ./IQLS -b 142201

6. **end position of the region in nucleotides** (default value is the position of the last marker in the marker data input file)

   This value is the end position (in nucleotides) of the region over which the sliding haplotype window is used. The default end position is the position of the last marker in the marker data input file. The value of the start position must be less than the value of the end position. To change the end position, use the command-line flag -e followed by the position. For example, to end at the second-to-last marker, first look up the position of this marker in the marker data input file — say it's 153707 — then you could type the command

   ./IQLS -e 153707

7. **speed vs. power option** (default value is b)

   The default, option (b), provides more powerful tests, but can be slower to run, while option (a) is generally faster but less powerful. In option (a), which uses MIQLS_a tests and IQL_a haplotype frequency estimates, the haplotype information for an individual is based only on that individual's own genotype data, while in option (b), which uses MIQLS_b tests and IQL_b haplotype frequency estimates, more haplotype information is extracted by also considering parental genotype data when available. The default option is (b). To use option (a), you could type the command

   ./IQLS -op a

8. **haplotype clustering threshold value** (default value is 15)

   This value is the threshold (which must be a nonnegative integer) for haplotype clustering to reduce the degrees of freedom in the full-df haplotype test. Rare haplotypes are grouped with more common ones that differ at a single site until every haplotype cluster has an expected haplotype count that is no less than this threshold. The 1-df IQLS haplotype association tests are performed only for haplotypes for which the expected haplotype counts are no less than this threshold.

   The default value is 15. To change this value, use the command-line flag -thresh followed by the threshold value. For example, to use a threshold value of 20, you could type the command

   ./IQLS -thresh 20

4 Output

1. **IQLStest.out** is the primary output file. It contains
   - Summary of phenotype file information: total number of individuals in phenotype file, number of independent families, number of individuals in each phenotype class (affected/unaffected/unknown)
   - Number of individuals in the genotype file and number of individuals who appear in both the genotype and phenotype files.
   - Prevalence value used in the MIQLS calculations.
   - *nmark*, the number of SNPs in the sliding haplotype window
   - For each haplotype window,
     - rs numbers of the SNPs in the window
     - Among those individuals with at least one SNP genotyped in the window, the numbers who are affected, unaffected, and of unknown phenotype, respectively,
     - If any haplotype clustering was done for the full-df haplotype association test (or even if no clustering was done but some haplotypes were dropped from the analysis because their estimated frequencies were below .00005), then the final haplotype clusters are reported.

- value of the full-df MIQLS statistic and corresponding p-value using the chi-squared null distribution,
- values of the 1-df MIQLS statistics and corresponding p-values for testing each individual haplotype
- haplotype frequencies in the full, affected, unaffected, and unknown samples, estimated using the IQLS method (Note: if any sample size is below 15, the corresponding frequency will not be estimated)
- haplotype frequencies in the full, affected, unaffected, and unknown samples, estimated using the EW method (Note: if any sample size is below 15, the corresponding frequency will not be estimated)

2. **IQLStest.fdf.top** lists the 20 haplotype windows with the smallest p-values based on the full-df MIQLS test. For each haplotype window, only the rs numbers of the SNPs in the window and p-value of the full-df MIQLS are given in IQLStest.fdf.top. The number of haplotype windows output to this file can be decreased or increased by the user by changing MAXTOP (currently set to 20) in the peddata.cc source file.

3. **IQLStest.1df.top** lists the 20 haplotype windows that have the smallest minimum p-values for the 1-df tests. For the ith haplotype window, consider the p-values from all the 1-df tests for that window, and let the minimum of these be p_i. Then IQLStest.1df.top reports the SNP haplotype windows having the 20 smallest values of p_i. For each haplotype window, only the rs numbers of the SNPs in the window and the minimum 1-df p-value are given in IQLStest.1df.top. The number of haplotype windows output to this file can be decreased or increased by the user by changing MAXTOP (currently set to 20) in the peddata.cc source file.

5 Tips

1. Input

   The program will stop if errors are detected in the formats of any of the input files. Please read Section 3 carefully and make sure the input files are in the correct format and have concordant information.

2. Parsimony in set of haplotypes considered in the analysis

   For haplotype analysis, it is common that not every possible haplotype actually occurs in the data. Whether or not a haplotype occurs in the data is often not directly observable when only unphased genotype data are available on the sample. Inclusion, in the IQLS analysis, of haplotypes with sample frequencies close to 0 can cause numerical issues (e.g. in the inversion of the covariance matrix). We resolve this problem as follows: we first use the IQLS method to estimate the haplotype frequency distribution in the entire case-control sample (including any individuals of unknown phenotype). If the estimated frequency for a haplotype of tag SNPs is below some threshold (currently set to $< .00005$), then this haplotype is dropped and is not considered further in the analysis.

6 References

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

2. Thornton, T., and 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.

3. Wang, Z., and McPeek, M.S. (2009a). An incomplete-data quasi-likelihood approach to haplotype-based genetic association studies on related individuals. J. Am. Stat. Assoc. 104, 1251-1260.

4. Wang, Z., and McPeek, M.S. (2009b). ATRIUM: Testing untyped SNPs in case-control association studies with related individuals. Am. J. Hum. Genet. 85, 667-678.