

Combined Linkage and Association Mapping of Quantitative Trait Loci (CLAM_QTL), Version 1.0

Ruzong Fan, Department of Statistics, Texas A&M University, April 2007

1 Overview

This document describes a C++ package to implement the methods and models for Combined Linkage and Association Mapping (CLAM) of quantitative trait loci (QTL). Section 2 briefly describes the installation and run the program. Section 3 explains how to run the program for population data with one example. Section 4 describes how to analyze pedigree data with one example.

Our group has been working on combined linkage and association mapping of QTL by variance component models, and this program targets to implement the models and methods we have developed. In writing this program, we adopt some features and structures of programs QTDT and MERLIN by Dr. Abecasis. **Actually, we have used QTDT-2.5.1 as the basis of our codes: some codes of QTDT-2.5.1 are removed since they are irrelevant to our work, some codes of QTDT-2.5.1 are kept and modified to implement our models. In particular, we use codes of libsrc in QTDT-2.5.1 to support our codes.** To analyze pedigree data, we recommend to use MERLIN to calculate ibd information and kinship coefficients first; this program **does NOT** provide them.

The theoretical basis for this program is given in our research papers in **References**. Please refer the appropriate ones if you use in any published work. In case of suggestions and questions and/or problems, you can contact us via e-mail (rfan@stat.tamu.edu).

2 Installation and Run the Program

The package is good in Linux. First, download the package “CLAM_QTL.zip”, and download libsrc in QTDT-2.5.1 from <http://www.sph.umich.edu/csg/abecasis/QTDT/download/index.html>. Use the following steps on Linux to install package:

- Put CLAM_qtl in a directory
- Generate your pedigree and data file (see data format in subsections 3.1 and 4.1 for detail)
- The options of command line are given in Table 1.

3 How to Run the Program for Population Data

The theoretical basis is based on Fan and Xiong (2002) and Fan et al. (2006).

3.1 Data Format

This program requires a pair of matched pedigree and data files. In pedigree file, each line describes a single individual, and includes a family identifier, a personal identifier, paternal and maternal identifiers, gender, marker information, quantitative traits and covariates. The data file describes the organization of the pedigree file. Possible types of pedigree data include marker genotypes (M), traits (T) and covariates (C). Each data item is described on a separate line. The data file must match with pedigree file.

3.2 Example

In subdirectory `data_pop`, we use the pedigree data “`sim_3_mrk.ped`”. It includes 12 items: 4 ID items, sex, 3 markers, one trait:

fam	id	dad_id	mom_id	sex	mrk1_1	mrk1_2	mrk2_1	mrk2_2	mrk3_1	mrk3_2	qtl
1	1	0	0	1	2	2	2	2	2	2	-1.08935
2	1	0	0	2	2	2	1	2	1	2	3.50249
3	1	0	0	2	1	1	1	2	1	2	0.90317
...											

The “`sim_3_mrk.dat`” is as follows

```
M SNP_1
M SNP_2
M SNP_3
T sim_phen
```

Using all three markers in the analysis to test the genotype effect (or both additive and dominance effect), run the program by

```
CLAM_qtl -d sim_3_mrk.dat -p sim_3_mrk.ped -x-99.999 -acad -w-
```

the output contains the following results for genotype effect:

The following models will be evaluated...

```
NULL MODEL
```

```
Means = Mu
```

```
FULL MODEL
```

```
Means = Mu + X_add + Z_dom
```

Estimating allele frequencies... [using founder genotypes]

```
SNP_1 SNP_2 SNP_3
```

Lumping alleles with frequencies of 0.05 or less...

Testing trait: sim_phen

=====

Allele	df(0)	RSS(0)	df(add+dom)	RSS(add+dom)	F	(df1, df2)	P-value	Sample-size
N/A	199	374.317	193	324.565	4.93	(6, 193)	0.0001	200

Using all three markers in the analysis to test the additive effect, run the program by

CLAM_qtl -d sim_3_mrk.dat -p sim_3_mrk.ped -x-99.999 -aca -w-

the output contains the following results for additive effect:

The following models will be evaluated...

NULL MODEL

Means = Mu

FULL MODEL

Means = Mu + X_add

Estimating allele frequencies... [using founder genotypes]

SNP_1 SNP_2 SNP_3

Lumping alleles with frequencies of 0.05 or less...

Testing trait: sim_phen

=====

Allele	df(0)	RSS(0)	df(add)	RSS(add)	F	(df1, df2)	P-value	Sample-size
N/A	199	374.317	196	348.266	4.89	(3, 196)	0.0027	200

Using all three markers in the analysis to test the dominance effect, run the program by

CLAM_qtl -d sim_3_mrk.dat -p sim_3_mrk.ped -x-99.999 -acd -w-

the output contains the following results for dominance effect:

The following models will be evaluated...

NULL MODEL

Means = Mu

FULL MODEL

Means = Mu + Z_dom

Estimating allele frequencies... [using founder genotypes]

SNP_1 SNP_2 SNP_3

Lumping alleles with frequencies of 0.05 or less...

Testing trait: sim_phen

=====

Allele	df(0)	RSS(0)	df(dom)	RSS(dom)	F	(df1, df2)	P-value	Sample-size
N/A	199	374.317	196	350.787	4.38	(3, 196)	0.0052	200

Using all three markers in the analysis to test the dominance effect given the additive effect, run the program by *CLAM_qtl -d sim_3_mrk.dat -p sim_3_mrk.ped -x-99.999 -aacd -bac -w* the output contains the following results for dominance effect:

The following models will be evaluated...

NULL MODEL

$$\text{Means} = \mu + X_{\text{add}}$$

FULL MODEL

$$\text{Means} = \mu + X_{\text{add}} + Z_{\text{dom}}$$

Estimating allele frequencies... [using founder genotypes]

SNP_1 SNP_2 SNP_3

Lumping alleles with frequencies of 0.05 or less...

Testing trait: sim_phen

=====

Allele	df(add)	RSS(add)	df(add+dom)	RSS(add+dom)	F	(df1, df2)	P-value	Sample-size
N/A	196	348.266	193	324.565	4.70	(3, 193)	0.0034	200

Using the marker one by one in the analysis, run the program by *CLAM_qtl -d sim_3_mrk.dat -p sim_3_mrk.ped -x-99.999 -aadm -w* the output contains the following results for dominance effect:

The following models will be evaluated...

NULL MODEL

$$\text{Means} = \mu$$

FULL MODEL

$$\text{Means} = \mu + X_{\text{add}} + Z_{\text{dom}}$$

Estimating allele frequencies... [using founder genotypes]

SNP_1 SNP_2 SNP_3

Lumping alleles with frequencies of 0.05 or less...

Testing trait: sim_phen

=====

Testing marker: SNP_1

Allele	df(0)	RSS(0)	df(add+dom)	RSS(add+dom)	F	(df1, df2)	P-value	Sample-size
All	199	374.317	197	351.866	6.28	(2, 197)	0.0023	200

Testing marker: SNP_2

Allele	df(0)	RSS(0)	df(add+dom)	RSS(add+dom)	F	(df1, df2)	P-value	Sample-size
All	199	374.317	197	359.273	4.12	(2, 197)	0.0176	200

Testing marker: SNP_3

Allele	df(0)	RSS(0)	df(add+dom)	RSS(add+dom)	F	(df1, df2)	P-value	Sample-size
All	199	374.317	197	345.088	8.34	(2, 197)	0.0003	200

3.3 Methods and Models

In this subsection, we briefly describe the theory of results of subsection 3.2. The related mathematical models and notations below can be found in Fan et al. (2006) for two marker case; if more than two markers are used in analysis, the models and notations can be extended similarly. The program reports F -Statistics and p-values of the following 4 tests:

1. Test of additive effect: fit model (13) of Fan et al. (2006) or the following regression

$$y = w\gamma + \alpha + \sum_{i=1}^{m-1} x_{Ai}\alpha_{Ai} + \sum_{k=1}^{n-1} x_{Bk}\alpha_{Bk} + e, \quad (1)$$

and test the hypothesis $H_{ABa0} : \alpha_{A1} = \dots = \alpha_{A(m-1)} = \alpha_{B1} = \dots = \alpha_{B(n-1)} = 0$.

2. Test of dominance effect: fit the following regression

$$y = w\gamma + \alpha + \sum_{1 \leq i < j \leq m} z_{Aij}\delta_{Aij} + \sum_{1 \leq k < l \leq n} z_{Bkl}\delta_{Bkl} + e, \quad (2)$$

and test hypothesis $H_{ABd0} : \delta_{A12} = \dots = \delta_{A1m} = \dots = \delta_{A(m-1)m} = \delta_{B12} = \dots = \delta_{B1n} = \dots = \delta_{B(n-1)n} = 0$.

3. Test of genotype effect: fit model (14) of Fan et al. (2006) or the following regression

$$y = w\gamma + \alpha + \sum_{i=1}^{m-1} x_{Ai}\alpha_{Ai} + \sum_{k=1}^{n-1} x_{Bk}\alpha_{Bk} + \sum_{1 \leq i < j \leq m} z_{Aij}\delta_{Aij} + \sum_{1 \leq k < l \leq n} z_{Bkl}\delta_{Bkl} + e, \quad (3)$$

and test hypothesis $H_{ABad0} : \alpha_{A1} = \dots = \alpha_{A(m-1)} = \alpha_{B1} = \dots = \alpha_{B(n-1)} = \delta_{A12} = \dots = \delta_{A1m} = \dots = \delta_{A(m-1)m} = \delta_{B12} = \dots = \delta_{B1n} = \dots = \delta_{B(n-1)n} = 0$.

4. Test of dominance effect given additive effect: fit above model (3) or model (14) of Fan et al. (2006), and test hypothesis $H_{ABd0|ABa0} : \delta_{A12} = \dots = \delta_{A1m} = \dots = \delta_{A(m-1)m} = \delta_{B12} = \dots = \delta_{B1n} = \dots = \delta_{B(n-1)n} = 0$.

4 How to Run the Program for Pedigree Data

The theoretical basis is Fan and Jung (2004), Fan et al. (2005, 2007), Fan and Xiong (2003), Jung et al. (2005), Xiong et al. (2002). By pedigrees, we mean nuclear families, sib-ships, and multi-generation pedigrees of any sizes and any types of relatives. For population data, one may treat each individual as an independent pedigree. The program can analyze pedigree data or combination of pedigree data and population data, for a combined linkage and association mapping of QTL.

4.1 Data Format and Command Options

To analyze pedigree data, the program requires four matched files: pedigree file, data file, and IBD (for identical by descent) file. The pedigree file and data file are similar to those of population data described in subsection 3.1. The difference is that only one individual is available for each family in population data, but more than one individuals are contained in a pedigree we are talking about here. The IBD file is a companion file of IBD estimates for the pedigree file, which can be calculated by MERLIN. One may want to notice that the IBD estimates are calculated based on the marker information. However, no marker information is needed for the kinship coefficients (Chapter 5, Lange 2002).

4.2 Example

In subdirectory `data_simu_Fig_B`, we use the pedigree data “`sim1.ped`”. The dataset is simulated using the template pedigree is pedigree b), Figure 1 of Fan, Spinka, Jin and Jung (2005) or Abecasis et al. (2000). The related “`sim_B.dat`” is the same as that of subdirectory `data_simu_Fig_A`.

Run the program by

```
CLAM_qtl -d sim_B.dat -p sim_B.ped -i merlin_B.ibd -x-99.999 -aa -c -wea
```

the results are

The following models will be evaluated...

```
NULL MODEL
```

Means = Mu

Variances = Ve + Va

FULL MODEL

Means = Mu + X_add

Variances = Ve + Va

Estimating allele frequencies... [using founder genotypes]

marker QTL

Testing trait: sim_phen1

=====

Testing marker: marker

Allele	df(0)	-LnLk(0)	df(add)	-LnLk(add)	ChiSq	P-value	Sample-size
1	809	2032.954	808	2027.008	11.89	0.0006	812
2	809	2032.954	808	2032.361	1.19		812
3	809	2032.954	808	2023.024	19.86	8e-06	812

Testing marker: QTL

Allele	df(0)	-LnLk(0)	df(add)	-LnLk(add)	ChiSq	P-value	Sample-size
1	897	2263.228	896	2241.183	44.09	3e-11	900
2	897	2263.228	896	2241.183	44.09	3e-11	900

Testing trait: sim_phen2

=====

Testing marker: marker

Allele	df(0)	-LnLk(0)	df(add)	-LnLk(add)	ChiSq	P-value	Sample-size
1	809	1801.163	808	1784.063	34.20	5e-09	812
2	809	1801.163	808	1800.439	1.45		812
3	809	1801.163	808	1776.466	49.39	2e-12	812

Testing marker: QTL

Allele	df(0)	-LnLk(0)	df(add)	-LnLk(add)	ChiSq	P-value	Sample-size
--------	-------	----------	---------	------------	-------	---------	-------------

```

1      897      1991.770                896          1938.735  106.07      7e-25  900
2      897      1991.770                896          1938.735  106.07      7e-25  900

```

If we want to impute the missing genotype by using the method of Fan et al. (2007), run the program by

```
CLAM_qtl -d sim_B.dat -p sim_B.ped -i merlin_B.ibd -x-99.999 -aar -c- -wea
```

the results are

```
Testing trait:                sim_phen1
```

```
=====
```

```
Testing marker:                marker
```

```
-----
```

Allele	df(0)	-LnLk(0)	df(add)	-LnLk(add)	ChiSq	P-value	Sample-size
1	897	2262.781	896	2257.389	10.79	0.0010	900
2	897	2262.781	896	2262.104	1.35		900
3	897	2262.781	896	2253.269	19.03	1e-05	900

```
.....
```

```
Testing trait:                sim_phen2
```

```
=====
```

```
Testing marker:                marker
```

```
-----
```

Allele	df(0)	-LnLk(0)	df(add)	-LnLk(add)	ChiSq	P-value	Sample-size
1	897	1994.029	896	1978.236	31.59	2e-08	900
2	897	1994.029	896	1993.474	1.11		900
3	897	1994.029	896	1972.046	43.97	3e-11	900

If we want to use multi-allelic marker model and to impute the missing genotype by using the method of Fan et al. (2007), run the program by

```
CLAM_qtl -d sim_B.dat -p sim_B.ped -i merlin_B.ibd -x-99.999 -aarm -c- -wea
```

the results are

```
Testing trait:                sim_phen1
```

```
=====
```

```
Testing marker:                marker
```

```

-----
Allele  df(0)  -LnLk(0)    df(add)    -LnLk(add)  ChiSq  P-value  Sample-size
    All   897   2262.781      895       2252.323   20.92   3e-05   900
Testing marker:                                QTL
-----

```

```

Allele  df(0)  -LnLk(0)    df(add)    -LnLk(add)  ChiSq  P-value  Sample-size
    All   897   2263.228      896       2241.183   44.09   3e-11   900

```

Testing trait: sim_phen2

=====

Testing marker: marker

```

-----
Allele  df(0)  -LnLk(0)    df(add)    -LnLk(add)  ChiSq  P-value  Sample-size
    All   897   1994.029      895       1967.519   53.02   3e-12   900
Testing marker:                                QTL
-----

```

```

Allele  df(0)  -LnLk(0)    df(add)    -LnLk(add)  ChiSq  P-value  Sample-size
    All   897   1991.770      896       1938.735  106.07   7e-25   900

```

If we want to use both markers in one analysis and to impute the missing genotype by using the method of Fan et al. (2007), run the program by

```
CLAM_qtl -d sim_B.dat -p sim_B.ped -i merlin_B.ibd -x-99.999 -aarmc -c- -wea
```

the results are

Testing trait: sim_phen1

=====

Using IBD information at marker: marker

```

-----
Allele  df(0)  -LnLk(0)    df(add)    -LnLk(add)  ChiSq  P-value  Sample-size
    N/A   897   2262.781      894       2240.365   44.83   1e-09   900
Using IBD information at marker:                QTL
-----

```

```

Allele  df(0)  -LnLk(0)    df(add)    -LnLk(add)  ChiSq  P-value  Sample-size

```

N/A 897 2263.228 894 2240.981 44.49 1e-09 900

Testing trait: sim_phen2

Using IBD information at marker: marker

Allele	df(0)	-LnLk(0)	df(add)	-LnLk(add)	ChiSq	P-value	Sample-size
N/A	897	1994.029	894	1939.094	109.87	1e-23	900

Using IBD information at marker: QTL

Allele	df(0)	-LnLk(0)	df(add)	-LnLk(add)	ChiSq	P-value	Sample-size
N/A	897	1991.770	894	1938.506	106.53	6e-23	900

If we want to include covariates in analysis and to impute the missing genotype by using the method of Fan et al. (2007), run the program by

CLAM_qtl -d sim_B.dat -p sim_B.ped -i merlin_B.ibd -x-99.999 -aarmd -cu -wead

the results are

The following models will be evaluated...

NULL MODEL

Means = $\mu + QTL_{ibd1} + QTL_{ibd2}$

Variances = $V_e + V_a + V_d$

FULL MODEL

Means = $\mu + QTL_{ibd1} + QTL_{ibd2} + X_{add} + Z_{dom}$

Variances = $V_e + V_a + V_d$

Estimating allele frequencies... [using founder genotypes]

marker QTL

Lumping alleles with frequencies of 0.05 or less...

Testing trait: sim_phen1

Testing marker: marker

Allele	df(0)	-LnLk(0)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
--------	-------	----------	-------------	----------------	-------	---------	-------------

All 894 2260.844 889 2248.725 24.24 0.0002 900

Testing marker: QTL

Allele	df(0)	-LnLk(0)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
All	894	2261.303	892	2239.180	44.25	2e-10	900

Using the two markers one by one in the analysis to test the dominance effect given the additive effect, run the program by

CLAM_qtl -d sim_B.dat -p sim_B.ped -i merlin_B.ibd -x-99.999 -aarmd -barm -wead

the results are

The following models will be evaluated...

NULL MODEL

Means = $\mu + X_{add}$

Variances = $V_e + V_a + V_d$

FULL MODEL

Means = $\mu + X_{add} + Z_{dom}$

Variances = $V_e + V_a + V_d$

Estimating allele frequencies... [using founder genotypes]

marker QTL

Lumping alleles with frequencies of 0.05 or less...

Testing trait: sim_phen1

Testing marker: marker

Allele	df(add)	-LnLk(add)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
All	894	2252.280	891	2250.666	3.23		900

Testing marker: QTL

Allele	df(add)	-LnLk(add)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
All	895	2241.139	894	2241.104	0.07		900

Testing trait: sim_phen2

=====

Testing marker: marker

Allele	df(add)	-LnLk(add)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
All	894	1967.519	891	1965.341	4.36		900

Testing marker: QTL

Allele	df(add)	-LnLk(add)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
All	895	1938.735	894	1938.734	0.00		900

Using all two markers in the analysis to test the dominance effect given the additive effect, run the program by

CLAM_qlt -d sim_B.dat -p sim_B.ped -i merlin_B.ibd -x-99.999 -aarmcd -barm -wead

the results are

Testing trait: sim_phen1

=====

Using IBD information at marker: marker

Allele	df(add)	-LnLk(add)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
N/A	893	2240.366	889	2238.290	4.15		900

Using IBD information at marker: QTL

Allele	df(add)	-LnLk(add)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
N/A	893	2240.945	889	2238.810	4.27		900

Testing trait: sim_phen2

=====

Using IBD information at marker: marker

Allele	df(add)	-LnLk(add)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
N/A	893	1939.094	889	1936.677	4.83		900

Using IBD information at marker: QTL

Allele	df(add)	-LnLk(add)	df(add+dom)	-LnLk(add+dom)	ChiSq	P-value	Sample-size
--------	---------	------------	-------------	----------------	-------	---------	-------------

5 References

1. Abecasis GR, Cookson WOC, and Cardon LR (2000) Pedigree tests of linkage disequilibrium. *Euro J Hum Genet* 8:545-551.
2. Fan RZ and Jung JS (2004) High resolution joint linkage disequilibrium and linkage mapping of quantitative trait loci based on sibship data. *Human Heredity* 56:166-187.
3. Fan RZ, Jung JS and Jin L (2006) High resolution association mapping of quantitative trait loci, a population based approach. *Genetics* 172:663-682.
4. Fan RZ, Liu L, Jung JS, and Zhong M (2007) Combined linkage and association mapping of quantitative trait loci with missing genotype data.
5. Fan RZ, Spinka C, Jin L, and Jung JS (2005) Pedigree linkage disequilibrium mapping of quantitative trait loci. *European Journal of Human Genetics* 13:216-231.
6. Fan RZ and Xiong MM (2002) High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. *European Journal of Human Genetics* 10:607-615.
7. Fan RZ and Xiong MM (2003) Combined high resolution linkage and association mapping of quantitative trait loci. *European Journal of Human Genetics* 11:125-137.
8. Jung JS, Fan RZ and Jin L (2005) Association mapping of quantitative trait loci by multiple markers. *Genetics* 170:881-898.
9. Lange K (2002) *Mathematical and Statistical Methods for Genetic Analysis*, 2nd edition. Springer.
10. Xiong MM, Fan RZ, and Jin L (2002) Linkage disequilibrium mapping of quantitative trait loci under truncation selection. *Human Heredity* 53:158-172.

Table 1: Command Line Options.

Column	Description
-d datafile	Data File Name
-p pedfile	Pedigree Name
-i ibdfile	IBD File Name
-a[regression model]	Association Model: -a- (do not model association), -aa (model additive effect), -ad (model dominance effect), -aad (model both additive and dominance effects); other three options: r (impute genotype assuming missing completely at random), m (model by multi-allelic marker instead of di-allelic marker), c (combine all markers in the model, instead of single marker one by one); the options r , m , and c can be combined with a , d , such as -aar , -aam , -aac , -aarc
-b[regression model]	Another Association Model for Comparison, such as -aarmcd -barm
-w[components]	Variance Components: e (error variance), a (additive variance of major locus), d (dominance variance of major locus), g (additive polygenic variance), c (common environmental variance), and combinations such as -wea , -wead , and -weadg
-v[components]	Another Variance Component for comparison, such as -we -vea
-c [covariates]	Covariates: -c- (do not use covariates in the analysis), -cu (user-defined covariates from pedigree and data files), -cs (sex), -cp (parental phenotypes used as condition to model offspring phenotypes), and combinations such as -cus , -cup , and -cups .
-x miss	Missing Value: specifies phenotype missing values