FullsibQTL Tutorial

Software for QTL mapping using a full-sib progeny

Rodrigo Gazaffi[1]

Antonio Augusto Franco Garcia[1*]

[1]Department of Genetics

Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ), Universidade de São Paulo (USP)

Av. Pádua Dias, 11 - Caixa Postal 83

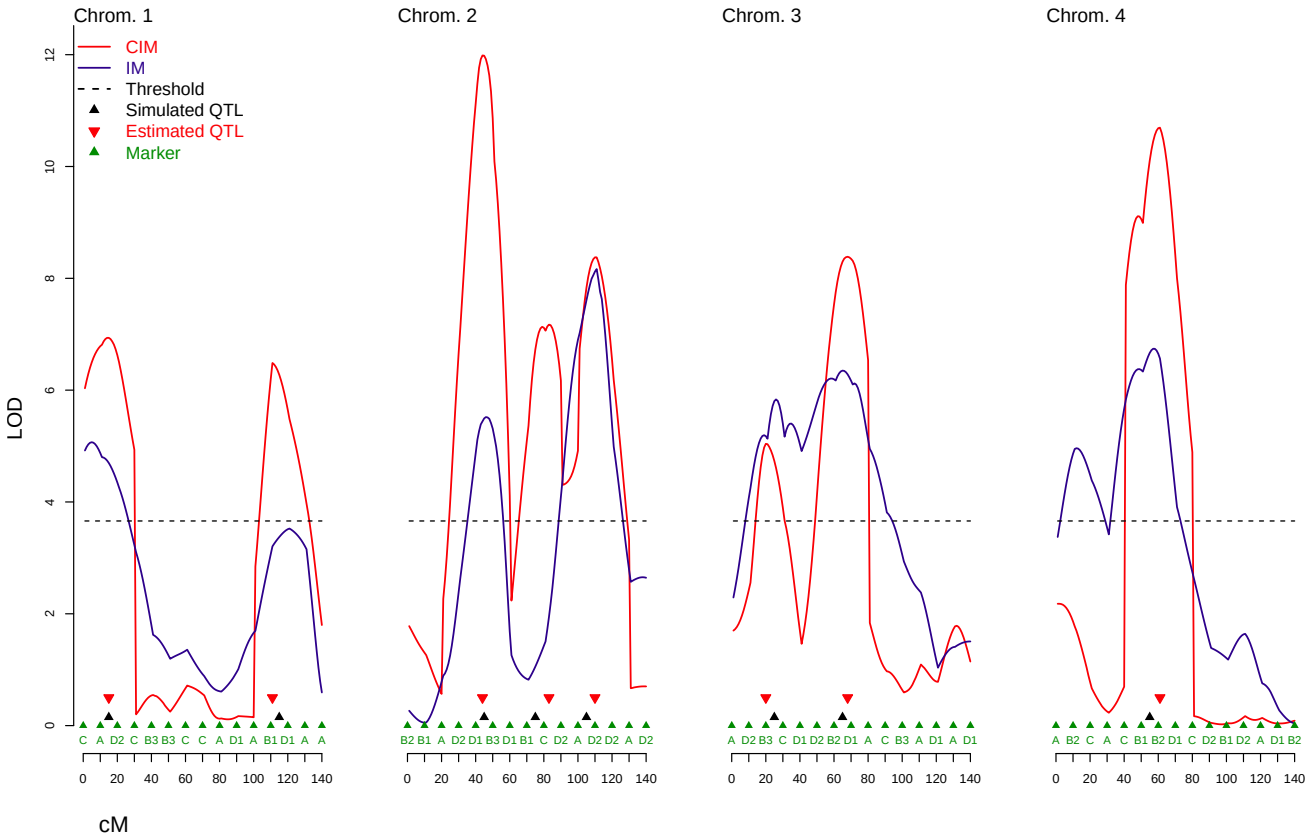
CEP: 13400-970 - Piracicaba - São Paulo - Brazil

Tel: +55 19 34294125

Fax: +55 19 34336706

E-mail: augusto.garcia@usp.br

[*]corresponding author



http://www.r-project.org

7 de outubro de 2013

1 Introduction

FullsibQTL is an `R` package to perform QTL mapping in outbred/outcrossing species. We consider as a mapping population a full-sib progeny (or F_1 population) derived by a biparental cross between two non homozygous parents, with a genetic map obtained with markers showing different segregation patterns. Here, we assumed the phenotypes are continuous having normal distribution and the genetic map was previously obtained with *onemap* package. If you are not familiar with *onemap*, we strongly encouraged to read *onemap* Vignettes (Margarido et al., 2007).

In our software, we developed tools to perform (composite) interval mapping. Briefly, we developed an statistical model with three genetics effects, one for each parent and an interation (dominance). To obtain these estimatives, we used maximum likelihood approach using mixture models and EM algorithm (Gazaffi et al., 2013). The QTL genotype probabilities were calculated using a multipoint technology, based on Hiddden Markov Models (Wu et al., 2002b). We also implemented a function to select cofactors using multiple linear regression and the information criteria. Permutation test (Churchill and Doerge, 1994) for threshold determination were also implemented, as well the modification proposed by Chen and Storey (2006) for having a relaxed threshold. Some extra functions were developed to provide a graphical or text output for the analysis helping the interpretation of QTL mapping. To exemplify the usage of *FullsibQTL* we will analyse the dataset present by Gazaffi et al. (2013).

The purpose of this tutorial is to help users dealing with this package and understanding the outputs. It is not our intention to teach the genetic basis of QTL mapping approach, see Gazaffi et al. (2013) for details. Remembering, users of *FullsibQTL* are supposed to have some experience with R, since the analysis is done using the command line, previous acknowledgement of *OneMap* is also desirably, once the genetic map is obtained with this software. *FullsibQTL* is available as source code for Windows™ and Unix. It is released under the GNU General Public License, is open-source and the code can be changed freely. It comes with no warranty. It is implemented as a package to be used under the freely distributed R software, which is a language and environment for statistical computing (www.r-project.org).

1.1 Citation

Gazaffi, R, Margarido, GRA, Pastina, MM, Mollinari, M, Garcia, AAF (submitted) A model for quantitative trait loci mapping, linkage phase and segregation pattern estimation for a full-sib progeny. *Tree Genetics and Genomes*.

1.2 Instalation

After installing R, *FullsibQTL* can be installed by opening R and issueing the command

The package can be also installed by downloading the appropriate files directly at the CRAN web site and following the instructions given in the section "6.3 Installing Packages" of the "R Installation and Administration" manual (http://cran.r-project.org/doc/manuals/R-admin.pdf).

1.3 Getting started

The following example is intended to show the usage of all *FullsibQTL* functions. With basic knowledge of R syntax, one should have no big problems using it. It is assumed that the user is running Windows™. Hopefully, these examples will be clear enough to help any user to understand its functionality and start using it.

1. Start R by double-clicking its icon.
2. Load *FullsibQTL* (after installing it):

   ```
   > library(fullsibQTL)
   ```

3. To save your project anytime, type:

   ```
   > save.image("C:/.../yourfile.RData")
   ```

   or access the toolbar File → Save Workspace.

4. To load your project, in a new section type:

   ```
   > load("C:/.../yourfile.RData")
   ```

   or access the toolbar File → Load Workspace.

5. To change the working directory, one can type:

   ```
   > setwd("C:/.../new_working_directory")
   ```

   or acess the toolbar File → Chance working directory.

6. To open the help file for a any function, type "?" followed by the function name:

   ```
   > ?im.scan
   ```

1.4 Functions

FullsibQTL package is composed by a small set of functions present in the Table 1. There are some other functions used internally by the software, which one don not use them directly.

Tabela 1: FullsibQTL functions that are directly called by users

Function type	Function name	Function description
Input	read.outcross.pheno	Reads the data file containing markers and phenotypic values
	create.fullsib	Creates the object to perform QTL
interval mapping	im.scan	Scan the genome using IM approach
	im.char	Provides the genetic effects, LOD Score and the p-values of complementary tests
cofactors	cof.selection	Selects cofactors using multiple linear regression
	cof.definition	Defines locations on the genome to be used as cofactors
composite interval mapping	cim.scan	Scan the genome using CIM approach
	cim.char	Provides the genetic effects, LOD Score and the p-values of complementary tests
summaries	print.fullsib	Prints a summary for the object of class 'fullsib'
	print.fullsib.scan	Prints the result of im.scan or cim.scan
	summary.fullsib.scan	Summarizes the QTL search for im.scan or cim.scan
	plot.fullsib.scan	Plot the QTL profile for the mapped groups
	summary.fullsib.perm	Provides the threshold values for permutations test
	plot.fullsib.perm	Plot the empiral distribution for permutation test
	draw.phase	Returns the linkage phase between QTL and markers
	get.segr	Returns the QTL segregation
	r2.ls	Provides the phenotypic proportion of each QTL mapped and altogether (least square estimation)

2 Required data

Para proceder o mapeamento de QTL é necessário a inclusão de um arquivo texto contendo os genótipos e os fenótipos dos indivíduos da população, além do mapa genético previamente construído com o software *onemap*.

O arquivo de entrada é análogo ao utilizado pelo *onemap*, ou seja, o mesmo arquivo considerado para a construção do mapa genético é utilizado acrescentando-se ao final os fenótipos dos indivíduos da população. Abaixo, mostra-se um exemplo de como seria um arquivo com 10 indivíduos, 5 marcadores e 2 fenótipos.

```
10 5 2
*M1 B3.7        ab,ab,-,ab,b,ab,ab,-,ab,b
*M2 D2.18       o,-,a,a,-,o,a,-,o,o
*M3 D1.13       o,a,a,o,o,-,a,o,a,o
*M4 A.4         ab,b,-,ab,a,b,ab,b,-,a
*M5 D2.18       a,a,o,-,o,o,a,o,o,o
*pheno1         4.8, 2.1, 10.6, 3.7, -3.7, -7.5, -, 3.9, 3.6, -5.8
*pheno2         5, -5.8, -, 14.8, -2.4, -1.9, -1.1, 8.1, -11.1, 4.8
```

Atenção: Recomenda-se ao usuário a não alterar a ordem dos marcadores deste arquivo ao utilizado para a construção do mapa de ligação, caso isto ocorra resultados não esperados podem ocorrer.

Nota-se que na primeira linha o terceiro elemento indica o número de fenótipos contido no arquivo e a sétima linha contém os fenótipos. A indicação do nome do trait inicia-se com o nome da carateristica precedido pelo caracter "*". Os valores fenotipícos para cada indivíduo devem ser separados por vírgulas (",") e números decimais representados por ponto (".") . Dado perdido pode ser representado por "-" ou "NA". Neste documento não explicaremos a codificação utilizada os marcadores moleculares já previamente apresentada no na seção *Creating the data file* do tutorial do *onemap*.

Uma vez que este arquivo esteja disponível, sua leitura deve ser feita com a função `read.outcross.pheno`, adaptado do pacote *onemap* para receber a entrada dos valores fenotípicos.

```
> fs.data <- read.outcross.pheno("C:/workingdirectory","filename.txt")
```

Caso, o arquivo de dados esteja no mesmo diretório de trabalho pode-se também digitar:

```
> fs.data <- read.outcross.pheno(file="filename.txt")
```

Outra informação necessária ao programa é o mapa genético construído. Para tanto deve-se disponibilizar ao software os grupos de ligação previamente ordenados, neste caso o objeto da classe "sequence" para cada grupo obtido.

3 Loading dataset example

Para exemplificar o uso do pacote iremos utilizar o conjunto de dados simulados presente em Gazaffi et al. (2013). Para acessá-lo, o usuário deve digitar:

```
> data(QTLexample)
```

Ao carregar o exemplo, o usuário pode digitar o comando `ls` para visualizar que cinco variáveis foram carregadas:

```
> ls()
```

```
[1] "fullsib.data" "LG1.end"      "LG2.end"      "LG3.end"      "LG4.end"
```

O objeto `fullsib.data` contém os marcadores moleculares e os fenótipos lidos com a função `read.outcross.pheno` e os objetos `LG1.end`, `LG2.end`, `LG3.end` e `LG4.end` correspondem aos grupos de ligação previamente ordenados utilizando os recursos do *onemap* (objetos da classe "sequence"). Digitando o nome das variáveis o usuário pode visualizar os objetos.

4 Creating the working object

Para iniciar as análises, o usuário precisa utilizar a função `create.fullsib` para criar um objeto da classe `fullsib`, o qual contém todos os objetos necessários ao mapeamento de QTLs. Para o exemplo, disponibilizado o usuário deve digitar:

```
> fsib <- create.fullsib(fullsib.data,
+                        map.list=list(LG1.end, LG2.end, LG3.end, LG4.end),
+                        step=1, map.function="kosambi")
```

O primeiro argumento, `fullsib.data` refere-se a variável criada com a função `read.outcross.pheno` (leitura dos dados), map.list é o argumento que indica os grupos de ligação construídos com o *onemap*. Para as situações que há mais de um grupo de ligação, estes devem ser armazenados em uma variável da classe `list`, ou seja, cada elemento dessa variável corresponde a um grupo de ligação. Para o cálculo das probabilidades condicionais multiponto, utiliza-se o argumento `step` define a distância em cM entre as quais as probabilidades são obtidas. Se `step=0` as probabilidades são calculadas apenas nas posições que contém marcadores. `map.function` indica que o mapa foi construído considerando a função de mapeamento de "kosambi". O usuário pode ver um resumo desse objeto digitando o nome da variável como indicado abaixo:

```
> fsib
```

```
This is an object of class 'fullsib'

The linkage map has 4 groups, with 560 cM and 60 markers
No. individuals genotyped:  300
Group 1 : 140 cM,  15 markers (A, B1, B3, C, D1, D2)
Group 2 : 140 cM,  15 markers (A, B1, B2, B3, C, D1, D2)
Group 3 : 140 cM,  15 markers (A, B2, B3, C, D1, D2)
Group 4 : 140 cM,  15 markers (A, B1, B2, C, D1, D2)
And 5 unlinked markers


2 phenotypes are avaliable for QTL mapping
Multipoint probability for QTL genotype was obtained for each 1 cM
```

De forma geral, o objeto contém o mapa genético composto por quatro grupos de ligação, em que o seu comprimento total e a segregação dos marcadores estão indicadas. Também apresenta-se o número de marcadores não ligados ao mapa de ligação, isto é, marcadores contidos no arquivo de entrada que não estão posicionados nos grupos de ligação. Isto pode ser útil para espécies em que o mapa genético não abrange o genoma como todo, podendo ter marcadores que estão em regiões não saturadas. Finalmente, o número de fenótipos disponíveis para mapeamento estão indicados, assim como a distância usada para a obtenção das probabilidades condicionais.

5 Interval mapping

Para realizar o mapeamento por intervalo, há duas funções disponíveis `im.scan` e `im.char`. A primeira percorre o genoma fazendo associação entre genótipos e fenótipos buscando detectar QTLs. Para os locais que há QTLs a segunda função é utilizada para mostrar os efeitos genéticos e seus respectivos testes significâncias e também os testes complementares para inferência da segregação do QTL.

5.1 Genome scan

The genome scan can be performed with the function `im.scan`, such as:

```
> im1 <- im.scan(fsib, lg="all", pheno.col=1, LOD=TRUE)
```

O primeiro argumento `fsib` foi criado com a função `create.fullsib`, o argumento `lg` indica quais os grupos deve ser escaneado para QTL (outra possível forma de representação é `lg=1:4`,

o default é percorrer todos os grupos de ligação, porém apenas alguns grupos de ligação podem ser analisados), `pheno.col` indica qual o fenótipo deve ser analisado. `LOD=TRUE` indica que o resultado do mapeamento deve ser mostrado em LOD Score, caso a opção seja `FALSE` a escala da curva de mapeamento será em $-log_{10}(p-value)$.

O resultado do mapeamento pode ser visualizado digitando o nome da variável:

```
> im1
```

ou usando a função `print` que permite uso do argument `lg` que se usado pode restringir a impressão dos dados pelo grupo de ligação definido pelo usuário:

```
> print(im1,lg=1)
```

O objeto retornado pela função `im.scan` é uma matrix composta por 4 colunas e k linhas, sendo k o número de posições no genoma que as probabilidades condicionais foram calculadas. De forma geral, o usuário vê a identificação da posição estudada pelo nome do marcador molecular ou caso seja entre marcadores (obtido um pseudo marcador), o usuário vê o local identificado por uma string iniciada por "loc". A seguir há a indicação do grupo de ligação que está posição está localizada, posição em centiMorgan, o valor de LOD Score ou $-log_{10}(p-value)$ e por fim o modelo considerado (see section 9).

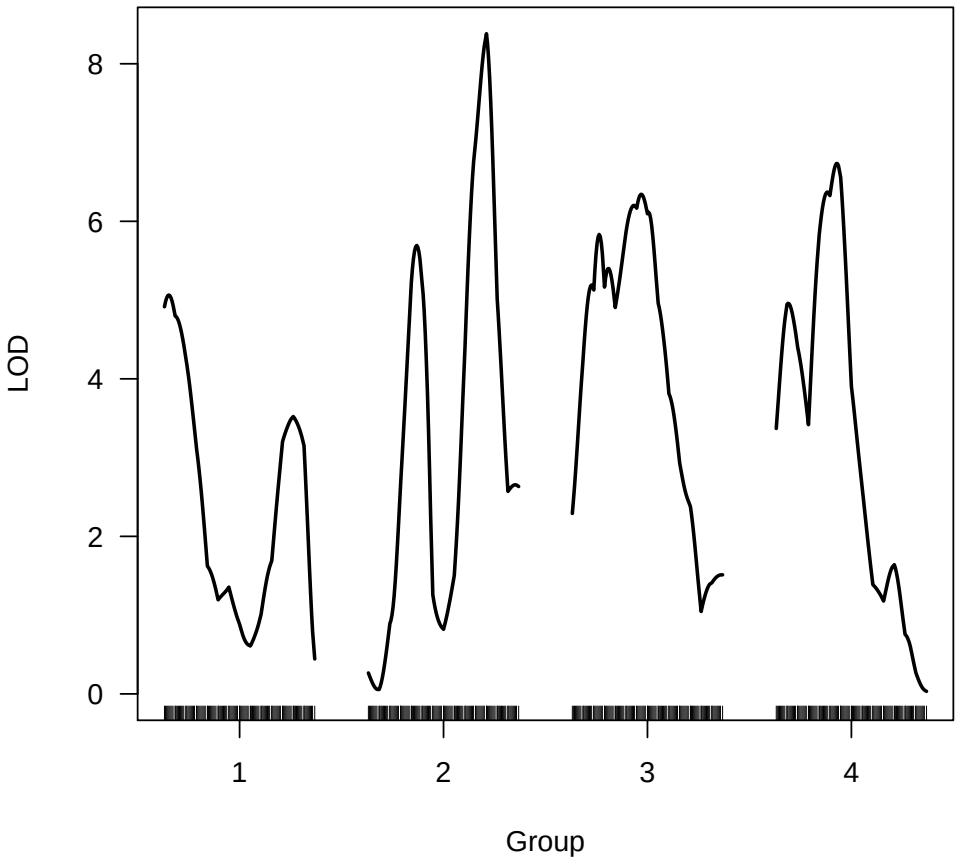
Uma forma sintética de visualizar os dados é utilizando a função `summary` que imprime o máximo valor de LOD Score ou $-log_{10}(p-value)$ para cada grupo de ligação. O argumento `thr` for utilizado, apenas o máximo valor acima desse threshold é impresso. Para as situações em que mais de um QTL é mapeado no mesmo grupo de ligação, o usuário deve buscar manualmente a localização dos demais QTLs, uma vez que identificar mais de um pico pode ser um processo muito subjetivo que pode variar de usuário para usuário.

```
> summary(im1)
```

```
      lg pos.cM      LOD model
loc4   1      4 5.063801     0
M27    2    110 8.381168     0
loc64  3     64 6.343023     0
loc56  4     56 6.733203     0
```

Ainda, o usuário pode-se visualizar graficamente o mapeamento de QTL utilizando a função `plot`:

```
> plot(im1)
```



O usuário pode customizar os gráficos utilizando os argumentos da função, por exemplo, em uma mesma área gráfica é possível representar até 5 perfis de mapeamento, além do usuário pode indicar a posição dos marcadores moleculares no mapa, alterar label dos grupos de ligação dentre outros recursos. Recomenda-se ao usuário digitar o comando `?plot.fullsib.scan` para acessar outros exemplos de uso desta função.

Por fim, para aumentar a flexibilidade no mapeamento é possível incluir covariável (de efeito fixo) ao modelo para controlar a variabilidade do fenótipo em estudo. Caso, este recurso seja utilizado deve-se usar o argumento `addcovar`, para entrar com a full-rank matrix que controlaria a variação do fenótipo, conforme exemplificado a seguir:

```
> covar <- matrix(rep(c(1,-1), each=150), ncol=1)
> im2 <- im.scan(fsib, pheno.col=2, addcovar=covar)
```

5.2 Permutation test

O threshold para detecção de QTL pode ser obtido através do uso do teste de permutações, o qual está implementado na função `im.scan`, nesse caso um objeto da classe `fullsib.perm` é

criado. Para executar o teste de permutações, o usuário deve considerar os mesmos argumentos utilizados para fazer o mapeamento, porém adicionando o argumento `n.perm`. Pelo fato do teste de permutação ser um processo de reamostragem, recomenda-se que o usuário defina uma semente (comando `set.seed`) antes da execução da função `im.scan` para que a permutação possa ser repetida se necessário. O argumento `write.perm` permite que o usuário defina uma string que será o nome de um arquivo texto contendo todas as permutações realizadas (válido para usuários que desejam acessar todos os valores gerados durante a análise), exemplificado abaixo:

```
> set.seed(123456789)
> perm1 <- im.scan(fsib, pheno.col=1, n.perm=1000)
> ## recording permutations values in text file
> perm2 <- im.scan(fsib, pheno.col=1, n.perm=1000, write.perm="permutations.txt")
```

Vale lembrar que o teste de permutações é uma análise intensiva a qual pode demandar algum tempo (Utilizando um ultrabbok com processador i5 e 1,70GHz, com 4Gigas de RAM e usando ubuntu 12.04 essa análise demorou 13,6 minutos). Para tanto, o usuário pode acompanhar o processo através da barra de progresso exibida na tela. Ao término do processo a variável `perm1` recebe uma matrix com p linhas, sendo p o número de permutações realizadas e duas colunas correspondentes aos dois maiores picos do genoma. A primeira coluna é nomeada como `peak.1` e mostra o maior valor obtido na curva de mapeamento, considerando todo o genoma e `peak.2` mostra o segundo maior pico do genoma. A definição do segundo utilizada aqui é análoga ao considerado no método dois de chen and storey (2006), isto é, corresponde ao maior valor de LOD Score ou $-log_{10}(p-value)$ obtido no genoma, desconsiderando o grupo de ligação em que o maior pico foi considerado. A obtenção do threshold pode ser utilizado com o comando `summary` aplicado ao objeto que recebeu os resultados dos testes de permutações, no caso o argumento `alpha` pode ser utilizado para a obtenção do threshold. Essencialmente essa função obtém o quantil $(1-\alpha) \times 100$ das amostras. 0 Default é mostrar valores de alpha igual a 0.90 e 0.95.

```
> summary(perm1, alpha=0.95)
```

Caso, o usuário queira ver a distribuição dos máximos valores pode-se aplicar a função `plot` ao objeto `perm1` no caso, um histograma mostra a distribuição dos valores para o primeiro pico. Se o argumento `peak` for alterado para '2', a distribuição dos segundos maiores picos será mostrada.

```
> plot(perm1)
> plot(perm1, peak=2)
```

5.3 Caracterização dos QTLs

Após determinar o número e a posição dos QTLs, o usuário pode caracterizar os QTLs, ou seja, identificar os efeitos significativos, sua segregação e a fase de ligação entre os alelos dos QTLs e marcadores flanqueadores. Para o mapeamento realizado com a função `im.scan` e armazenada em `im1` pode-se destacar a presença de pelo menos 4 QTLs:

```
> summary(im1)
```

```
      lg pos.cM      LOD model
loc4   1      4 5.063801     0
M27    2    110 8.381168     0
loc64  3     64 6.343023     0
loc56  4     56 6.733203     0
```

A caracterização destas quatro regiões é realizada com a função `im.char`, conforme indicada a seguir:

```
> qtl1 <- im.char(fsib, pheno.col=1, lg=1, pos="loc4")
> qtl2 <- im.char(fsib, pheno.col=1, lg=2, pos="M27")
> qtl3 <- im.char(fsib, pheno.col=1, lg=3, pos="loc64")
> qtl4 <- im.char(fsib, pheno.col=1, lg=4, pos="loc56")
```

Nota-se que as posições estudadas “loc4”, “M27”, “loc64” e “loc56” estão apresentadas como sendo nome das linhas presentes dentro da variável `im1` e também mantidas ao aplicar a função `summary`. O objeto que contém a caracterização do QTL localizado no grupo de ligação 1 é:

```
> qtl1
```

```
                  M1-M2
LG            1.00000000
pos           4.00000000
-log10(pval)  4.46566439
LOD_Ha        5.06392065
mu           -0.12075333
alpha_p       1.85844505
LOD_H1        4.34832177
alpha_q       0.43196490
LOD_H2        0.23481276
delta_pq     -0.51729787
```

```
LOD_H3          0.29268801
H4_pvalue       0.01402575
H5_pvalue       0.03295641
H6_pvalue       0.89215754
model           0.00000000
attr(,"class")
[1] "fullsib.char" "matrix"
```

O conteúdo da variável `qtl1` é uma matriz com 1 coluna e 15 linhas. A primeira linha contém, o grupo de ligação que está região está localizada, a segunda indica a posição em centiMorgan que está região está localizada. Na terceira e quarta coluna apresentam-se respectivamente os $-log_{10}(pvalue)$ and LOD Score, valores similares aos obtidos com a função `im.scan`. A quinta linha ("mu") indica o efeito do intercepto do modelo. As linhas "alpha.p", "alpha.q" e "delta.pq" representam os efeitos genéticos dos QTLs e "LOD.alpha.p", "LOD.alpha.q" e "LOD.delta.pq" correspondem aos testes marginais destes efeitos. Vale lembrar que este três testes foram feitos com 1 grau de liberdade (testando apenas se um efeito difere de zero), nesse caso um LOD de 0,83 indicaria se um dado efeito é significativo (LOD=0,83 corresponde a LRT=3.81 que sob a distribuição de χ^2 com 1 Grau de liberdade tem ível de significancia de 5%). As linhas "H4.pvalue", "H5.pvalue" e "H6.pvalue" indicam os p-valores para os testes complementares necessários para identificar a segregação dos QTLs. Para maiores detalhes ver Gazaffi et al. (2013). Finalmente, a linha "model" indica qual o modelo foi possível de ser estudado, detalhes na seção 9. Para as situações em que alguns dos efeitos e/ou testes não podem ser conduzidos devido a falta de informatividade dos marcadores, as células contém NA.

A identificação das fases de ligação permite identificar a localização dos alelos que aumentam ou reduzem o fenótipo, no caso é feito com a interpretação dos valores de "LOD.alpha.p" e "LOD.alpha.q". Para facilitar esse procedimento, a função `draw.phase` foi desenvolvida e pode ser utilizada conforme a seguir:

```
> draw.phase(fsib, qtl1, 0.05)

Printing QTL and its linkage phase between markers across the LG:

Markers          Position          Parent 1       Parent 2

M1                 0.00           a |  | o       a |  | o
QTL                4.00          P1 |  | P2     Q0 |  | Q0
M2                10.00           a |  | b       c |  | d
M3                20.00           c |  | c       a |  | b
```

```
M4                30.00          a |  | o       a |  | o
M5                40.00          a |  | b       a |  | b
M6                50.00          a |  | b       a |  | b
M7                60.00          a |  | o       a |  | o
M8                70.00          a |  | o       a |  | o
M9                80.00          a |  | b       c |  | d
M10               90.00          a |  | b       c |  | c
M11              100.00          a |  | b       c |  | d
M12              110.00          a |  | b       a |  | o
M13              120.00          a |  | b       c |  | c
M14              130.00          a |  | b       c |  | d
M15              140.00          a |  | b       c |  | d

P1 and Q1 has positive effect (increase phenotypic value)
P2 and Q2 has negative effect (reduce phenotypic value)
P0 and Q0 has neutral effect (non signif.)
```

O argumento alpha igual a 0.05 indica que para indicar os efeitos aditivo dos genitores, caso sejam superiores a 0.05. Caso isto ocorra, o alelo com super escrito 1 aumenta o fenótipo e o sobre escrito 2 reduz, alelos com sobre escrito 0 tem efeito nao significativo considerando o nível de significancia (alpha) de 95%

Para inferir a segregação do QTL deve-se utilizar a função `get.segr`, por default considera-se que os efeitos marginais ("LOD.alpha.p", "LOD.alpha.q" e "LOD.delta.pq") devem ser significativo considerando alpha de 95%, assim como os testes complementares também a 5% ("H4.pvalue", "H5.pvalue" e "H6.pvalue"). O uso da função está exemplificado abaixo:

```
> get.segr(qtl1)

QTL segregation is 1:1

> get.segr(qtl2)

QTL segregation is 1:2:1

> get.segr(qtl3)

QTL segregation is 1:1

> get.segr(qtl4)

QTL segregation is 1:2:1
```

6 Cofactor selection

Antes de realizar o mapeamento por intervalo composto é necessário definir os marcadores que serão utilizados como cofator. Para tanto, há disponível duas funções: `cof.selection` e `cof.definition` que serão apresentadas a seguir:

6.1 Seleção dos cofatores usando regressão linear múltipla

A função `cof.selection` permite selecionar cofatores através do uso regressão linear multipla via stepwise e para determinar a entrada e/ou saída dos cofatores utiliza-se o critério de informação (akaike ou bayesiano). Essencialmente, esta função prepara os dados para o processo de seleção que é realizado pela função `step` nativa do R. A sua utilização da forma mais simplificada é mostrada abaixo:

```
> cofs.fs <- cof.selection(fsib, pheno.col=1, k = log(300), n.cofactor=10)
```

O argumento `fsib` foi obtido pelo uso da função `create.fullsib` e `pheno.col` indica o fenótipo utilizado para a análise. `k` é a penalidade utilizada para o cálculo do critério de informação. Se `k=2` corresponde ao critério de informação de Akaike, caso seja $log(n)$ (sendo n o tamanho da população de mapemento), corresponde ao critério de informação Bayesiano. Caso o usuário queira considerar outros valores de penalidade, recomenda-se ver o manual do software QTL Cartographer v1.17 (http://statgen.ncsu.edu/qtlcart/manual.pdf), página 76 e o help da função `step`. Para algumas situações o processo de seleção pode recomendar a entrada de muitos cofatores, levando a super parametrização do modelo, logo o argumento `n.cofactor` permite que o usuário restrinja o número máximo de cofatores selecionados no modelo, ou seja, o procedimento de entrada de cofatores caso não seja parado naturalmente é paralizado ao atingir o número definido em `n.cofactor` (o default também é considerar limite de 10 cofatores).

O exemplo a seguir mostra como proceder a seleção de cofatores, para situações que covariaveis fixas são utilizadas para controlar a variabilidade do caráter em estudo, utilizando o argumento `addcovar`, utilizando de forma análoga como na função `im.scan`:

```
> covar <- matrix(rep(c(1,-1), each=150), ncol=1)
> cofs.fs2 <- cof.selection(fsib, pheno.col=2, addcovar=covar, k=2, thres.effect=0.0
```

A opção `thres.effect=0.05` verifica ao final da seleção dos cofatores se há efeitos não significativos (com pvalores maiores que 0.05). Caso isto ocorra, os efeitos não significativos são removidos mantendo-se no modelo apenas os efeitos significativos para os cofatores. Isto é útil para a redução do número de graus de liberdade do modelo, uma vez que cada cofator pode ter até 3 efeitos genéticos, em que nem todos os três podem ser significativos. O default

desse argumento é 1, ou seja, não proceder nenhuma remoção de marcadores não significativos (recomendamos que se esse argumento for utilizado que seja por critério do usuário e não por default do programa.

Por fim, o usuário pode visualizar os marcadores digitando o nome da variável que recebeu a seleção dos cofatores ou graficamente através do uso da função `plot`:

```
> cofs.fs

  This is an object of class 'fullsib'

  The linkage map has 4 groups, with 560 cM and 60 markers
  No. individuals genotyped:  300
  Group 1 : 140 cM,  15 markers (A, B1, B3, C, D1, D2)
  Group 2 : 140 cM,  15 markers (A, B1, B2, B3, C, D1, D2)
  Group 3 : 140 cM,  15 markers (A, B2, B3, C, D1, D2)
  Group 4 : 140 cM,  15 markers (A, B1, B2, C, D1, D2)
  And 5 unlinked markers



  2 phenotypes are avaliable for QTL mapping
  Multipoint probability for QTL genotype was obtained for each 1 cM



  Cofactor selection was done for pheno.col = 1
  Markers selected for CIM analysis: 8
   LG     Marker
   1      M13
   1      M2
   2      M20
   2      M24
   2      M27
   3      M33
   3      M37
   4      M52

> plot(cofs.fs)
```

Para outros exemplos de uso desta função recomenda-se que o usuário veja o help page da função `cof.selection`.

6.2 Definição (ad-hoc) dos marcadores cofatores

cof.definition() seleção com P-valor por conta e risco

7 CIM

7.1 varredura

7.2 permutacao

7.3 caracterização

7.4 r2.ls

8 Proposed exercise

8.1 mapa

8.2 QTL

9 QTL mapping with partially informative markers

The model presented by Gazaffi et al. (2013) was developed considering a given locus that segregates in 1:1:1:1 fashion. This would result in four different conditional probabilities for QTL genotypes (P^1Q^1, P^1Q^2, P^2Q^1, P^2Q^2). However, for situations in which the genetic map is composed with partially informative markers may be limitations on the level of information that the odds may contain.

Uma situação muito comum que isto pode ocorrer é se uma região do genoma for

que apresentava quatro classes genotípicas Inicialmente, o modelo genético-estatístico (Equação **??**) foi desenvolvido considerando um loco que apresentava quatro classes genotípicas, sendo possível de ser modelado através de três efeitos genéticos. A segregação do QTL é inferida através do número de efeitos genéticos significativos e por suas magnitudes [**?**]. Entretanto, algumas regiões de um mapa de ligação pode não ser informativas para a estimação de todos os efeitos genéticos, mesmo utilizando probabilidades multiponto. No presente trabalho, por exemplo, isto pode acontecer quando grupos de ligação são compostos por apenas um ou dois tipos de marcadores (C e D_1 ou C e D_2).

Um procedimento em cada posição do mapa genético pode ser realizada, visando identificar regiões com problemas de colinearidade entre as probabilidades condicionais dos QTLs. O critério adotado foi a utilização em medidas de diagnóstico baseadas na decomposição de valor

singular [**?**], utilizando o indíce de condição. Apesar dos autores sugerirem que problemas de colinearidades ocorrem para indíces de condição acima de 30, resultados preliminares sugerem que para o presente contexto, a utilização de 3.5 como threshold fornece bons resultados.

Nas situações em que não se verifica problemas com colinearidade utilizam-se os constrastes definidos nas colunas 1, 2 e 3 (matriz $\mathbf{D}$) para a obtenção do modelo de mapeamento. A segregação e a fase de ligação são inferidos de acordo com procedimento apresentados por [**?**] Porém, quando identifica-se alguma relação de dependência entre as probabilidades dos genótipos dos QTLs sugere-se a remoção ou adaptação dos contrastes utilizados para definição dos efeitos genéticos (colunas 4 a 13).

De forma sucinta, quando considera-se os três contrastes para realizar o mapeamento de QTLs, a segregação é inferida com base em hipóteses complementares. Primeiramente, deve-se verificar a significância dos efeitos α_p^*, α_q^* e δ_{pq}^* através das hipóteses, $H_{01} : \alpha_p^* = 0$, $H_{02} : \alpha_q^* = 0$ e $H_{03} : \delta_{pq}^* = 0$, respectivamente. Nas situações em que observa-se mais de um efeito significativo, são consideradas hipóteses para verificar a igualdade das estimativas: Hipóteses $H_{04} : \alpha_p^* = \alpha_q^*$ (ou $\alpha_p^* = -\alpha_q^*$, ou $-\alpha_p^* = \alpha_q^*$, ou $-\alpha_p^* = -\alpha_q^*$), $H_{05} : \alpha_p^* = \delta_{pq}^*$ (ou $\alpha_p^* = -\delta_{pq}^*$, ou $-\alpha_p^* = \delta_{pq}^*$, ou $-\alpha_p^* = -\delta_{pq}^*$) e $H_{06} : \alpha_q^* = \delta_{pq}^*$ ou ($\alpha_q^* = -\delta_{pq}^*$, ou $-\alpha_q^* = \delta_{pq}^*$, ou $-\alpha_q^* = -\delta_{pq}^*$). Caso apenas um efeito genético for significativo, assumi-se que o QTL tem segregação 1:1. Para as situações em que há dois efeitos genéticos significativos, um teste complementar é realizado para inferir se a segregação é 1:2:1 (Hipótese complementar não-rejeitada) ou 1:1:1:1 (Hipótese complementar rejeitada). Se os três efeitos forem significativos, considera-se que o QTL tem segregação 3:1 se as três hipóteses complementares não forem rejeitadas, 1:2:1 se apenas uma dentre as três hipóteses forem não-rejeitadas e 1:1:1:1 para os demais casos. Já a fase de ligação entre QTL e marcadores é inferida através da interpretação dos sinais de α_p^* e α_q^*. Para a definição do modelo foi considerada a configuração $P_m^1 P^1 P_{m+1}^1 / P_m^2 P^2 P_{m+1}^2 \times Q_m^1 Q^1 Q_{m+1}^1 / Q_m^2 Q^2 Q_{m+1}^2$, a qual é definida como associação em ambos genitores, caso isto ocorra as estimativas de α_p^* e α_q^* são positivas, nas situações em que outras configurações sejam verificadas, as estimativas apresentaram sinais invertidos [**?**].

Por outro lado, em regiões pouco informativas do genoma, não é possível considerar um modelo com todos os parâmetros, nesse caso há limitações para caractizar corretamente os QTLs. Na Tabela 2 é possível inferir a partir das relações de dependência entre os genótipos dos QTLs quais são os efeitos estimáveis, assim como suas esperança e também a possível caracterização do QTL. Para os casos i e ii, somente é possível estimar um dos efeitos genéticos dos genitores, o que permitiria detectar um QTL com segregação 1:1, análogo ao encontrado em uma população de retrocruzamento. Contudo, algumas considerações devem ser salientadas: i) Para um QTL localizado numa região com este tipo de colinearidade, somente é possível testar a presença de um QTL segregando em um genitor, nesse caso, mesmo que o QTL apresentasse mais efeitos significativos não seriam possíveis de serem estimados, devido a falta de

informatividade dos marcadores, indicando que a segregação observada do QTL na progênie seja parcialmente condizente, com a segregação real do QTL. ii) Apesar de algumas regiões permitirem detectar um QTL com efeito genético similar a um retrocruzamento, o que poderia ser comparado com a abordagem de *duplo pseudo testcross* em progênies de irmãos completos, porém o presente modelo tem como vantagem a possibilidade de incorporar como cofatores, regiões com diferentes padrões de segregação em relação ao local em mapeamento, aumentando assim o poder estatístico para a detecção de QTLs.

Para o caso iii é possível estimar o efeito α_p^*, mas devido às probabilidades condicionais aos genótipos P^1Q^1 e P^1Q^2 serem iguais, estatisticamente não seria possível estimar separadamente os efeitos α_q^* e δ_{pq}^*. Neste caso, a estratégia foi testar a existência de um efeito genético α_q^*, dentro dos genótipos informativos (P^2Q^1 e P^2Q^2), representado pelo contraste 4 indicado na matriz **D**. Entretanto, este contraste testa uma possível combinação entre os efeitos α_q^* e δ_{pq}^*, o que dificultaria a identificação a origem do efeito significativo e consequentemente a sua correta segregação. Neste caso, se apenas o efeito α_p^* for significativo, o QTL segregaria nesta progênie com proporção 1:1. Se apenas o efeito α_q^* for significativo a segregação observada na progênie seria na proporção 1:2:1. Para os casos em que ambos efeitos forem significativos a segregação observada será 1:2:1 ou 3:1. O padrão 3:1 surge quando a estimativa para α_q^* (contrastes modificado) é o dobro do valor de α_p^* (contrastes original), independente de seu sinal. Nota-se que neste tipo de colinearidade, a identificação da segregação do QTL, assim como sua fase de ligação entre os marcadores é parcialmente informativa, pois estão condicionados a marcadores que não permitem estimar todos os efeitos genéticos separadamente. Consequentemente, a futura integração de marcadores com diferentes padrões de segregação em regiões de colinearidade poderiam fornecer resultados mais precisos para o mapeamento de QTLs, inclusive alterando o tipo da segregação inferida. Esta mesma situação pode ser extrapolada para os casos iv, v, vi, vii e viii.

Para o caso viii, se for assumido uma autofecundação ou cruzamento entre dois genitores geneticamente iguais, pode-se considerar uma situação análoga a um F_2, assim α_p^* e α_q^* seriam iguais, fornecendo uma estimativa não viesada para a segregação, nem para a fase de ligação entre QTLs e marcadores ($P^1P^2 \times Q^1Q^2$, se α_p for positivo ou $P^2P^1 \times Q^2Q^1$, se α_p for negativo). Para obter a estimativa de dominância, conforme proposto por [**?**], deve-se considerar metade do efeito δ_{pq}^*.

Para os casos de ix a xii, a colinearidade é originada pela semelhança entre três classes genotípicas, o que permitiria apenas estimar um efeito genético que representaria uma combinação linear entre os efeitos α_p^*, α_q^* e δ_{pq}^*, o que inviabiliza o correto entendimento da segregação e o reconhecimento das fases de ligação entre QTLs e marcadores. No presente estudo, considerou-se apenas o caso ix, quando identificado colinearidade entre três genótipos, pois ao considerar o tamanho amostral utilizado ($n = 100$) a detecção de marcadores do tipo C ligados em repulsão

são difíceis de serem observados. Ainda, ressalta-se que para definir estes contrastes adotou-se o coeficiente $1/3$ para os genótipos que são similares e -1 para o genótipo que difere das demais classes genotípicas, sendo que estes contrastes não são ortogonais em relação aos demais, o que pode tornar oneroso o mapeamento de QTLs, por exemplo, através de modelos MIM visando estudar epistasia, logo outras estratégias podem ser consideradas quando este tipo de situação é verificada.

Tabela 2: Identificação dos efeitos estimáveis e caracterização dos QTLs, a partir das possíveis relações de colinearidade existente entre os genótipos do QTL mapeado

Case	Colinearity	**D** matrix Contrasts	Estimable effects	Contrast Espected mean	Rejected Hypothesis	Observed Segregation[(a)]
i	$P^1Q^1 = P^1Q^2$ and $P^2Q^1 = P^2Q^2$	(1)	α_p^*	α_p^*	H_{01}: $\alpha_p^* = 0$	1:1
ii	$P^1Q^1 = P^2Q^1$ and $P^1Q^2 = P^2Q^2$	(2)	α_q^*	α_q^*	H_{02}: $\alpha_q^* = 0$	1:1
iii	$P^1Q^1 = P^1Q^2$	(1)	α_p^*	α_p^*	H_{01}	1:1
		(4)	α_q^*	$(1/2)(\alpha_q^* - \delta_{pq}^*)$	H_{02}	1:2:1
					H_{01} and H_{02}	1:2:1 or 3:1
iv	$P^2Q^1 = P^2Q^2$	(1)	α_p^*	α_p^*	H_{01}	1:1
		(5)	α_q^*	$(1/2)(\alpha_q^* + \delta_{pq}^*)$	H_{02}	1:2:1
					H_{01} and H_{02}	1:2:1 or 3:1
v	$P^1Q^1 = P^2Q^1$	(6)	α_p^*	$(1/2)(\alpha_p^* - \delta_{pq}^*)$	H_{01}	1:2:1
		(2)	α_q^*	α_q^*	H_{02}	1:1
					H_{01} and H_{02}	1:2:1 or 3:1
vi	$P^1Q^2 = P^2Q^2$	(7)	α_p^*	$(1/2)(\alpha_p^* + \delta_{pq}^*)$	H_{01}	1:2:1
		(2)	α_q^*	α_q^*	H_{02}	1:1
					H_{01} and H_{02}	1:2:1 or 3:1
vii	$P^1Q^1 = P^2Q^2$	(8)	α_p^*	$(1/2)(\alpha_p^* - \alpha_q^*)$	H_{01}	1:2:1
		(3)	δ_{pq}^*	δ_{pq}^*	H_{03}: $\delta_{pq}^* = 0$	1:1
					H_{01} and H_{03}	1:2:1 or 3:1
viii	$P^1Q^2 = P^2Q^1$	(9)	α_p^*	$(1/2)(\alpha_p^* + \alpha_q^*)$	H_{01}	1:2:1
		(3)	δ_{pq}^*	δ_{pq}^*	H_{03}	1:1
					H_{01} and H_{03}	1:2:1 or 3:1
ix	$P^1Q^1 = P^1Q^2 = P^2Q^1$	(10)	α_p^*	$(4/3)(\alpha_p^* + \alpha_p^* - \delta_{pq}^*)$	H_{01}	3:1
x	$P^1Q^1 = P^1Q^2 = P^2Q^2$	(11)	α_p^*	$(4/3)(\alpha_p^* - \alpha_p^* + \delta_{pq}^*)$	H_{01}	3:1
xi	$P^1Q^1 = P^2Q^1 = P^2Q^2$	(12)	α_p^*	$(4/3)(-\alpha_p^* + \alpha_p^* + \delta_{pq}^*)$	H_{01}	3:1
xii	$P^1Q^2 = P^2Q^1 = P^2Q^2$	(13)	α_p^*	$(4/3)(-\alpha_p^* - \alpha_p^* - \delta_{pq}^*)$	H_{01}	3:1

[(a)]: Devido a presença de colinearidade entre as probabilidades dos genótipos dos QTLs não é possível estimar a real segregação do QTL, uma vez que nem todos os efeitos são possíveis de serem estimados de forma independente, assim o QTL pode ser caracterizado apenas parcialmente, a partir da informação disponível.

[(b)]: For clarity reasons, only the QTL alleles are represented; P^1P^2 represents the configuration $P_m^1P^1P_{m+1}^1/P_m^2P^2P_{m+1}^2$, and so on.

[(c)]: Caso seja considerado um cruzamento entre dois genitores P e Q geneticamente iguais (ou autofecundação), a segregação real seria equivalente a segregação observada na população de mapeamento (1:2:1), com fases de ligação conhecida entre QTLs e marcas $P^1P^2 \times Q^1Q^2$ ou $P^2P^1 \times Q^2Q^1$.