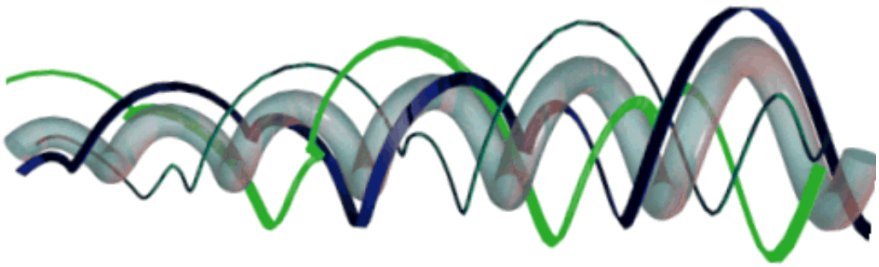
INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE

Supplement to MCQTL reference manual : Outbred families

Brigitte Mangin
Remy Cathelin, David Delannoy, Stéphane Lambert, Jacques Marcel, Baba Ngom, Marie-Françoise Jourjon
Sylvain Jasson

RAPPORT UBIA TOULOUSE N° 2007/3 - OCTOBER 2007

DÉPARTEMENT DE MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES



INRA

MCQTL_Outbred is an additional part of MCQTL software package which permits to perform QTL mapping in multiple outbred families.

Implemented methods in MCQTL are available for outbred families except the epistasis search.
Main differences are :

– the input data files whose format is based upon MapQTL and JoinMap 3.0 input files (Van Ooijen and Voorrips, 2001),
– the modelling of QTL effects which is specific to outbred families. Indeed, in each position along the genome the QTL is assumed to have four different alleles.

Table des matières

1 Model

The intra-family model is a usual regression model with genetic cofactors and a single QTL. Genetic cofactors and the single QTL are assumed to have four different alleles.
Let c denotes the cross between two parent i, j, the phenotypic value Y_{ck} of the kth individual is modelled by

$$Y_{ck} = \mu_c + \sum_{l=1}^{L} \sum_{a=1}^{2} \sum_{a'=1}^{2} p^l_{ck,i_a j_{a'}} \theta^l_{c,i_a j_{a'}} + \epsilon_{ck}$$

where μ_c is the global mean in the cross c, $L-1$ is the number of genetic cofactors, $p^l_{ck,i_a j_{a'}}$ is the probability of the kth individual having genotype $i_a j_{a'}$ at the QTL or cofactor locus l given the marker information, $\theta^l_{c,i_a j_{a'}}$ is the mean of the $i_a j_{a'}$ genotype at locus l in cross c and ϵ_{ck} the residual error.
Moreover, the QTL or genetic cofactor effect is decomposed in two effects, the additive and the dominance ones

$$\theta^l_{c,i_a j_{a'}} = \alpha^l_{c,i_a} + \alpha^l_{c,j_{a'}} + \delta^l_{c,i_a j_{a'}}$$

Constraints
– for additive parameters and both choices of parameter dependency

$$\alpha^l_{[c],i_1} = -\alpha^l_{[c],i_2}\ \alpha^l_{[c],j_1} = -\alpha^l_{[c],j_2}\ \forall c$$

– for dominance parameters and both choices of parameter dependency

$$\delta^l_{[c],i_1 j_1} = \delta^l_{[c],i_2 j_2} = -\delta^l_{[c],i_2 j_1} = -\delta^l_{[c],i_1 j_2}\ \forall c$$

In an additive model, the dominance parameters are all assumed to be equal to zero.

2 Data entry files

Four ASCII files with a single (no other point character) and mandatory extension .loc .qua .map and .inf are necessary to run the TranslateData program. They contain the marker data and information about the family (.loc), the quantitative trait data (.qua) and a consensus map (.map). In the last file (.inf), names of the parent are provided. Files with extension .loc and .map are similar respectively to the locus genotype file and the map file of JoinMap 3.0 (Van Ooijen and Voorrips, 2001).

2.1 .loc file

This file contains the marker data and information about the family. Its format is similar to the JoinMap 3.0 format of the locus genotype file.

Comment lines are allowed in .loc and .map files. They begin by a `;` character.

⚠ Contrarily to JoinMap 3.0 format, the so-called *whitespace* is limited to a sequence of spaces and mustn't contains tab, newline or carriage-return characters.

The header of the file contains four instructions :

`name = P1xP2`	name of the family
`popt = CP`	code for outbred family
`nloc =`	number of marker loci
`nind =`	number of individuals

⚠ The name of the family must be written carefully, as it is used to link families which share one parent and to assign the number of alleles and the linkage phase of the marker loci for both parents. The name of the family must be formed by the parent names given in the .inf file separated by a `x` character. You must respect the order given in the .inf file.

Next, the marker information follows on 2 lines with the following syntax :

```
locusname <SEG> {PHASE}
gen1 gen2 ...
```

Code for the type of segregation, the locus phase and the possible observed marker genotypes are the same as JoinMap 3.0 code for family type CP. They are summarized in table 1. However incomplete genotype information have been extended in order to take into account either dominant markers or missing alleles. The code for these incomplete genotype are presented in table 2.

Table 1

<SEG>	{PHASE}	Possible genotypes
<abxcd>	{00}, {01},{10} or {11}	ac, ad, bc, bd, --
<abxac>	{00}, {01},{10} or {11}	aa, ac, ba, bc, --
<abxab>	{00}, {01},{10} or {11}	aa, ab, bb, --
<abxaa>	{0-} or {1-}	aa, ab, --
<aaxab>	{-0} or {-1}	aa, ab, --

Table 2

<SEG>	Incomplete genotype code	Possible genotypes
<abxcd>	a-	ac or ad
	b-	bc or bd
	-c	ac or bc
	-d	ad or bd
<abxac>	a-	aa or ac
	b-	ba or bc
	-a	aa or ba
	-c	ac or bc
	a*	aa, ac or ba
<abxab>	a-	aa, ab or ba
	b-	bb, ab or ba

⚠ **Be careful of the phase of the marker loci of both parents. Indeed, in designs of multiple related families, each common parent is assumed to have a coherent phase for all**

the marker loci in each family they are involved in. There is no procedure to check the coherence of the phases. Il you have doubt about the phase coherence, use disconnected as value for the interpop attribute of MODEL tag. .

example of a .loc file

```
; data from a cross between two heterogeneously heterozygous and homozygous diploid parents
name = P1xP2
popt = CP
nloc = 6
nind = 6
mark1 <abxcd> {00}
ac ac ad bc b- bd
mark2 <aaxab> {-0}
aa ab aa -- aa ab
mark3 <abxab> {10}
ab ab ab aa aa ab
mark4 <abxab> {00}
bb bb a- a- a- bb
mark5 <abxaa> {1-}
 ab -- aa ab aa ab
mark6 <abxac> {10}
a* bc b- bc -- a*
```

2.2 .qua file

This file contains the quantitative trait data. Its format is similar to the MapQTL format of the quantitative data file.
The header contains three instructions :

`ntrt =` number of traits
`nind =` number of individuals
`miss =` missing value code

⚠ Do not use the star character as a missing code

Then follow the trait names and the observed trait data grouped per individual (contrarily to the .loc file in which data are grouped per locus). So be careful to respect the order of individuals between the .qua file and the .loc file. There is no constraints on the trait names. On the contrary, the trait data must be numeric (except for the missing code). Each data must be separated by space(s) and all the data concerning one individual must be filled in a single line.

example of a .qua file

```
; map for family P1xP2
ntrt = 3
nind = 6
miss = -
index
```

```
trait1
trait2
1  1.2  3
2  -   -2
3  -   -
4  0.2 -1.5
5  2.   2.1
6 -3.0  2
```

2.3 .map file

This file contains a consensus map for families that are going to be analyzed together with MultiPop. Its format is similar to JoinMap 3.0 format of the map file. It is a completely line-structured file.

Each linkage group begins by the instruction `group` followed by the name of the group. Then follows, in an ascending order, a name of a marker and its positions in the map until the group is complete.

⚠ The position must be a Haldane distance given in centiMorgans.

example of a .map file

```
group chrom1
mark1 0.2
mark4 15
mark3 17.98
mark6  32.0
group chrom2
mark2 0.0
mark5 20.0
```

2.4 .inf file

This file contains the names of the parents given by the following instructions :

`*lineA P1` name of the first parent
`*lineB P2` name of the second parent

⚠ Parent names are crucial for MultiPop. Indeed, phenotypes of families that share a parent name are modelled with equal QTL or cofactor additive parameters.

⚠ Be careful to respect the order of the parents between .inf and .loc files.

3 LOD support

The default value of the LOD support is 1.

4 References

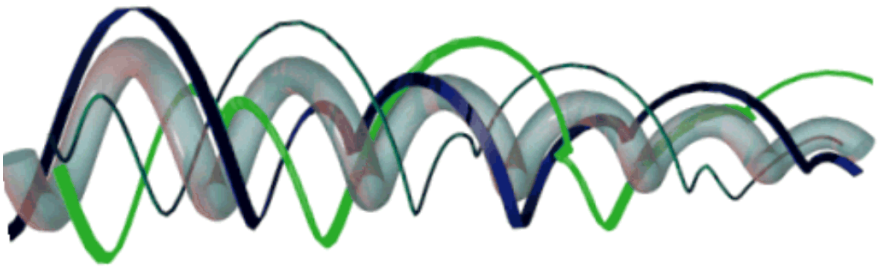
Van Ooijen J. W., Voorrips R. E. 2001. JoinMap 3.0, Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, the Netherlands.

L'unité de Biométrie et Intelligence Artificielle de Toulouse a pour mission de développer et d'appliquer des méthodes de statistiques et d'intelligence artificielle dans le cadre des grands axes de recherche de l'INRA. L'unité travaille en particulier sur l'analyse par modélisation, simulation et optimisation de systèmes en agronomie, écologie, épidémiologie, gestion forestière, et dans le domaine de la biologie sur la localisation et l'identification d'éléments fonctionnels dans les génomes des bactéries, plantes et animaux, aux niveaux génétique, moléculaire et de l'expression de gènes.
Ces recherches s'accompagnent d'une activité de production de logiciels pour leur valorisation et d'une activité de formation pour leur diffusion.

The Biometry and Artificial Intelligence laboratory at Toulouse is part of INRA (French National Institute for Agricultural Research). Its mission is to develop and apply statistical and artificial intelligence methodologies to the specific research domains of INRA. The laboratory develops methods for modeling, simulating and optimizing systems in agronomy, ecology, epidemiology, forest management. In biology, models and methods are developed that aim at locating and characterizing functional elements inside the genomes of bacteria, plants and animals, at the genetic, sequence and expression data levels. This also includes teaching and software development.

INRA - UBIA, BP 52627 Auzeville F-31326 Castanet Tolosan cedex
Tél : +33(0)5.61.28.52.75 - Fax : +33(0)5.61.28.53.35 - http://mia.toulouse.inra.fr

Image de synthèse : Benjamin Auriol
Publication : Jérôme Juille (INRA - SDAR)



INRA
Institut National de la Recherche Agronomique