QTLMap 0.9.6

User's guide

21/06/13

1. Introduction

QTLMap is a software dedicated to the detection of QTL from **experimental designs** in **outbred population**. QTLMap software is developed at INRA (French National Institute for Agronomical Research). The statistical techniques used are linkage analysis (LA), linkage disequilibrium analysis (LD) and linkage disequilibrium linkage analysis (LDLA) using **interval mapping**. Different versions of the LA are proposed from a quasi Maximum Likelihood approach to a fully linear (regression) model. The LDLA and LD analyses are regression approaches (Legarra and Fernando, 2009). The population may be sets of **half-sib families** or **mixture of full- and half- sib families in daughter or grand-daughter design**. The computations of **Phase and Transmission probabilities** are optimized to be rapid and optimised (Elsen *et al*., 2011; Favier *et al*., 2010). QTLMap is able to deal with large numbers of markers (SNP) and traits (eQTL).

QTLMap sources (Fortran language) are freely available.
Up to now, the following functionnalities have been implemented:

- ✓ daughter or grand daughter design
- ✓ QTL detection in half-sib families or mixture of full- and half-sib families
- ✓ One or several linked QTL segregating in the population
- ✓ Single trait or multiple trait analyses
- ✓ Nuisance parameters (e.g. sex, batch, weight...) and their interactions with QTL can be included in the analysis
- ✓ Gaussian, discrete or survival (Cox model) data
- ✓ Familial heterogeneity or homogeneity of variances (homo/heteroscedasticity)
- ✓ Can handle eQTL analyses
- ✓ Computation of transmission and phase probabilities adapted to high throughput genotyping (SNP)
- ✓ Empirical thresholds are estimated using simulations under the null hypothesis or permutations of trait values
- ✓ Computation of power and accuracy of your design or any simulated designs

2. Contributors

Pascale Le Roy, UMR1348 PEGASE, INRA, Rennes, France
Jean-Michel Elsen, UR0631 SAGA, INRA, Toulouse, France
Hélène Gilbert, UMR0444 LGC, INRA, Jouy-en-Josas, France
Carole Moreno, UR0631 SAGA, INRA, Toulouse, France
Andres Legarra, UR0631 SAGA, INRA, Toulouse, France
Olivier Filangi, UMR1348 PEGASE, INRA, Rennes, France

3. Support

Subsribe and post any message/question to the qtlmap-users list:

mailto:qtlmap-users@listes.inra.fr

4. Theoretical background

QTLMap is a software dedicated to marker assisted genetic dissection of quantitative traits recorded in experimental populations.

Typically the analysed populations must be presented as a collection of full or half-sib families each comprising a sire ($i = 1 \cdots ns$) and its mates ($j = 1 \cdots nd_i$) each giving birth to one or more progenies $\left(k = 1 \cdots np_{ij}\right)$.There is a total of ns sires, $nd = \sum_i nd_i$ dams and $np = \sum_i np_i$ (with $np_i = \sum_j np_{ij}$) progenies. The parents form the G1 generation, and their progenies the G2 generation. Extra data may be given about the grand parents (G0), their ancestors (G-1) and the descendants (G3+) of the G2 generation.

4.1. The basic model

4.1.1. Information

Three groups of information are needed in the analysis.

The **pedigree information** ***P*** describes the familial structure along the generations, *i.e.*, for each individual (say the l^{th} in the list), its ID (P_l) and the ID of its sire ($P_{s(l)}$) and dam ($P_{d(l)}$). The only mandatory information are the trios ($P_l, P_{s(l)}, P_{d(l)}$) of G2 individuals.This information is assembled in a "pedigree" file. Animals without parental information are the founders and do not figure in this pedigree file. When available and useful, information about other generations (G-1, G0, G1, G3+) may be given.The table lists these extra cases

Available extra information	The file containing the trio $P_l, P_{s(l)}, P_{d(l)}$ must be given for $l \in$				
	G-1	G0	G1	G2	G3+
Full pedigree	Yes	Yes	Yes	Yes	No
G1 markers	No	No	Yes	Yes	No
G0 and G1 markers	No	Yes	Yes	Yes	No
P3+ phenotype	No	No	No	Yes	Yes

The **marker information** ***M*** describes, for each individual (l), a list of alleles pairs observed at a set of nm markers: $mx_l = \{mx_{lsa}\}_{s=1..nm,a=1,2}$.
The only mandatory information concern the G2 individuals ($mx_l = mp_{ijk}$ for $l = ijk$). However, when available, extra information about G1 (*i.e.* ms_i and md_{ij}) and G0 (mgs_i, mgd_i, mgs_{ij} and mgd_{ij}) will be used.All data concerning the i^{th} sire family (*i.e.* the markers genotypes of the sire, its mates and their progeny) is pooled in the $\boldsymbol{M}_i$ table. In the simplest situation $\boldsymbol{M}_i = \left\{mp_{ijk}, j = 1 \cdots nd_i, k = 1 \cdots np_{ij}\right\}$.

It is important to realize that, before running QTLMap, the parental "phases", that is the way the marker alleles are positionned on their chromosomes, are supposed unknown (in fact, an option allows the input of phased marker phenotypes from external software). If, for instance, sire i is said to carry the marker genotypes A/C, T/G, A/A at loci 1, 2 and 3, the "reading order" which gives the trios ATA and CGA may <u>not</u> be the way the alleles are carried by sire i chromosomes 1 and 2.

The **Trait phenotype information** ***Y*** describes, for each individual (l), a quantitative (possibly discrete) performance, or a vector of nt quantitative performances: $yx_l =$

$\{yx_{lt}\}_{t=1..nt}$.

The only mandatory information are the G2 individuals quantitative traits phenotypes (yp_{ijkt} for $l = ijk$), assembled in a vector $\boldsymbol{yp_i} = \{yp_{ijk}, j = 1 \cdots nd_i, k = 1 \cdots np_{ij}\}$.These vectors form $\boldsymbol{yp} = (\boldsymbol{yp}_1 \cdots \boldsymbol{yp}_i \cdots \boldsymbol{yp}_{ns})$. However, when available, extra information from G3+ will be used.

4.1.2. Statistical formulation

In the basic model of QTLMap the hypothesis is tested that one QTL affecting a single trait is located at a position *x* in a linkage group (*e.g.* a chromosome). Successive positions on this linkage group are scanned. The test is performed with the interval mapping technique applied to an approximation of the likelihood (Knott *et al*, 1996; Elsen *et al*, 1999, Le Roy *et al*,1998).

In this family of modelling, all parents are supposed heterozygous at the QTL, with specific alleles, giving a total of $2(ns + nd)$ QTL effects α_{la} $(a = 1,2)$: $\alpha_{la} = \alpha_{ia}$ $(i = 1 \cdots ns)$ for the sires and $\alpha_{la} = \alpha_{ija}$ $(j = 1 \cdots nd_i)$ for the dams. When the l^{th} parent is homozygous at the QTL we get $\alpha_{l1} = \alpha_{l2}$, a situation which may be statistically tested.

An other parametrization of the model describes performances expectations as the sum of parental means values μ_l ($l = i$ or ij) and deviations α_l to this mean due to the QTL, with $\alpha_{l1} = \mu_l - \alpha_l$ and $\alpha_{l2} = \mu_l + \alpha_l$ which can be summarized by $\alpha_{la} = \mu_l + (-1)^a \alpha_l$.It was proposed by Soller and Genizi (1974), but not kept here.

In the basic model, it is assumed that the parents are unrelated, the markers in linkage equilibrium and the trait normally distributed.

As proposed by Goffinet *et al* (1999) in the case of populations structured in half sib families, and by Le Roy *et al* (1998) when the population is a mixture of half and full sib families, the residual variance of the quantitative trait σ_i^2 is estimated within sire. This heteroskedastic parametrization better fits different (between sires) patterns of segregation of other QTLs, unlinked to the tested position.

The likelihood is given by

$$L1^x = \prod_{i=1}^{ns} \sum_{hs_i} p(\boldsymbol{hs_i} / \boldsymbol{M}_i) \prod_{j=1}^{nd_i} \sum_{hd_{ij}} p(\boldsymbol{hd_{ij}} / \boldsymbol{hs_i}, \boldsymbol{M}_i) \prod_{k=1}^{np_{ij}} L1_{ijk}^x$$

With $\boldsymbol{hs_i}$ and $\boldsymbol{hd_{ij}}$ the sire and dam phases, $\boldsymbol{M}_i$ the marker information for the ith sire family.In the full likelihood, the element $L1_{ijk}^x$ is:

$$L1_{ijk}^x = \sum_{t_s=1,2} \sum_{t_d=1,2} p(\boldsymbol{t}_{ijk}^x = (t_s, t_d) / \boldsymbol{hs_i}, \boldsymbol{hd_{ij}}, \boldsymbol{M}_i) . \varphi(\mu + \alpha_{it_s} + \alpha_{ijt_d}, \sigma_i).$$

with :

- ✓ $\varphi(\mu, \sigma)$ a normal density with a μ mean and σ^2 variance.
- ✓ μ is the fixed general mean
- ✓ $\boldsymbol{t}_{ijk}^x = (t_{ijks}^x, t_{ijkd}^x)$ the vector of transmission event (1 or 2) from the sire and dam to the

progeny, *i.e.* from which parental chromosome originated the x segment transmitted to ijk

- ✓ α_{it} (resp. α_{ijt}) the effect of the within sire (resp. dam) t^{th} QTL allele
- ✓ σ_i^2 the within sire residual variance.

In the following, the double summation $\sum_{t_s=1,2}\sum_{t_d=1,2}$ will be summarized by $\sum_{t_s,t_d}$.

It must be emphasized that the α_{it} and α_{ijt} effects include both the sire and dam QTL effect and polygenic deviations to the general mean.In the alternative parametrization (Soller and Genizi, 1974), those effects would have been replaced by $\mu_i + (-1)^{t_s}\alpha_i$ and $\mu_{ij} + (-1)^{t_d}\alpha_{ij}$

In QTLMap, this part of the likelihood is linearly approximated by $\widehat{L1_{ijk}^x} = \varphi(\mu_{ijk}^x, \sigma_i)$, with

$$\mu_{ijk}^x = \mu + \sum_{t_s,t_d} p(\boldsymbol{t}_{ijk}^x = (t_s,t_d)/\boldsymbol{hs_i}\,\boldsymbol{hd_{ij}}, \boldsymbol{M}_i).\left[\alpha_{it_s} + \alpha_{ijt_d}\right]$$

As described in Mangin *et al* (1999), this approximation allows a much faster computation of the likelihood, with marginal losses of power and parameters estimation precision (this last point not being true when the number of markers is very limited, 3 in the Mangin et al (1999) paper).

A likelihood ratio test LRT^x compares this $L1^x$ likelihood under the *H1* hypothesis to the $L0$ likelihood under *H0*: "there is no QTL segregating on the linkage group".

$$LRT^x = -2\ln(L0/L1^x)$$

with

$$L0 = \prod_{i=1}^{ns}\prod_{j=1}^{nd_i}\prod_{k=1}^{np_{ij}}\varphi\left(\mu + \alpha_i + \alpha_{ij}, \sigma_i\right).$$

It must be noted that the elements α_i and α_{ij} only represent the sire and dam polygenic deviations to the general mean.

4.1.3. Simplified family structure

In some designs the experimental population is made of sire half-sibs, each dam producing only one progeny. A more frequent situation corresponds to a nested struture of large sire families with very small offspring sizes for the dams. In these situations the dam parameters are very difficult to estimate, and must be omitted in the likelihood formulation. The formulae are adjusted accordingly:

$$L1^x = \prod_{i=1}^{ns}\sum_{hs_i} p(\boldsymbol{hs_i}\,/\,\boldsymbol{M}_i)\prod_{k=1}^{nd_i} L1_{ik}^x$$

$$L1_{ik}^x = \sum_{t_s=(1,2)} p(t_{ik}^x = t_s/\boldsymbol{hs_i}\,, \boldsymbol{M}_i).\,\varphi\left(\mu + \alpha_{it_s}, \sigma_i\right).$$

t_{ik}^x the transmission event (1 or 2) from the sire to the progeny

Again, this part of the likelihood is linearly approximated by $\widehat{L1_{ik}^{x}} = \varphi(\mu_{ik}^{x}, \sigma_i)$, with

$$\mu_{ik}^{x} = \mu + \sum_{t_s=(1,2)} p(t_{ik}^{x} = t_s / \boldsymbol{hs_i}, \boldsymbol{M}_i).\alpha_{it_s}$$

Under *H0* the likelihood $L0$ becomes

$$L0 = \prod_{i=1}^{ns} \prod_{k=1}^{nd_i} \varphi(\mu + \alpha_i, \sigma_i).$$

4.1.4. Computation of elements

4.1.4.1. Parental phases

In the current version of QTLMap, only the most probable sire phase (given the $\boldsymbol{M}_i$) is considered : sire families being large, it is supposed that enough information is available for a correct phase inference.The efficiency of this approximation was demonstrated in Mangin *et al* (1999). Practically, finding the most probable phase can be described as the maximisation of a quadratic function of binary variables (Favier *et al*, 2010). This optimisation belongs to the Binary Weighted Constraint Satisfaction Problem area, making possible the use of a very efficient algorithm (Larrosa and Schiex, 2004).

Two strategies are proposed for the computation of the dam phase probability, $p(\boldsymbol{hd_{ij}} / \boldsymbol{hs_i}, \boldsymbol{M}_i)$. When the number of markers on the linkage group is small (less than 15), possible phases can be exhaustively listed and all phase probabilities estimated. When this number is high, the BWCSP approach is used, giving only the most probable dam phase.

4.1.4.2. Parents to progeny transmission

Probabilities of transmission $p(\boldsymbol{t}_{ijk}^{x} / \boldsymbol{hs_i}, \boldsymbol{hd_{ij}}, \boldsymbol{M}_i)$ are calculated following Elsen *et al* (2009) algorithm. This algorithm needs only very limited computational resources, both in terms of time and space. It limits the exploration of the linkage group to the markers informative for a given position to be traced, and thus performs very fast.

It must be noted that $p(\boldsymbol{t}_{ijk}^{x} / \boldsymbol{hs_i}, \boldsymbol{hd_{ij}}, \boldsymbol{M}_i)$ is a joint probability of transmission events and not the product of the marginals. Indeed, when all genotypes (for parents and progeny) at a marker are heterozygous and identical (say 1/2), the origins of the alleles received by the progeny are not independant($prob(t_s = 1, t_d = 2) = 0.5 \neq prob(t_s = 1) \times prob(t_d = 2) = 0.25$).

4.1.4.3. Penetrance

In the basic model, the exponent of the exponential in the penetrance $\widehat{L1_{ijk}}$ is algebraically developped and elementary statistics (ES) are computed only once, allowing a fast computation of the likelihood during the optimisation process.

The penetrance is $\varphi(\mu_{ijk}^{x}, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} exp\left\{-\frac{1}{2}\frac{\left(yp_{ijk} - \mu_{ijk}^{x}\right)^2}{\sigma_i^2}\right\}$. Its mean μ_{ijk}^{x} can be writen $\mu_i + \mu_{ij} + b_{ijk}^{s}\alpha_i + b_{ijk}^{d}\alpha_{ij}$. The element $\left(yp_{ijk} - \mu_{ijk}^{x}\right)^2$ is

$yp_{ijk}^2 - 2yp_{ijk}\left(\mu_i + \mu_{ij} + b_{ijk}^{s}\alpha_i + b_{ijk}^{d}\alpha_{ij}\right) + (\mu_i + \mu_{ij} + b_{ijk}^{s}\alpha_i + b_{ijk}^{d}\alpha_{ij})^2$

The ES for the part of the likelihood $\prod_{k=1}^{np_{ij}} L1_{ijk}^{x}$ corresponding to a sire i –dam ij –dam phase hd_{ijk} are:

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
ES	$\sum_k yp_{ijk}^2$	$\sum_k yp_{ijk}$	$\sum_k yp_{ijk}\, b_{ijk}^s$	$\sum_k yp_{ijk}\, b_{ijk}^d$	np_{ij}	$\sum_k b_{ijk}^s$	$\sum_k b_{ijk}^d$	$\sum_k b_{ijk}^s b_{ijk}^d$	$\sum_k {b_{ijk}^s}^2$	$\sum_k {b_{ijk}^d}^2$
for	constant	μ_i and μ_{ij}	α_i	α_{ij}	μ_i^2,μ_{ij}^2 and $\mu_i\mu_{ij}$	$\mu_i\alpha_i$, $\mu_{ij}\alpha_i$	$\mu_i\alpha_{ij}$, $\mu_{ij}\alpha_{ij}$	$\alpha_i\alpha_{ij}$	α_i^2	α_{ij}^2

During the optimization process, the space of the unknown parameters $(\mu_i, \mu_{ij}, \alpha_i, \alpha_{ij})$ is explored, using the $\{c_e\}_{e=1,10}$ computed only once to estimate the exponent as

$$c_1 + (\mu_i + \mu_{ij})(-2c_2 + 2c_6\alpha_i + 2c_7\alpha_{ij}) + c_5(\mu_i + \mu_{ij})^2 - 2c_3\alpha_i - 2c_4\alpha_{ij} + 2c_8\alpha_i\alpha_{ij} + c_9\alpha_i^2 + c_{10}\alpha_{ij}^2$$

4.1.5. Optimisation

Parameters of $L0$ likelihood are directly computed using standard formulae,
$\hat{\mu}_i = \sum_j \sum_k yp_{jk}/np_i$
$\hat{\mu}_{ij} = \sum_k yp_{jk}/np_{ij} - \hat{\mu}_i$
$\hat{\sigma}_i^2 = \sum_{jk}(yp_{jk} - \hat{\mu}_i - \hat{\mu}_{ij})^2/np_i$
With $np_i = \sum_j np_{ij}$

Parameters of $L1^x$ likelihood are estimated for each tested QTL position *x*, using a derivative free numerical optimiser. As numerical difficulties may occur depending on the data structure and type of analysis / model chosen by the user, a panel of optimisers are proposed by QTLMap.

4.1.6. Conclusion about the basic model

The computation framework corresponding to the basic model described so far (*i.e.* Within sire/dam regression, using elementary statistics develoment for the computation of the penetrance, and assuming the sire phase correctly found from the marker information) will be qualified the **standard** framework.This is the first we used, following Elsen *et al* (1999), Mangin *et al* (1999) and Goffinet *et al* (1999) recommandations.

Many alternative formulations and enrichments of the analysis were proposed in a second time and are now described.They are available both for the half sib and general family structures.

4.2. Alternative formulations of the likelihood

4.2.1. Full linearisation : the regression model

A fully linear approximation of the likelihood, generalizing the Haley and Knott (1992) or Haley et al (1994) regression models is proposed to the user. Under *H1*, the likelihood is given by

$$L1^x = \prod_{i=1}^{ns} \prod_{j=1}^{nd_i} \prod_{k=1}^{np_{ij}} \varphi(\mu_{ijk}^x, \sigma_i).$$

$$\mu_{ijk}^{x} = \mu + \sum_{hs_i} p(\boldsymbol{hs_i} / \boldsymbol{M}_i) \sum_{hd_{ij}} p(\boldsymbol{hd_{ij}} / \boldsymbol{hs_i}, \boldsymbol{M}_i) \sum_{t_s,t_d} p(\boldsymbol{t}_{ijk}^{x} = (t_s, t_d)/\boldsymbol{hs_i}, \boldsymbol{hd_{ij}}, \boldsymbol{M}_i).\left[\alpha_{it_s} + \alpha_{ijt_d}\right]$$

this is equivalent to $\mu_{ijk}^{x} = \mu + \sum_{t_s,t_d} p(\boldsymbol{t}_{ijk}^{x} = (t_s, t_d)/\boldsymbol{M}_i)\left[\alpha_{it_s} + \alpha_{ijt_d}\right]$.

In this situation, the parameters may be estimated using the standard linear models framework.

Let focus on the ith family: $\boldsymbol{yp}_i = \left(yp_{ijk}\right)_{j=1\cdots n_i, k=1\cdots nd_i}$ is the vector of progeny phenotypes, $\boldsymbol{\alpha}_i = \left(\alpha_{i1}, \alpha_{i2}, \left(\alpha_{ij1}, \alpha_{ij2}\right)_{j=1\cdots nd_i}\right)$ the vectors of unknown first moment parameters, $\boldsymbol{X_i}$ and $\boldsymbol{W}_i^{x}$ the corresponding $nd_i \times (1 + nd_i)$ incidence matrices, and $\boldsymbol{V}_i = \sigma_i^2 \boldsymbol{I}_{np_i \times np_i}$ the covariance matrix.

The non nul $\boldsymbol{W}_i^{x}$ elements are given by $p(\boldsymbol{t}_{ijk}^{x}/\boldsymbol{M}_i)$: the 1^{st} and 2^{nd} elements of the $\boldsymbol{W}_i^{x}$ line corresponding to progeny ijk are $p\left(t_{ijks}^{x} = 1/\boldsymbol{M}_i\right)$ and $p\left(t_{ijks}^{x} = 2/\boldsymbol{M}_i\right)$ while the $(1 + 2j)^{th}$ and $(2 + 2j)^{th}$ elements to $p(t_{ijkd}^{x} = 1/\boldsymbol{M}_i)$ and $p(t_{ijkd}^{x} = 2/\boldsymbol{M}_i)$.Note that in this linear context, the non independance between sire and dam transmissions has not to be considered.

Finally, we have the linear model

$$\boldsymbol{yp}_i = \boldsymbol{X_i}\boldsymbol{\mu_i} + \boldsymbol{W}_i^{x}\boldsymbol{\alpha}_i + \boldsymbol{e}_i$$

with $\boldsymbol{e}_i$ the random residual, supposed to be distributed in $\mathcal{N}(0, \boldsymbol{V}_i)$

Extention to the ns sire families is straightforward.Let
$\boldsymbol{yp} = (\boldsymbol{yp}_1 \cdots \boldsymbol{yp}_i \cdots \boldsymbol{yp}_{ns})$ be the vector of performances,
$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \cdots \boldsymbol{\alpha}_i \cdots \boldsymbol{\alpha}_{ns})$ the vector of QTL effects,
$\boldsymbol{X}' = (\boldsymbol{X'}_1 \cdots \boldsymbol{X'}_i \cdots \boldsymbol{X'}_{ns})$
$\boldsymbol{W^x} = \oplus_{i=1,ns} \boldsymbol{W}_i^{x}$ the incidence matrices extended to the whole set of sires
$\boldsymbol{V} = \oplus_{i=1,ns} \boldsymbol{V_i}$ the total covariance matrix.

The linear model is

$$\boldsymbol{yp} = \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{W^x}\boldsymbol{\alpha} + \boldsymbol{e}$$

The least square equations are (all $\boldsymbol{W_i}$ matrices depend on the x position but the corresponding superscript was omitted) :

$$\begin{pmatrix} \widehat{\boldsymbol{\mu_1}} \\ \widehat{\boldsymbol{\alpha_1}} \\ \widehat{\boldsymbol{\mu_2}} \\ \widehat{\boldsymbol{\alpha_2}} \\ \vdots \\ \widehat{\boldsymbol{\mu_{ns}}} \\ \widehat{\boldsymbol{\alpha_{ns}}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{V}_1^{-1}\boldsymbol{W}_1 & 0 & 0 & & 0 & 0 \\ \boldsymbol{W}_1'\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & \boldsymbol{W}_1'\boldsymbol{V}_1^{-1}\boldsymbol{W}_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \boldsymbol{X}_2'\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \boldsymbol{X}_2'\boldsymbol{V}_2^{-1}\boldsymbol{W}_2 & & 0 & 0 \\ 0 & 0 & \boldsymbol{W}_2'\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \boldsymbol{W}_2'\boldsymbol{V}_2^{-1}\boldsymbol{W}_2 & & 0 & 0 \\ & & \vdots & & \ddots & \vdots & \\ 0 & 0 & 0 & 0 & & \boldsymbol{X}_{ns}'\boldsymbol{V}_{ns}^{-1}\boldsymbol{X}_{ns} & \boldsymbol{X}_{ns}'\boldsymbol{V}_{ns}^{-1}\boldsymbol{W}_{ns} \\ 0 & 0 & 0 & 0 & \cdots & \boldsymbol{W}_{ns}'\boldsymbol{V}_{ns}^{-1}\boldsymbol{X}_{ns} & \boldsymbol{W}_{ns}'\boldsymbol{V}_{ns}^{-1}\boldsymbol{W}_{ns} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{V}_1^{-1}\boldsymbol{yp}_1 \\ \boldsymbol{W}_1'\boldsymbol{V}_1^{-1}\boldsymbol{yp}_1 \\ \boldsymbol{X}_2'\boldsymbol{V}_2^{-1}\boldsymbol{yp}_2 \\ \boldsymbol{W}_2'\boldsymbol{V}_2^{-1}\boldsymbol{yp}_2 \\ \vdots \\ \boldsymbol{X}_{ns}'\boldsymbol{V}_{ns}^{-1}\boldsymbol{yp}_{ns} \\ \boldsymbol{W}_{ns}'\boldsymbol{V}_{ns}^{-1}\boldsymbol{yp}_{ns} \end{pmatrix}$$

In this case where all parameters are defined within sire family, the equations simplified, for each sire, to

$$\begin{pmatrix} \widehat{\boldsymbol{\mu}}_i \\ \widehat{\boldsymbol{\alpha}}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_i'\boldsymbol{X}_i & \boldsymbol{X}_i'\boldsymbol{W}_i \\ \boldsymbol{W}_i'\boldsymbol{X}_i & \boldsymbol{W}_i'\boldsymbol{W}_i \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}_i'\,\boldsymbol{yp}_i \\ \boldsymbol{W}_i'\,\boldsymbol{yp}_i \end{pmatrix}$$

And the variances are estimated by

$$\hat{\sigma}_i^2 = (\boldsymbol{yp}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}}_i - \boldsymbol{W}_i\hat{\boldsymbol{\alpha}}_i)'(\boldsymbol{yp}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}}_i - \boldsymbol{W}_i\hat{\boldsymbol{\alpha}}_i)/np_i$$

4.2.2. More complex models

If linear models are much easier to handle, estimations may be more accurate when using mixture models (*e.g.* Knott *et al*, 1996 or Mangin *et al*, 1999). Four levels of likelihood linearization will be available in QTLMap.

Full mixture, exact likelihood

$$\prod_{i=1}^{ns}\sum_{\boldsymbol{hs}_i} p(\boldsymbol{hs_i}/\boldsymbol{M}_i)\prod_{j=1}^{nd_i}\sum_{\boldsymbol{hd}_{ij}} p(\boldsymbol{hd_{ij}}/\boldsymbol{hs_i},\boldsymbol{M}_i)\ \prod_{k=1}^{np_{ij}}\sum_{t_s,t_d} p(\boldsymbol{t}_{ijk}^x=(t_s,t_d)/\boldsymbol{hs_i}\,\boldsymbol{hd_{ij}},\boldsymbol{M}_i).\varphi\left(\mu+\alpha_{it_s}+\alpha_{ijt_d},\sigma_i\right)$$

Within sire/dam regression (the basic model)

$$\prod_{i=1}^{ns}\sum_{\boldsymbol{hs}_i} p(\boldsymbol{hs_i}/\boldsymbol{M}_i)\prod_{j=1}^{nd_i}\sum_{\boldsymbol{hd}_{ij}} p(\boldsymbol{hd_{ij}}/\boldsymbol{hs_i},\boldsymbol{M}_i)\ \prod_{k=1}^{np_{ij}}\varphi\left(\mu+\sum_{t_s,t_d} p(\boldsymbol{t}_{ijk}^x=(t_s,t_d)/\boldsymbol{hs_i}\,\boldsymbol{hd_{ij}},\boldsymbol{M}_i).\left[\alpha_{it_s}+\alpha_{ijt_d}\right],\sigma_i\right)$$

Within sire regression

$$\prod_{i=1}^{ns}\sum_{\boldsymbol{hs}_i} p(\boldsymbol{hs_i}/\boldsymbol{M}_i)\prod_{j=1}^{nd_i}\prod_{k=1}^{np_{ij}}\varphi\left(\mu+\sum_{t_s,t_d} p(\boldsymbol{t}_{ijk}^x=(t_s,t_d)/\boldsymbol{hs_i},\boldsymbol{M}_i).\left[\alpha_{it_s}+\alpha_{ijt_d}\right],\sigma_i\right)$$

With $p(\boldsymbol{t}_{ijk}^x=(t_s,t_d)/\boldsymbol{hs_i},\boldsymbol{M}_i)=\ \sum_{\boldsymbol{hd}_{ij}} p(hd_{ij}/\boldsymbol{hs_i},\boldsymbol{M}_i)p(\boldsymbol{t}_{ijk}^x=(t_s,t_d)/\boldsymbol{hs_i},\boldsymbol{hd_{ij}},M_i)$

Fully linear model

$$\prod_{i=1}^{ns}\prod_{j=1}^{nd_i}\prod_{k=1}^{np_{ij}}\varphi\left(\mu+\sum_{t_s,t_d} p(\boldsymbol{t}_{ijk}^x=(t_s,t_d)/\boldsymbol{M}_i).\left[\alpha_{it_s}+\alpha_{ijt_d}\right],\sigma_i\right)$$

With $p\left(\boldsymbol{t}_{ijk}^x/\boldsymbol{M}_i\right)=\ \sum_{\boldsymbol{hs}_i} p(\boldsymbol{hs_i}/\boldsymbol{M}_i)\ \sum_{\boldsymbol{hd}_{ij}} p(\boldsymbol{hd_{ij}}/\boldsymbol{hs_i},\boldsymbol{M}_i)p(\boldsymbol{t}_{ijk=(t_s,t_d)}^x/\boldsymbol{hs_i},\boldsymbol{hd_{ij}},\boldsymbol{M}_i)$

When only the most probable sire phases are considered, Within sire regression and Fully linear model are identical. When only the most probable dam phases are considered, Within sire/dam regression and Within sire regression are identical. When this restriction is applied to both sire and dam phases, the three last models are identical.

4.3. Alternative genetic hypotheses

4.3.1. Assuming more than one QTL

Extensions of the basic model to more than one QTL acting additively are available.

4.3.1.1. Two linked non interacting QTL

Following Gilbert and Le Roy (2007), the $L1_{ijk}^x$ part of the likelihood is extended to

$$L1_{ijk}^x=\sum_{t_s^1,t_d^1}\sum_{t_s^2,t_d^2} p(\boldsymbol{t}_{ijk}^{x_1}=(t_s^1,t_d^1),\boldsymbol{t}_{ijk}^{x_2}=(t_s^2,t_d^2)/\boldsymbol{hs_i},\boldsymbol{hd_{ij}},\boldsymbol{M}_i).\varphi\left(\mu+\sum\nolimits_{q=1,2}\left[\alpha_{qit_s^q}+\alpha_{qijt_d^q}\right]+,\sigma_i\right)$$

With

- ✓ $\boldsymbol{t}_{ijk}^{x_q}=\left(t_{ijks}^{x_q},t_{ijks}^{x_q}\right)$ the vectors of transmission events (1 or 2) from the sire and dam to the progeny at QTL located at x_q $(q=1,2)$ on the scanned chromosome. The two first summations thus extends on 16 situations.
- ✓ α_{1i} and α_{2i} (resp. α_{1ij} and α_{2ij}) the effects of the QTL located at x_1 and x_2 in the i^{th} sire (resp. ij^{th} dam)

It must be noted that the probability of transmission events is their joint probability (and not the product of marginals) accounting for the linkage between tested positions.

In the current version of QTLMap, this two linked QTL hypothesis is only available in the basic model framework.

4.3.1.2. Two linked epistatic QTL

As the number of parameters to be estimated in this genetic hypothesis may be very large, this option was only made available for half sib family structure where the only sire effect (polygenic and QTL effect) are estimated. In this situation, the $L1_{ik}^{x}$ part of the likelihood is

$$L1_{ik}^{x} = \sum_{t_s^1, t_s^2} p(t_{ik}^{x_1} = \boldsymbol{t}_s^1, t_{ik}^{x_2} = \boldsymbol{t}_s^2 / \boldsymbol{hs_i}, \boldsymbol{M}_i).\varphi\left(\mu + \alpha_{1it_s^1} + \alpha_{2it_s^2} + \alpha\alpha_{\boldsymbol{i}(\boldsymbol{t}_s^1 + 2\boldsymbol{t}_s^2 - 2)}, \sigma_i\right)$$

With

- ✓ $t_{ik}^{x_q}$ the transmission event (1 or 2) from the sire to the progeny at QTL located at x_q $(q = 1,2)$ on the scanned chromosome. The summation thus extends on 4 situations.
- ✓ α_{qit} the within sire i effect of the t^{th} allele at the q^{th} QTL located at x_q
- ✓ $\alpha\alpha_{ih}$ the epistatic effect for the i^{th} sire : $\alpha\alpha_{i1}$ if the sire transmitted alleles 1 and 1, $\alpha\alpha_{i2}$ if it transmitted 2 and 1, $\alpha\alpha_{i3}$ if it transmitted 1 and 2, $\alpha\alpha_{i4}$ if it transmitted 2 and 2.

4.3.1.3. Any number (nq) of unlinked QTL

The amount of computation increasing very rapidly with this number, further approximations were made to face this burden: the fully linearized version of the likelihood is retained and the transmission events are supposed independant between all simultaneously tested QTLs.

With these approximations, the likelihood turns to be:

$$\prod_{i=1}^{ns} \prod_{j=1}^{nd_i} \prod_{k=1}^{np_{ij}} \varphi\left(\mu + \sum_{q=1,n_q} \sum_{t_s^q, t_d^q} p\left(\boldsymbol{t}_{ijk}^{x_q} = (t_s^q, t_d^q)/\boldsymbol{M}_i\right)\left[\alpha_{qit_s^q} + \alpha_{qijt_d^q}\right], \sigma_i\right)$$

4.3.2. Modeling the polygenic background (not yet fully available in QTLMap)

In the basic model, all parents are supposed unrelated, a situation not realistic in livestock populations.When pedigree information (about ancestors) is available, population structure due to familial relationships may be considered in performances description.

This extension is proposed in the fully linearized version of the likelihood. The linear model is extended to:

$$\boldsymbol{yp} = \boldsymbol{X\mu} + \boldsymbol{W^x\alpha} + \boldsymbol{Za} + \boldsymbol{e}$$

where

- ✓ $\boldsymbol{X\mu} = \boldsymbol{1}\mu$ is a column of μ, the general mean**.**
- ✓ $\boldsymbol{\alpha}$ is the vector of QTL effects, following the first parametrization, *i.e.*considering for each parents the two effects α_{i1} and α_{i2} or α_{ij1} and α_{ij2}

- ✓ $\boldsymbol{W^x}$ is the correponding incidence matrix, whose elements are the $p\left(t^x_{ijks}/\boldsymbol{M}_i\right)$ and $p(t^x_{ijkd}/\boldsymbol{M}_i)$
- ✓ $\boldsymbol{a}$ is the random animal effect, distributed in $\mathcal{N}(0, A\sigma^2_a)$
- ✓ $\boldsymbol{Z} = \boldsymbol{I}$
- ✓ $\boldsymbol{e}$ the random residual,distributed in $\mathcal{N}(0, I\sigma^2_e)$

In this mixed linear model, as between families heterogeneity is considered through the A matrix, the homoskedastic situation is kept (only one variance for the residuals).

In principle, the first moments $(\mu, \boldsymbol{\alpha}, \boldsymbol{a})$ and second moments (σ^2_a, σ^2_e) should be estimated at each tested QTL location.

Following the FASTA approach of Aulchenko presented in the GENABEL software (Aulchenko, 2011), QTLMap does not re-estimate the heritability coefficient (more precisely, the ratio $\lambda = \sigma^2_e/\sigma^2_a$) along the genome scan. These parameters must be given by the user, and are easily estimable from standard approaches as ASREML.

At each location, the following mixed model equations are solved

$$\begin{pmatrix} \boldsymbol{X'X} & \boldsymbol{X'W^x} & \boldsymbol{X'} \\ \boldsymbol{W^{x\prime}X} & \boldsymbol{W^{x\prime}W^x} & \boldsymbol{W^{x\prime}} \\ \boldsymbol{X} & \boldsymbol{W^x} & \boldsymbol{I + \lambda A^{-1}} \end{pmatrix} \begin{pmatrix} \mu \\ \boldsymbol{\alpha} \\ \boldsymbol{a} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'yp} \\ \boldsymbol{W^{x\prime}yp} \\ \boldsymbol{yp} \end{pmatrix}$$

Or

$$\left(\begin{pmatrix} \boldsymbol{X'} \\ \boldsymbol{W^{x\prime}} \end{pmatrix} \boldsymbol{H} (\boldsymbol{X} \quad \boldsymbol{W^x}) \right) \begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'} \\ \boldsymbol{W^{x\prime}} \end{pmatrix} \boldsymbol{H\, yp}$$

with $\boldsymbol{H} = \boldsymbol{\lambda}(\boldsymbol{A} + \boldsymbol{\lambda I})^{-1}$, a matrix invariant to the location x which has to be calculated only once.

The residual variance is estimated at each location by

$$\widehat{\sigma^2_e} = \boldsymbol{\lambda} \left(\boldsymbol{yp} - (\boldsymbol{X} \quad \boldsymbol{W^x}) \begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} \right)' (\boldsymbol{A} + \boldsymbol{\lambda I})^{-1} \left(\boldsymbol{yp} - (\boldsymbol{X} \quad \boldsymbol{W^x}) \begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} \right)$$

$$\widehat{\sigma^2_e} = \left(\boldsymbol{yp} - (\boldsymbol{X} \quad \boldsymbol{W^x}) \begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} \right)' \boldsymbol{H} \left(\boldsymbol{yp} - (\boldsymbol{X} \quad \boldsymbol{W^x}) \begin{pmatrix} \hat{\mu} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} \right)$$

The $\boldsymbol{A^{-1}}$ matrix (which gives the $\boldsymbol{H} = \boldsymbol{\lambda A^{-1}}(\boldsymbol{I} + \boldsymbol{\lambda A^{-1}})^{-1}$ matrix) is estimated following the usual Henderson rules

4.4. Alternative penetrance functions

A few alternatives are available for unitrait and multitrait analyses.

4.4.1. Unitrait-uniQTL situations

4.4.1.1. Nuisance effects

The penetrance $\varphi(\mu_{ijk}, \sigma_i)$ may be enriched considering nuisance factors, fixed effects or covariables. In this case, the mean μ_{ijk} becomes

$$\mu_{ijk} = \boldsymbol{X}_{ijk}\boldsymbol{\beta} + \sum_{t_s,t_d} p(\boldsymbol{t}^x_{ijk} = (t_s, t_d)/\boldsymbol{hs_i}, \boldsymbol{hd_{ij}}, \boldsymbol{M}_i).\left[\alpha_{it_s} + \alpha_{ijt_d}\right]$$

With $\boldsymbol{X}_{ijk}$ the incidence vector corresponding to the ijkth progeny, and $\boldsymbol{\beta}_l$ the vector assembling the general mean μ and nuisance factors (fixed effects and covariables).

It is possible to create interactions between nuisance factors (when defined as fixed effects) and the QTL effects.

This extension of the basic model has been implemented for two likelihood options, Within sire/dam regression and Fully linear model. The penetrance $\widehat{L1}_{ijk}$ in the Within sire/dam regression is estimated directly from the classical formulation $\left(\frac{1}{\sqrt{2\pi}\sigma_i} exp\left\{-\frac{1}{2}\frac{\left(yp_{ijk}-\mu^x_{ijk}\right)^2}{\sigma_i^2}\right\}\right)$. Its decomposition in Elementary Statistics should be available soon.

As the nuisance effects may affect performances of individuals belonging to different sire families, the within sire likelihoods are no more independant. The linear model corresponding to the within sire/dam regression is changed accordingly.

Let
$\boldsymbol{yp} = (\boldsymbol{yp}_1 \cdots \boldsymbol{yp}_i \cdots \boldsymbol{yp}_{ns})$ be the vector of performances,
$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \cdots \boldsymbol{\alpha}_i \cdots \boldsymbol{\alpha}_{ns})$ the vector of QTL effects,
$\boldsymbol{X'} = (\boldsymbol{X'}_1 \cdots \boldsymbol{X'}_i \cdots \boldsymbol{X'}_{ns})$
$\boldsymbol{W^x} = \oplus_{i=1,ns}\boldsymbol{W}^x_i$ the incidence matrices extended to the whole set of sires
$\boldsymbol{V} = \oplus_{i=1,ns}\boldsymbol{V}_i$ the total covariance matrix.

The linear model is

$$\boldsymbol{yp} = \boldsymbol{X\beta} + \boldsymbol{W^x}\,\boldsymbol{\alpha} + \boldsymbol{e}$$

Parameters maximising the likelihood can be obtained in an iterative two steps procedure:

At iteration +1 :

- Step 1 = Solving the linear system:

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{(t+1)}\\ \widehat{\boldsymbol{\alpha}}_{(t+1)}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X'V^{-1}}_{(t)}\boldsymbol{X} & \boldsymbol{X'V^{-1}}_{(t)}\boldsymbol{W^x}\\ \boldsymbol{W^{x'}V^{-1}}_{(t)}\boldsymbol{X} & \boldsymbol{W^{x'}V^{-1}}_{(t)}\boldsymbol{W^x}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{X'V^{-1}}_{(t)}\boldsymbol{yp}\\ \boldsymbol{W^{x'}V^{-1}}_{(t)}\boldsymbol{yp}\end{pmatrix}$$

- Step 2 = Estimating the within sire family variances:

$$\hat{\sigma}^2_{i(t+1)} = \left(\boldsymbol{yp}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_{(t+1)} - \boldsymbol{W}^x_i\widehat{\boldsymbol{\alpha}}_{i_{(t+1)}}\right)'\left(\boldsymbol{yp}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_{(t+1)} - \boldsymbol{W}^x_i\widehat{\boldsymbol{\alpha}}_{i_{(t+1)}}\right)/np_i$$

The steps are repeated until convergence, detected for instance when $\left\|\widehat{\boldsymbol{\beta}}_{(t+1)} - \widehat{\boldsymbol{\beta}}_{(t)}\right\| < \epsilon, \left\|\widehat{\boldsymbol{\alpha}}_{(t+1)} - \widehat{\boldsymbol{\alpha}}_{(t)}\right\| < \epsilon$, $\left\|\widehat{\boldsymbol{\sigma^2}}_{(t+1)} - \widehat{\boldsymbol{\sigma^2}}_{(t)}\right\| < \epsilon$

As estimability of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ elements varies with the tested location x, this information is checked for each x before likelihood estimation. This information, initially developped for the fully linear model, is also used in the Within sire-dam regression to avoid numerical difficulties.

To check estimability parameters, the incidence matrix $\boldsymbol{M^x}$ corresponding to the linear model $\boldsymbol{yp = X\beta + W^x\alpha + e = M^x\theta + e}$ is built at each location $\boldsymbol{x}$, choising as an order for the elements of $\boldsymbol{\theta}$, (i) the QTL effects $\boldsymbol{\alpha}$ first, (ii)the parental means $\boldsymbol{\mu}$ and (iii) the nuisance factors $\boldsymbol{\beta}$. A Cholesky decomposition of the incidence matrix is performed eliminating $\boldsymbol{M^x}$ elements corresponding to parameters linearly dependant on previously considered ones[1].

4.4.1.2. Non gaussian models

4.4.1.2.1. Discrete traits

Ordered qualitative phenotypes (codded with integer figures) are analysed using the liability threshold model of Falconer (1989). For the basic model, the penetrance in this case is (*e.g.* Moreno 2003):

$$L1^x_{ijk} = \int_{\lambda_{yp_{ijk}}}^{\lambda_{yp_{ijk}+1}} \frac{1}{\sqrt{2\pi}\sigma_i} exp\left\{-\frac{1}{2}\frac{\left(t-\mu^x_{ijk}\right)^2}{\sigma_i^2}\right\} dt$$

With

- ✓ λ_y and $\lambda_{y+1}$the lower and upper thresholds corresponding to a y phenotype
- ✓ μ^x_{ijk} the expectation of the underlying distribution, which is a linear function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

$$\mu_{ijk} = \boldsymbol{X}_{ijk}\boldsymbol{\beta} + \sum_{t_s,t_d} p(\boldsymbol{t}^x_{ijk} = (t_s,t_d)/\boldsymbol{hs_i}, \boldsymbol{hd_{ij}}, \boldsymbol{M}_i).\left[\alpha_{it_s} + \alpha_{ijt_d}\right]$$

- ✓ σ_i^2 the residual variance for the sire i family

The general picture is that this liability model needs a much longer computing time than the gaussian model but gives similar results (in terms of power and parameters estimations). We recommand the use of this discrete traits approach only when (1) there is very few (2 or 3) classes on the discrete scale, (2) their frequencies are very unequal and (3) the data set is large enough to avoid that only a few individuals represent a given rare class.

4.4.1.2.2. Survival and Time-to-events phenotypes

These phenotypes, also called failure times, describe the length of interval between a point of origine and an end-point. They are characterized by the presence of censored data, *i.e.* indi-

[1] The Cholesky decomposition aims at transforming $\boldsymbol{M^{x\prime}M^x}$ in the product $\boldsymbol{L^{x\prime}L^x}$ with $\boldsymbol{L^x}$ a upper triangular matrix. The transformation is processed using $L^x_{ii} = \sqrt{M^x_{ii} - \sum_{k=1,j-1} {L^x_{ik}}^2}$ and $L^x_{ij} = \left(M^x_{ij} - \sum_{k=1,j-1} L^x_{ik}L^x_{jk}\right)/L^x_{jj}$. Here, the i^{th} line and column of $\boldsymbol{M^{x\prime}M^x}$ (and $\boldsymbol{L^{x\prime}L^x}$) were suppressed when $L^x_{ii} < \varepsilon$.

viduals not reaching the end-point during the recording period. Truncated data are also present for individuals without point of origine. Different approaches were developped for the analysis of such information, including parametric (the Weibull regression model of Kalbfleisch and Prentice, 1980) and semi parametric (the Cox model, Cox, 1972) models. Moreno *et al* (2005) extended those models to QTL detection. QTLMap offers the Cox model for QTL detection. In the extension of Moreno *et al* (2005), the Cox model is approximated to make computations feasible.This model is developped only within the full mixture framework, still assuming that the sire phase is known:

$$\prod_{i=1}^{ns^*}\prod_{j=1}^{nd_i^*}\sum_{\boldsymbol{hd_{ij}}} p(hd_{ij}/\boldsymbol{hs_i},\boldsymbol{M}_i)\ \prod_{k=1}^{np_{ij}^*}\sum_{t_s,t_d} p(\boldsymbol{t}_{ijk}^x=(t_s,t_d)/\boldsymbol{hs_i},\boldsymbol{hd_{ij}},\boldsymbol{M}_i).\varphi\left(\boldsymbol{X}_{ijk}\boldsymbol{\beta}+\alpha_{it_s}+\alpha_{ijt_d},\sigma_i\right)$$

Only uncensored individuals are considered in the ijk list (possibly reducing the numbers ns, nd_i, np_{ij} to ns^*, nd_i^*, np_{ij}^*).The penetrance function $\varphi\left(\boldsymbol{X}_{ijk}\boldsymbol{\beta}+\alpha_{it_s}+\alpha_{ijt_d},\sigma_i\right)$ weights the risk observed for the ijk individual dying at a time yp_{ijk} by the mean risk of individuals still alived at this date.This weighting gives:

$$\frac{exp\left\{\mu_{ijkt_{ijk}^x}\right\}}{\sum_{i_e}\sum_{j_e}\sum_{\boldsymbol{hd_{i_ej_e}}}\left[p(hd_{i_ej_e}\,/\,\boldsymbol{hs_{i_e}},\boldsymbol{M}_{i_e})\times\left(\sum_{k_e\in R(yp_{ijk})}\sum_{\boldsymbol{t}_{i_ej_ek_e}^x} p(\boldsymbol{t}_{i_ej_ek_e}^x/\boldsymbol{hs_{i_e}},\boldsymbol{hd_{i_ej_e}},\boldsymbol{M}_{i_e})\,exp\left\{\mu_{i_ej_ek_e\boldsymbol{t}_{i_ej_ek_e}^x}\right\}\right)\right]}$$

With

- ✓ $\mu_{ijkt_{ijk}^x}=\boldsymbol{X}_{ijk}\boldsymbol{\beta}+\left[(-1)^{t_{ijks}^x}\alpha_i+(-1)^{t_{ijkd}^x}\alpha_{ij}\right]$
- ✓ $R(yp_{ijk})$ the set of individuals known to be alive just prior to time yp_{ijk}. These individuals are pointed as k_e with $k_e\in R(yp_{ijk})$

4.5. Accounting for linkage desequilibrium in the parental generation

4.5.1. Association analysis

In the basic model, it was assumed that all loci (marker loci as well as QTLs) were in linkage equilibrium in the parents: the allele carried by a chromosome at a locus is independant of the allele the same chromosome possesses at any other locus. However, this hypothesis is not sustainable at very short distances: it is now well known that, due to various reasons (mutation, migration, selection, drift...), observation of alleles carried at two close loci are not independant. This was clearly demonstrated for marker loci (*e.g.* Farnir *et al*, 2000, in the Bovine species), and is certainly true between QTLs and very close marker loci.This disequilibrium between allele frequencies justifies so called Association or Linkage Desequilibrium Analyses (LDA). In their simplest form, these methods consider the population as a set of unrelated individuals and test the direct effect of genetic information (may be allelic, genotypic or haplotypic effect) on the quantitative trait variability.

QTLMap being dedicated to experimental populations, characterized by a family structure, the "LDA Decay" approach described by Legarra and Fernando (2009) was implemented. In this approach, parental haplotypes are pooled in classes, the classification being open to the user decision. Here, only the most probable sire and dam phases ($\widehat{\boldsymbol{hs_i}},\widehat{\boldsymbol{hd_{ij}}}$) were considered, and the classes (following the example given by Legarra and Fernando, 2009) were simply defined

by the haplotype IBS status (to a class δ corresponds a single haplotype).
To a given class δ corresponds a specific effect γ_δ on the quantitative trait.The quantitative performance of a progeny depends on the haplotypes as found in the parental chromosomes from which the putative QTL alleles are originating and not to the (possibly recombinated) haplotypes the progeny itself is carrying.Thus, the trait expectation is given by:

$$E\left[yp_{ijk}\,/\widehat{\boldsymbol{hs}}_{\boldsymbol{i}},\widehat{\boldsymbol{hd}}_{\boldsymbol{ij}}\right]=\ \mu_i+\mu_{ij}\ +\sum_{t_s,t_d}p(\boldsymbol{t}_{ijk}^{x}=(t_s,t_d)/\widehat{\boldsymbol{hs}}_{\boldsymbol{i}},\widehat{\boldsymbol{hd}}_{\boldsymbol{ij}},\boldsymbol{M}_i)\,.\,[\gamma_{\delta(\widehat{hs}_{it_s})}+\ \gamma_{\delta\left(\widehat{hs}_{ijt_d}\right)}]$$

Where $\gamma_{\delta(\widehat{hs}_{it_s})}$ is the effect of the class of t_s^{th} haplotype (1 or 2) the sire i, knowing its most probable phase $\widehat{\boldsymbol{hs}}_{\boldsymbol{i}}=\left\{\widehat{hs}_{i1},\widehat{hs}_{i2}\right\}$.

4.5.2. Linkage Desequilibrium Linkage Analysis

Association analyses suffer a lack of robustness to hidden structures.Familial structures may be accounted for adding to the model description a random individual effect with a covariance matrix computed from pedigree or dense marker information (see Teyssedre *et al*, 2011 for a review). More generally, modelling both the association (linkage desequilibrium) and transmission (linkage) in a single "Linkage Desequilibrium Linkage Analysis" as been recommanded as a solution for a better control of first type errors. Family Based Association Tests (Abecassis *et al*, 2000) and Mixed models including a random QTL effect (Meuwissen and Goddard, 2000) are generally used.

QTLMap being dedicated to experimental populations, characterized by a family structure, the "LDLA" approach described by Legarra and Fernando (2009) was implemented. This approach combines the LD Decay(§5.1) and regression (§2.1)models, the QTL effect being defined within the parental haplotype effect.The performance expectation, $E\left[yp_{ijk}\,/\widehat{\boldsymbol{hs}}_{\boldsymbol{i}},\widehat{\boldsymbol{hd}}_{\boldsymbol{ij}}\right]$ becomes

$$\mu_i+\mu_{ij}\ +\sum_{t_s,t_d}p(\boldsymbol{t}_{ijk}^{x}=(t_s,t_d)/\widehat{\boldsymbol{hs}}_{\boldsymbol{i}},\widehat{\boldsymbol{hd}}_{\boldsymbol{ij}},\boldsymbol{M}_i)\,.\,[\gamma_{\delta(\widehat{hs}_{it_s})}+\ \gamma_{\delta\left(\widehat{hs}_{ijt_d}\right)}+\alpha_{it_s}+\alpha_{ijt_d}]$$

Thus some flexibility between families around the mean haplotype effect is given.

4.5.3. The half-sib case

When the dams have only one progeny, or a very small offspring, (1) dam QTL effects cannot be correctly estimated (see § 1.3) and (2) dam phases cannot be inferred from the available marker information. However, in these situations the number of dams is large and a lot of information about QTL segregation can be extracted, thanks to the linkage desequilibrium. In dense map situations, local haplotypes (defined by segments comprising a limited number of marker loci) are very generally fully transmitted (without recombination) by the dams to their progenies.These dam to progeny transmitted haplotypes, easily deduced from the progeny marker genotypes and the sire transmitted haplotypes, are good approximations of genuine dam haplotypes, and considered as such in the previous model.

5. Setting up QTLMap

Pre-requisites

- ✓ The GNU compiler collection: gfortran 4.6, gcc
- ✓ Cmake 2.8, cross-platform, open-source build system.

Compilation

```
>cd ${QTLMAP_DIR}
>mkdir build
>cd build
>cmake -DCMAKE_BUILD_TYPE=Release ..
>cmake -DCMAKE_Fortran_COMPILER=gfortran ..
>make
```

The binary qtlmap is created in the ${QTLMAP_DIR}/build/src directory.

To install the qtlmap binary in the bin directory ${QTLMAP_DIR}/bin:

```
>make install
```

OpenMP support

Supports multi-platform shared-memory parallel programming.

To define the number of threads:

```
>export OMP_NUM_THREADS=8
```

NVIDIA GPU acceleration support

QTLMap is the ability to use NVIDIA GPU cards (Tesla C20XX series) to massively accelerate analyses and simulations for QTL detection.

```
>cmake -DCUDA_TOOLKIT_ROOT_DIR=/path-cuda-toolkit-dir -
DGENCODE_CUDA="arch=compute_20,code=sm_20" ..
```

6. Input files

To carry out an analysis, you need a minimum of 4 data files (Marker map file, Pedigree file, Marker genotypes file, Performance file), a file describing the performances (Model file) and a file describing the input, ouput and options (**parameter file**). A file describing the breed origins of parents or grand parents have to be provided when within breed haplotype effects are considered.

6.1. Pedigree file

The file contains pedigree information for the 2 last generations of a design which comprises 3 generations, *i.e.* parents and progeny. It must not contain the grand parental pedigree information.

Each line is made of an alphanumeric ID triplet (individual, sire, dam). A fourth information gives the generation number: « 1 » for the parental generation; « 2 » for the progeny generation. An animal missing one or both parents ID has not to be included in the file. The missing value code (given in the parameterization of the analyses, see 6.2) cannot be used in the pedigree file. When an animal is parent in a sire family and offspring in another one, it has to be duplicated in the pedigree file. One line with generation=1 and another with generation=2.

```
922961 911287 902206 1
944547 924758 911714 1
944985 922961 915321 1
944985 922961 915321 2
961924 922961 944547 2
961925 922961 944547 2
961926 922961 944547 2
963187 922961 944985 2
963188 922961 944985 2
963189 922961 944985 2
```

Box 1: Example of a pedigree file

In this example, the pedigree includes 7 progeny born from 1 sire and 3 dams. Sire 922961 is the son of sire 911287 and dam 902206 etc. The id 944985 is dam and offspring then it is duplicated with generation number 1 and 2.

6.2. Population file (optional)

This file gives the population category of parents or grand parents (breed, strain...).

- ✓ First column: Parents ID
- ✓ Second column: name of the population category

This information is used to determine different origins of parental haplotypes. Identical haplotypes coming from different origins will be considered different in the LD QTL analysis.

```
911287 lac
902206 rom
924758 lac
911714 rom
922961 lac
915321 rom
```

Box 2: Example of a population file

6.3. Marker map file

This file gives the locations of the markers on the chromosome(s). Each line corresponds to a single marker, and gives (order to be followed):

- ✓ marker name (alphanumerique) ;
- ✓ name of the chromosome carrying the marker (alphanumerique) ;
- ✓ marker position of the marker on the average map (in Morgan) ;

- ✓ marker position of the marker on the male map (in Morgan) ;
- ✓ marker position of the marker on the female map (in Morgan) ;
- ✓ inclusion key (=1 if the marker has to be included in the analysis, 0 if not)

```
SW552   1  0.08 0.05 0.09 1
SW64    1  0.24 0.24 0.25 0
CGA     1  0.49 0.45 0.55 1
snp1   15  0.50 0.37 0.59 1
snp2   15  0.58 0.49 0.63 1
```

Box 3: Example of a marker map file

In this example, marker SW552 is on chromosome 1, at position 0.08 on the average map, 0.05 on male map and 0.09 on the female map, and will be included in the analysis of chromosome 1, etc.

6.4. Marker genotypes file

This file contains the animals phenotypes at the markers. The first line gives the marker names, the markers must belong to the marker map file. For each animal, a line gives its ID (as described in the pedigree file) followed by the markers phenotypes, ranked following the first line order. Each phenotype is made of 2 alleles, unordered. When an animal has no phenotype for a marker, both alleles must be given the missing value code as given in the parametrisation of the analysis (see 6.2).

```
SW552 SW64 CGA snp1 snp2
911714 2 5 3 1 4 13 A T G A
912892 8 2 6 5 4 13 A T G A
924758 2 5 6 1 12 5 A T G A
922961 2 2 3 1 12 13 A T G A
944547 2 5 1 3 12 4 A A G A
944985 2 8 1 5 12 4 T T G G
961924 2 5 0 0 13 4 A A G A
961925 * * 0 0 13 4 A T A G
961926 2 5 0 0 0 0 A A G A
963187 2 8 0 0 12 4 T T G G
963188 2 2 3 1 13 4 A T A G
963189 2 2 1 1 12 4 * * G A
963190 2 8 1 5 12 4 T T A G
```

Box 4: Example of a marker genotypes file

In this example, amongst the 5 grand parents, 3 were genotyped (911714, 912892 et 924758). For instance, grand dam 911714 is heterozygous « 2 5 » at marker SW552, the individual 961925 has no genotype at marker mark1 …etc.

6.5. Performance file

This file gives the phenotypes of the traits to be analysed. The progeny performances only are considered in the analysis and must be given in the file.

For each animal, its ID (identical to the ID given in the pedigree file) is followed by information about nuisance effects (fixed effect levels, covariate value) and then by three items for each trait: performance, CD and IC . In grand daughter designs, CD is the square of the EBV accuracy. In daughter designs, CD indicates if (CD=1) or not (CD=0) the trait was measured for

this animal and must be included in the analysis. 0/1 variable (IC) which indicates if (IC=0) it was censored or not (IC=1), this IC information being needed for survival analysis (by default IC=1).

```
944985 2 10,3    5,5   1 1   75,2  1 1
961924 1 10.43   7.8   1 1   77.6  1 1
961925 2 5.34    0.0   0 1   90.   1 1
961926 1 12.34   11.3  1 1   103.  1 1
963187 2 9.45    12.7  1 1   98.   1 1
963188 1 11.10   13.5  1 1   0.0   0 1
963189 2 10.11   10.   1 1   94.8  1 1
```

Box 5: Example of a quantitative trait values file

This file describes 2 traits. For progeny 961924, the recorded information are: sexe 1 (fixed effect), body weight 10.43 (covariate), backfat thickness 7.8mm (trait 1) and fatening period of 77.6 days (trait 2) etc.

Special case: performance file for expression quantitative traits:

When performances are expression data, another format is required. This file gives the phenotypes expression traits to be analysed. The header line is the list of animals phenotyped. The following lines are the fixed effects, covariates and finally the phenotype.

The format of the nuisances effects and phenotype line is:

<IDANIMAL> <VALUE_ANIMAL1><VALUE_ANIMAL2>...

For missing data, insert a character string which is not interpretable as a numeric (*e.g.* n/a).

```
944985 961924 961925 961926 963187 963188 963189 963190
sexe 1 1 1 1 1 1 1 1
cov1 0.3 0.4 0.3 0.5 0.5 0.6 0.3 0.2
gen1 0.0184170490684831 -0.143560443113406 -0.118137020630747 -0.06666521254513
0.0642879011796014 -0.255460347400393 -0.189477060869665 -0.25462868498086
gen2 -0.127806826817031 -0.163876647400758 0.0184043832497863 -0.296146098377366 -
0.112715209230912 -0.0684375510992924 -0.180990247175303 -0.182892021501701
gen3 -0.259405679027549 -0.365184085691961         n/a       -0.104403755609133 -
0.154653751085067 -0.213511162284327 -0.190633612968503 -0.344837877148359
gen4 0.151093991655429 0.10964888434473 0.15832262904679 0.284848089326391
0.0808434990010986 0.306550168430082 0.00906573426897184 0.10731093171816
```

Box 6: Example of a expression quantitative trait values file

In this example, animal 6380 have a missing data for the gen3.

6.6. Model file

In this file, the information on model for analysis of each trait is described:

- ✓ Number of traits
- ✓ Number of fixed effects (nf), Number of covariates (nc)
- ✓ Names of the fixed effects and covariates
- ✓ Name of the 1st trait, nature of trait ('r' for real value or 'i' discrete ordered data) model for this trait symbolized by 0/1 indicators for each fixed effect (nf first indicators), each covariate (nc following) and each interaction between the QTL and the fixed effects (nf last

indicators). Fixed effect, covariate or interaction will be included in the analysis if its indicator is 1, will not be if it is 0.

✓ Name of the 2nd trait,...

✓ For simulations or grand daughter design: the <keyword> "CORRELATION_MATRIX" should be used. Following the key word, the correlation matrix is given: the heritability as diagonal elements, below the phenotypic correlations and above the genetic correlations (not available for expression traits) . If this information is missing, h^2=0.5 and correlations=0.5 are assumed.

Options for trait information:

✓ To give the same model for all traits use the <keyword> "ALL" instead of one model for each trait (only for expression data).

✓ "TRAITS" <keyword> is an option to select analysed traits: a list of traits to be kept in the analysis.

Example 1: standard situation.

```
3              ! Number of traits
1 1            ! Number of  fixed effects and covariates
sexe poids     ! Names of the fixed effects and covariates
malade r 1 1 0 ! 1st trait, (nature: real value) model
malcor r 0 0 1 ! 2nd trait,(nature: real value) model
third  r 0 0 0 ! 3nd trait,(nature: real value) model

CORRELATION_MATRIX
0.35 0.28 0.29
0.20 0.32 0.28
0.20 0.20 0.33
```

Box 7: Example 1 of a model file

This model file describes the performance file where one fixed effect, one covariate and three performances are referenced for each animals.

The model for each performance is:

malade = μ + sexe + β . poids + ε
malcor =μ + QTL × sexe + ε
third =μ + ε

Example 2: Use of <keyword> ALL, in particular in expression data analyses.

```
 10000          ! Number of traits
1 1             ! Number of  fixed effects and covariates
sexe cov1       ! Names of the fixed effects and covariates
ALL r 1 1 0     ! all is a word key: the model will be applied for all
                ! the 10000 expression trait
```

Box 8: Example 2 of a model file

Example 3: Use of <keyword> TRAITS.

Only traits “third” and “malcor” will be analysed

```
3              ! Number of traits
1 1            ! Number of  fixed effects and covariates
sexe poids     ! Names of the fixed effects and covariates
malade r 1 1 0 ! 1st trait, (nature: real value) model
malcor r 0 0 1 ! 2nd trait,(nature: real value) model
third  r 0 0 0 ! 3nd trait,(nature: real value) model

CORRELATION_MATRIX
0.35 0.28 0.29
0.20 0.32 0.28
0.20 0.20 0.33

TRAITS
third malcor
```

Box 9: Example 3 of a model file

Example 4: Use of key words TRAITS and ALL, to select a list of expression traits to be analysed. Here only genes, named gen3 and gen4, will be analysed.

```
10000          ! Number of traits
1 1            ! Number of  fixed effects and covariates
sexe cov1      ! Names of the fixed effects and covariates
ALL r 1 1 0
TRAITS
gen3 gen4
```

Box 10: Example 4 of a model file

6.7. Parameter file

All information needed to run an analysis is given in the **parameter file** *p_analyse*:

- ✓ Files paths and names
  - o input files:
    - pedigree (*cf.* 5.1)
    - population origin of prents or grand parentes (*cf.* 5.2)
    - markers map (*cf.* 5.3)
    - markers genotypes (*cf.* 5.4)
    - trait performances (*cf.* 5.5)
  - o file giving the performances model (*cf.* 5.6)
  - o ouput files:
    - full information analysis result file
    - summary of the analysis
    - sire and dam family **likelihood ratio test (LRT)** along the linkage group
    - sire and dam **QTL effect estimations** along the linkage group (under hypothesis H1 = 1 QTL and H2 = 2 QTL)
    - grand parental **segment transmission** marginal and joint probabilities
- ✓ compulsory parameters:

- explored chromosomes id
- step length of the scan
- minimum size of a full sib above which the dam effects (QTL and polygenic) are estimated
- minimal probability for a paternal and maternal phase to be considered in the analysis
- missing genotype value

The **parameter file** use the format <key>=<value>. None of the characters after the character '#' are interpreted (useful to add comments).

```
#qtlmap --help-panalyse: for more information
##### USER FILES
in_map=carte
in_genealogy=genea
in_genotype=typage
in_traits=perf
in_model=model

##### ANALYSIS PARAMETERS
# analysis step:  in Morgan
opt_step = 0.1
# minimal number of progeny by dams
opt_ndmin=20
#Minimal paternal phase probability
opt_minsirephaseproba=0.80
# overload:
opt_minsirephaseproba=0.90
#Minimal maternal phase probability
opt_mindamphaseproba=0.10
# chromosome to analyse
opt_chromosome=7
#for several chromosomes
#opt_chromosome=7,8,Y
#missing phenotype marker value
opt_unknown_char=0
###### OUTPUT
out_output=./OUTPUT/result
out_summary=./OUTPUT/summary
out_lrtsires=./OUTPUT/sires
out_lrtdams=./OUTPUT/dams
out_pded=./OUTPUT/pded
out_pdedjoin=./OUTPUT/pdedjoin
out_pateff=./OUTPUT/pateff
out_mateff=./OUTPUT/mateff
out_phases=./OUTPUT/phases
out_haplotypes=./OUTPUT/haplotypes
```

Box 11: Example of a parameter file

Several keys may be defined (compulsory keys in grey):

Key	Description	Default value
INPUT FILES KEYS		
in_map	Input map	
in_genealogy	Input genealogy	
in_genotype	Input genotype	
in_traits	Input trait	
in_model	Input model describing the performances and model factors	
in_paramsimul	Input simulation parameters	
in_pop	Optional input to give population names	
OUTPUT FILES KEYS		
out_output	Full information about the results	
out_summary	Short information about the results	
out_lrtsires	Sire family likelihood ratio test	
out_lrtdams	Dam family likelihood ratio test	
out_pded	Grand parental segment transmission marginal probabilities	
out_pdedjoin	Grand parental segment transmission joint probabilities	
out_phases	Parental phases information	
out_freqall	Allele frequency of markers retained in the analysis	
out_phases_offspring	Offspring haplotypes with the parental origin: the entire chromosome is considered without option about begining and end of region	
out_haplotypes	Haplotypes	
out_pateff	Sire QTL effect estimations under H1	
out_mateff	Dam QTL effect estimations under H1	
out_maxlrt	Simulation report (Position and max LRT)	
out_grid2qtl	Sire QTL effect estimations under Hypothesis H2	
out_coeffda	Trait weights of linear combinations at each tested chromosomal location (multivariate analyses)	
out_informativity	Informativity at each tested chromosomal locations	
GENERAL OPTIONAL KEYS		
opt_step	Step length of the genome scan (Morgan). When opt_step=0 analysis is done at each marker position	0.05

opt_ndmin	Minimum number of progeny per dam: offspring size above which the polygenic and QTL effects of the dam are estimated	10000
opt_minsirephaseproba	Minimal paternal phase probability: the analysis is interrupted if for a sire, none of its phases reaches this threshold	0.90
opt_mindamphaseproba	Minimal maternal phase probability: threshold above which the probable maternal phases will be considered in the analysis	0.10
opt_unknown_char	Unknown genotype value	'0'
opt_chromosome	Linkage group (most often chromosome name)	
opt_phases_offspring_marker_start	Name of the marker at the begining of the offspring haplotypes (option of out_phases_offspring)	
opt_phases_offspring_marker_end	Name of the marker at the end of the offspring haplotypes (option of out_phases_offspring)	
OPTIONAL KEYS FOR advanced users		
opt_eps_cholesky	coeff cholesky decomposition	0.5
opt_eps_confusion	Threshold to test between factors confusion from the contingence matrix	0.70
opt_eps_hwe	Threshold to check the equilibrium of marker transmission within each family	0.001
opt_eps_linear_heteroscedastic	Threshold for convergence in the heteroscedastic linear model	0.5
opt_max_iteration_linear_heteroscedastic	Maximum iteration in the heteroscedastic linear model to avoid infinity loop	5
opt_eps_recomb	Minimum probability of recombination events knowing the recombination rate between 2 markers to detect mapping errors	0.05
opt_nb_haplo_prior	Maximum number of haplotypes at a given position above which the runtime execution is stopped (computing ressource may become problematic)	200
opt_pro_haplo_min	Minimum frequency under which an haplotype is added to the rare haplotypes group	0.00001
opt_longhap	Haplotype length in LD and LDLA (number of markers)	4
opt_optim_maxeval	Maximum number of objective function estimations	1000000
opt_optim_maxtime	Maximum time to optimize the objective function	1000000
opt_optim_tolx	Finite difference variables values used in estimating function gradient (non linear methods)	0.00005
opt_optim_tolf	Stopping criteria lower bound of the objective function	0.00005
opt_optim_tolg	Stopping criteria lower bound of the gradient	0.00005
opt_optim_h_precision	Precision to obtain the gradient	0.00005

Remark1: opt_ndmin:

The maximum likelihood methods implemented in QTLMap considers the population as being a mixture of half sib and full sib families. The sires and the dams are supposed unrelated. A sire (*resp.* a dam) may be mated to more than one dam (*resp.* sire). Thus, two animals of the second generation may be unrelated, half sibs or full sibs. A polygenic and a QTL effect are estimated for each parent having a large enough family. To avoid numerical difficulties, these effects are not estimated for dams having too small offspring. In this case, the dam progeny are considered as sire half sibs only. A control of the structure is allowed through the option number of progeny *opt_ndmin* which is given in the **parameter file**.

Remark 2: opt_mindamphaseproba and opt_minsirephaseproba

In the current release QTLMap considers only one phase for the sire, excepted when the probabilities of all possible sire and dam phases are computed with the running option –haplotype=1,2,3 (see below). If none of those probabilities for the sire exceed a given threshold (opt_minsirephaseproba in the **parameter file**) the process is aborted.

Remark3: optimisation options

Optimisation methods can be fine tuned by expert users changing, from their default values, the keys *opt_optim_maxeval, opt_optim_maxtime ,opt_optim_tolx, opt_optim_tolf , opt_optim_tolg , opt_optim_h_precision* (see in point 6.5 –optim=).

7. Run the software with the different running options for analyses

```
>${QTLMAP_PATH}/qtlmap parameter_file
<--calcul=  ,
 --haplotype=  ,
 --optim=  , -
 --qtl=  ,
 --snp,
 --data-transcriptomic,
 --print-all,
 --permut,
 --nsim =  ,
 -- disable-sire-qtl
 -- ci =
 -- ci-nsim = >
```

7.1. Option `--calcul=`: choice of the QTL analyses

Option `--calcul` allows to perform LA, LD or LDLA analysis using a Gaussian distribution for one trait. For all these analyses, the variance within sire families can be considered identical or heterogeneous between families (homoscedastic/heteroscedastic). For LA only, additional models are available: joint analysis of several traits either considering a multivariate gaussian distribution or using a discriminant analysis approach ; censored analysis using a cox model.

Available options for `--calcul`:

Option	Type of Analysis	# traits	# QTL	residual variance	Trait distribution	within dam likelihood	nuisance effect	prohibited running options	specific key in **parameter file**
1	LA	unitrait	1 or 2 QTL	Within sire	gaussian	mixture analysis	No	--hdam,--disable-sire-qtl	
2	LA	unitrait	1 QTL	Within sire	gaussian, discrete	mixture analysis	possible		
3	LA	unitrait	1 ,n QTL	general	gaussian	linear approximation	possible		
4	LA	unitrait	1,n QTL	Within sire	gaussian	linear approximation	possible		
5	LA	multitrait (multivariate analysis)	1 QTL	Within sire	gaussian	mixture analysis	no		
6	LA	multitrait (discriminant analysis)	1 QTL	Within sire	gaussian	mixture analysis	no		
7	LA	unitrait	1 QTL	no variance	Cox model	approximation	possible	--qtl=,--print-all --data-transcriptomic -- disable-sire-qtl	
21	LA	unitrait	1,n QTL	Within sire	gaussian, , discrete	mixture analysis	possible		
23**	LA	multitrait (multivariate analysis)	1 ,n QTL	Within sire	gaussian	mixture analysis	possible		
24**	LA	multitrait (multivariate analysis)	1 ,n QTL	Within sire	gaussian	mixture analysis	possible		
25	LD	unitrait	1 ,n QTL	general	gaussian	linear approximation	possible		in_race
26	LD	unitrait	1 ,n QTL	Within sire	gaussian	linear approximation	possible		in_race
27	LDLA	unitrait	1 ,n QTL	general	gaussian	linear approximation	possible		in_race
28	LDLA	unitrait	1 ,n QTL	Within sire	gaussian	linear approximation	possible		in-race

Options for `--calcul` (in development for advanced users only):

8	LD	unitrait	1 QTL	Within sire	gaussian	mixture analysis	possible		
9	LDLA	unitrait	1 QTL	Within sire	gaussian	mixture analysis	possible		

* n is the number of QTL: recommended value<=3. **23 and 24 approaches are similar but 24 includes a LU factorization of traits correlation matrix

7.2. Option `--haplotype=`: parental phase identification

To chose the parental phases identification and the grand parental segment transmission methods. The methods are based on various algorithms, with different balance between computation speed and precision.

--haplotype=	Sire and Dam phase probability	Transmission probabilities from parents to offspring	Recomandation for sparse/dense map
0	Phases are read in the markers genotype file: for each locus 1st (resp. 2nd)allele read on the 1st chromosome (resp. 2nd)	Rapid and optimised transmission probabilities (Elsen et al, 2009)	Sparse or dense
1	Exact probability of phases by enumeration All possible phases are considered in turn and their probability computed	Exact transmission probabilities are computed using all available information	Sparse
2 (not recommended)	Exact probability of phases by enumeration All possible phases are considered in turn and their probability computed	Approximate transmission probabilities are computed using all available information	Sparse
3 (default)	Exact probability of phases by enumeration All possible phases are considered in turn and their probability computed	Rapid and optimised transmission probabilities (Elsen et al, 2009)	sparse
4	Very fast but approximate identification of the most probable phases based on closest marker information (Windig and Meuwissen)	Rapid and optimised transmission probabilities (Elsen et al, 2009)	dense
5 (eq --snp)	Fast and almost exact identification of the most probable phases based on closest marker information (Favier et al, 2010)	Rapid and optimised transmission probabilities (Elsen et al, 2009)	dense

7.3. Option `--snp`: fast phasing in dense genotyping situations

This option allows to determine phases rapidly and is a good option for dense markers maps. In some cases, convergence may be diffcult if not impossible. This situation may happen due to genotyping errors.

This option --snp is equivalent to --haplotype=5.

Example:

```
>${QTLMAP_PATH}/qtlmap parameter_file –calcul=1 --snp
```

7.4. Option `--qtl=`: number of qtl detection available

For most of the analyses (controlled by the runtime option --calcul), only 1 QTL is considered in the model. However, this number may be increased to 2 or more depending of the `--calcul` option. The number of QTL is given by the --qtl runtime option. Practically, as computing time increases rapidly with the number of QTL, we do not recommend testing for more than 3 qtl.

Analysis `--calcul`	QTL test detection `--qtl`
1	1,2
2,5,6,7,8,9	1
3,4,23,25,26,27,28	>=1

Example:

```
>${QTLMAP_PATH}/qtlmap p_analysis-calcul=1 --qtl=1
```

7.5. Option `--optim=`: Optimisation method

The `–optim` runtime option allows a control of the optimisation procedure. Many methods are proposed:

--optim=	Description	External packages needed
1	E04JYF NAG routine - quasi-Newton	NAGG
2	L-BFGS routine - **the Broyden–Fletcher–Goldfarb–Shanno** quasi-Newton	no
5,...,11	LUKSAN optimisation	no
12,...,47	NLOPT Optimisation	GCC

Methods may be parametrized with the following options:

- ✓ opt_optim_maxeval: maximum number of objective function
- ✓ opt_optim_maxtime: maximum time to find the solution of the objective function
- ✓ opt_optim_tolx: tolerance lower bound of a step
- ✓ opt_optim_tolf: stopping criteria lower bound of the objective function
- ✓ opt_optim_tolg: stopping criteria lower bound of the gradient
- ✓ opt_optim_h_precision: precision to obtain the gradient

7.6. Option `--disable-sire-qtl`

This option allows to estimate the dam QTL effect only (needed when exploring the X chromosome in mammals).

7.7. Options –-ci & --ci-nsim=

To obtain confidence interval of QTL position, four methods are available, informed by the –ci runtime option.

- ✓ 1 : Drop-off method
- ✓ 2 : Boostrap resampling method
- ✓ 3 : Boostrap resampling method (keep exactly the number of progeny within a family)
- ✓ 4 : Hengde Li method using the Relative Frequency Ratio

The number of simulation or resampling for Confidence Intervals/boostrap methods is given by the runtime option --ci-nsim=, the default value being 1000.

Example:

```
>${QTLMAP_PATH}/qtlmap parameter_file –-calcul=1 --ci=3,4 –-ci-nsim=500
```

7.8. Options --data-transcriptomic & --print-allReport output mode: eQTL analyses (to analysis transcriptomic data)

When looking for eQTL the number of traits to be analysed becomes very large. In this case, specific routines are needed, and *ad hoc* output are produced. To get this situation, the runtime option data-transcriptomic must be indicated.

When performing eQTL analyses (using –data-transcriptomic command) or corresponding eQTL simulations the output is minimised. To force the classical reporting format, use the runtime option –-print-all.

Example:

```
>${QTLMAP_PATH}/qtlmap p_analysis–calcul=1 –data-transcriptomic --print-all
```

7.9. Options for the control of process information

To get the maximum information during the process, add –v (or --verbose) to the command

```
>${QTLMAP_PATH}/qtlmap p_analysis–calcul=1 -v
```

When debuging the software , add –d (or --debug) to the command

```
>${QTLMAP_PATH}/qtlmap p_analysis–calcul=1 -d
```

To avoid outpout, add –q (or --quiet) to the command

```
>${QTLMAP_PATH}/qtlmap p_analysis–calcul=1 –q
```

8. Control of first and second type errors in existing designs

Simulations or permutations can be organised to empirically estimate the rejection thresholds of the test statistic and to measure the detection power of an existing design.

To perform these estimations <Keywords> must be defined in the **parameter file**:

The **simulation parameters file** name is given in the parameter analysis file with the key `in_paramsimul` (not for permutations).

A second key (optional) `out_maxlrt` specifies the name of a file reporting the maximum likelihood ratio test values for the simulations or permutations.

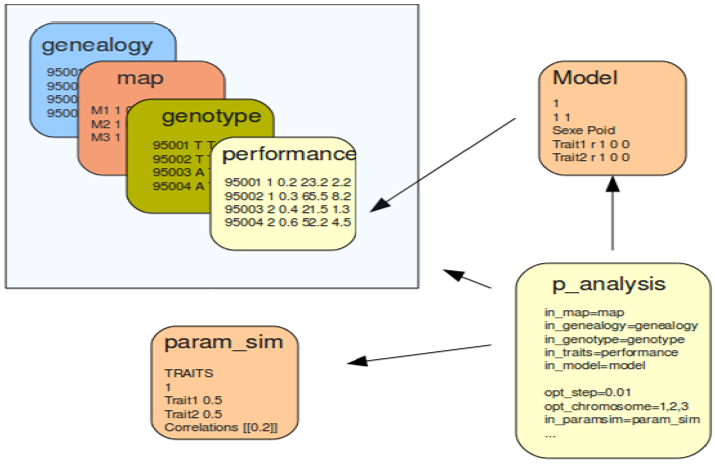
Sections 7.1 and 7.2 describe how to compute empirical distributions of test statistics while accounting for the missing data structure on phenotypes existing in the data set. Section 7.3 proposes an alternative to simulate data for all progeny independantly to recorded traits.

8.1. Simulations with respect of missing data structure

When no QTL is simulated (null hypothesis "No QTL on the linkage group") all needed information (heritabiilty and correlations) is provided in the model file.

Under H1 or H2, a specific file, **simulation parameter file** , defined by *`in_paramsimul`* in the **parameter file***(*example *`param_sim)`* must be provided by the user. This file contains the information needed for the simulation:

- ✓ QTL information (keyword `QTL`)
- ✓ Trait information (keyword `TRAITS`)



When N QTL are simulated, N≠0, (rejection thresholds for the test of H_N "N QTL" *vs.* H_{N+q} "N+q QTLs" segregating, or analyses for power of detection), the QTL is supposed to be biallelic with alleles Q1,Q2. f1 is the frequency of the first allele in the grand sire population, simulated as being equal to the frequency of the second allele in the grand dam population. As a result, the expected genotype frequencies in the parental population are Q1Q1: f1.(1-f1), Q1Q2: f1.f1+(1-

f1).(1-f1), Q2Q2: (1-f1).f1. To get for instance all parents heterozygous, the frequency f1 must be given the value 1. or 0.

The specific "`QTL`" <keyword> on the first line is mandatory to simulate QTL effects. Next, the number of QTLs to be simulated is given.

The user defines for the QTL with a format <keyword> <value>:

- ✓ After <keyword> `Position`, positions of the chromosme (in Morgan unit)
- ✓ After <keyword> `chromosome`, chromosome where they are located
- ✓ After <keyword> `frequency`, frequency of one of the QTL allele in grand sire population

The specific "`TRAITS`" <keyword> starting the second section is mandatory to simulate phenotypes. On the next line, the number of traits to be simulated is given.

Then, for continuously distributed traits: the name of each trait to simulate (as referenced in the model file), with one line per trait.

For discrete traits: the name of each discrete trait to simulate (as referenced in the model file), with one line per trait, followed on the same line by:

- ✓ trait heritability
- ✓ number of modalities
- ✓ frequency of each modality

Only if one or more QTL is to simulate, after the <keyword> `qtleffect`: the QTL effects (real values) are listed: trait 1-QTL1, trait1-QTL2 .... trait2-QTL1, trait2-QTL2....

Example 1: a **parameter file** for the estimation of the rejection threshold of the test « There is one qtl on the linkage group» *versus* « there is no QTL », with the corresponding model file.

```
TRAITS    ! <keyword>
2         ! number of traits to simulate
imf       ! name of the first trait as given in the model file (see Box 12.2)
bardiere  ! name of the second trait as given in the model file (see Box 12.2)
```

Box 12.1: Example of simulation parameter file with no QTL effect

```
2
0 0
nofix nocov
imf r 0 0 0
bardiere r 0 0 0
```

Box 12.2: Corresponding example of Model file

Example 2: a **parameter file** for the estimation of the rejection thresholds for the test « There are two QTL on the linkage group» *versus* « there is one QTL at the position 0.6 Morgan on the first chromosome ».

In this example, the QTL simulated has an effect of 0.4 on the first trait and 0.5 on the second trait. The QTL alleles are fixed in the grand parental populations.

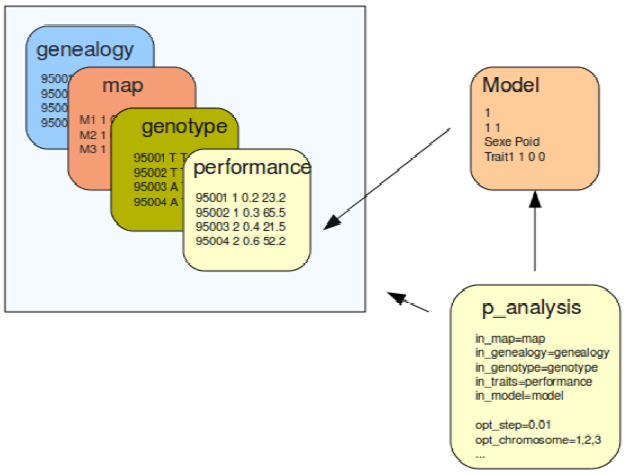
```
QTL                 ! <keyword>
1                   ! number of QTL to simulate
Position 0.6        ! <keyword> position of each QTL in Morgan
chromosome 1        ! <keyword> chromosome location of each QTL
frequency 1.0       ! <keyword> frequency f1 of one QTL allele in a grand parental
population

TRAITS              ! <keyword>
2                   ! number of traits to simulate
imf                 ! name of the first trait as given in the model file (see Box 12.2)
bardiere            ! name of the second trait as given in the model file (see Box 12.2)

qtleffect 0.4 0.5   ! <keyword> QTL effect on trait 1 and 2
```

Box 13: Example of a simulation parameter file to simulate data with a QTL effect affecting traits referenced in the model file

8.2. Permutations



In the QTLMap software, the permutation option allows to permute the nuisance effects and phenotypes between genotyped animals within full and/or half sib families to empirically estimate the distribution of test statistic under the null hypothesis "no QTL is segregating on the linkage group". The permutation procedure, proposed by Churchill and Doerge (1994) is an intuitive method for estimating thresholds which accurately reflects the specificities of an experimental situation. However, when the permutation groups are small, the number of permutation possibilities decrease and the simulation method is more adapted to estimate the distribution of the test statistic under H0. In order to prevent unsuited calculations, an arbitrary threshold for family sizes was fixed to 10 to allow permutations. Different permutation situations were considered:

– When the full sib family size is higher than the `nd_min` key (or 10 if `nd_min` <10), genotyped animals are permuted within the sire full sib family

– When the full sib family is smaller than `nd_min` (or 10 if `nd_min`<10), the permutation is performed within half sib family.

– When a half sib family is smaller than 10, no permutation is performed and an error message is printed.

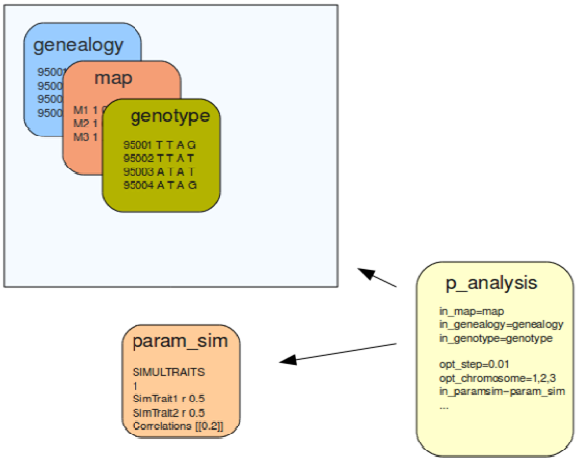
In case of a multitrait analysis, only phenotyped animals are permuted. In case of a unitrait

analysis, animals with at least one phenotyped trait among the traits of the performance file are permuted.

Permutations of performances is available with the runtime option `--permute`

```
>${QTLMAP_PATH}/qtlmap p_analysis-calcul=1 –nsim=100 –permute
```

8.3. Simulations without reference to data structure



The simulations can be carried out with no reference to existing traits, which allow simulating phenotypes for all progeny without missing data. In this case, the **parameter file** does not need the <keywords> *`in_model`* and *`in_trait`*.

The **simulation parameter file** should have a specific <keyword> to start the trait section: `SIMULTRAITS`. This section is identical to the `TRAITS` section in paragraph 7.1, but additional information about the nature of the trait is provided to compensate the absence of the model file. This information is given next to the trait name:

- ✓ Trait name, « `r` » for real data, heritability of the trait
- ✓ Trait name, « `i` » for integer (ordered discrete data), heritability of the trait, number of classes and frequencies of each class

If QTL are simulated, the **simulation parameter file** should start with the `QTL` section as presented in paragraph 7.1.

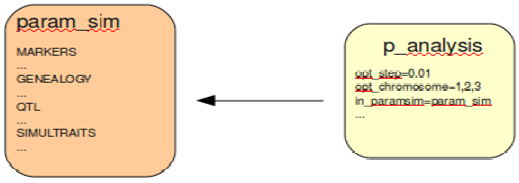
Example: a **simulation parameter file** for the estimation of the rejection thresholds for the test « There are two QTL on the linkage group» versus « there is one QTL at the position 0.6 Morgan on the 7th chromosome ». The QTL simulated has an effect of 0.5 on the first trait (normally distributed with h^2=0.5) and 0.5 on the second trait (discrete, distribtuted in 4 classes with h^2=0.50). The QTL alleles are fixed in the grand parental populations.

```
QTL
1                    ! nb QTL
position      0.6    ! position QTL1...
chromosome     7     ! chromosome ID
frequency     1.0    ! frq alleles QTL dans P1 et P2 - biallelique pour QTL1...

SIMULTRAITS
2                                   ! nb QTL
traitsimul1 r 0.50                  ! name, nature (real) and heritability of the 1st
trait
traitsimul2 i 0.50 4 0.3 0.3 0.3 0.1  ! name, nature (discrete), heritability, numbr of
classes, classes frequencies of the second trait
correlation [ [0.0] ]               ! between traits phenotypic correlation
qtleffect   0.5 0.5                 !   QTL effects
```

Box 14: Example of a simulation parameter file to simulate data with no reference to existing trait data (no missing phenotype structure)

9. Simulate and design a new protocol



QTLMap offers the possibility to simulate all the data (markers, genealogy, traits) in order to plan a new experiment. The output file (named by the `out_maxlrt` option in the following example) provides the value of the LRT resulting from the simulations, allowing an estimation of the power of the design.

To perform these simulations, two specific sections must be created in the **simulation parameters file** in addition to the sections `QTL` and `SIMULTRAITS` previously described:
The first, with the <keyword> `MARKERS`, must give on a single line three items: the marker density (M), the number of alleles/marker, the map size (Morgan)
The second, with the <keyword> `GENEALOGY`, followed on the next line by the <keyword> `F2`, `BC` or `OUTBRED` depending on the type of population to simulate, and a line giving three items: the number of sires, number of dam/sire and number of progeny/dam to simulate.

Example : simulation of an F2 protocol with 10 sires, 7 dams per sire and 12 progeny per dam (total 840 progeny), with one chromosome and 101 SNP evenly distributed on 100cM, a QTL located in position 70.5cM. The QTL is not fixed in the grand parental population. Two real traits are simulated with correlation -0.4.

```
MARKERS
0.01 2 1              ! marker density (M), number of alleles/marker, map size (Morgan)

GENEALOGY
F2                    ! type of design
10 7 12               ! number of sires, number of dam/sire and number of progeny/dam
```

```
QTL
1
Position    0.705
chromosome  1
frequency    0.7

SIMULTRAITS
2
simtrait1 r 0.25
simtrait2 r 0.35

correlation [ [ -0.4 ] ]
qtleffect 0.1 0.5
```

Box 15: Example of a simulation parameter file to simulate a completely new protocol F2

10. Output files

A set of files is proposed as the result of an analysis or a simulation:

- ✓ The main output (analysis report, simulation report)
- ✓ A summary

Additional files (optional) in for reporting analyses:

- ✓ Likelihood ratio test profile (per sire family, per dam family, general)
- ✓ QTL effect estimation at each tested position (sire family and dam family)
- ✓ Parental phases
- ✓ Alleles frequencies
- ✓ Haplotypes assigned from parents to progeny
- ✓ Parental segment transmission marginal probabilities
- ✓ Parental segment transmission joint probabilities

Specific files (advanced users):

- ✓ Coefficients of the discriminant analysis along the linkage group

Additional file (optional) in a simulation/permutation case:

- ✓ Maximum Likelihood Ratio Test and its position for each simulation/permutation

10.1. Main output for phenotype analysis

The main output files comprises five sections.

- The **first section** describes the **data** as read by the software.
  - Description of the parameters file

The name of the corresponding file is provided by the user with the key *out_output* in the **parameter file**. The list of runtime option keys used by the application (runtime environment) is given (all keys are described at the end of this document).

```
    ***
       DATE      = 2013/06/20-17:12:02
       Release-build      = 0.9.6-C-26.04.2013-17.11.08
       ARGUMENTS   = p_analyse --calcul=2
       --CALCUL  =            2  (MODLIN ANALYSIS)
 OMP_NUM_THREADS  =            6


    *****************  PARAMETERS ANALYSE FILE SUMMARY *****************

out_output                                =./OUTPUT/result
out_lrtsires                              =./OUTPUT/sires
out_lrtdams                               =./OUTPUT/dams
out_pded                                  =./OUTPUT/pded
out_pdedjoin                              =./OUTPUT/pdedjoin
out_pateff                                =./OUTPUT/pateff
out_mateff                                =./OUTPUT/mateff
out_grid2qtl                              =./OUTPUT/grid2qtl
out_summary                               =./OUTPUT/summary
opt_chromosome                            =7
opt_ndmin                                 =20
opt_step                                  = 0.01
opt_unknown_char                          =0
opt_mindamphaseproba                      =0.10
opt_minsirephaseproba                     =0.90
opt_eps_cholesky                          =0.01
opt_eps_confusion                         =0.70
opt_eps_hwe                               =0.01
opt_eps_linear_heteroscedastic            =0.5
in_map                                    =map
in_genealogy                              =genealogy
in_traits                                 =phenotypes
in_genotype                               =genotype
in_model                                  =model
in_paramsimul                             =param_sim_real
out_maxlrt                                =./OUTPUT/all_simul
opt_max_iteration_linear_heteroscedastic=30
opt_eps_recomb                            =0.05
out_phases                                =./OUTPUT/phases
opt_nb_haplo_prior                        =200
opt_prob_haplo_min                        =0.00001
opt_long_min_ibs                          =4
opt_longhap                               =4
opt_optim_maxeval                         =1000000
opt_optim_maxtime                         =1000000
opt_optim_tolx                            =0.00005
opt_optim_tolf                            =0.00005
opt_optim_tolg                            =0.00005
opt_optim_h_precision                     =0.00005
```

Box 16.1: Example of main output file information, first section (data characteristics)

- Description of the genealogy

Number of parents, grand-parents and progeny are given.

```
        *****************  GENEALOGY DESCRIPTION *****************

 The pedigree file includes      20 parents born from      5 grand sires and       5
grand dams
 and     236 progeny born from      4 sires and      16 dams
```

Box 16.2: Example of main output file information, first section (simple statistics about genealogy)

- Description of the markers information

Charateristics of marker data read in input files are given :

- ✓ Number of genotyped individuals
- ✓ Number and names of the genetic markers, of alleles of each marker, allele frequencies
- ✓ If unbalanced allelic segregations are observed (for all markers, the deviation to 0.5 of heterozygous frequency in the offspring of heterozygous sires is tested with a Fisher test), a warning about potential transmission distorsion for the marker within the family.

```
   ***************** MARKER DESCRIPTION *****************

  236 animals are present in the genotype file
 animal890738 900848 of genotype file are not in the pedigree file
 where all animals are genotyped for at least one marker.
markers were selected among   10 markers
There are  236 genotyped animals

** Check the equilibrium of marker transmission within each family **

 Marker [m10] for sire :20 not in HWE :           92  heterozygous progeny
amongst          150
 Marker [m49] for sire :17 not in HWE :           56  heterozygous progeny
amongst          150
 Marker [m139] for sire :11 not in HWE :           58  heterozygous progeny
amongst          150
…
```

Box 16.3: Example of main output file information, first section (simple statistics about marker data)

Description of performance traits

For each quantitative traits, simple statistics are edited:

- ✓ Names of the quantitative traits, for each trait:
- ✓ Number of individuals measured
- ✓ Number of individuals having for both performance values and marker genotypes
- ✓ Mean, variance, minimum and maximum
- ✓ Names of fixed effect, if any, with the list of levels
- ✓ Names of the covariates, if any, with their mean, variance, minimum and maximum

```
     ***************** TRAITS DESCRIPTION *****************


     NUMBER OF PHENOTYPED ANIMALS   :   236
     NUMBER OF PHENOTYPED AND GENOTYPED ANIMALS :   236
     NUMBER OF TRAITS               :     3
     NUMBER OF FIXED EFFECTS        :     0
     NUMBER OF COVARIABLES          :     0

 TRAIT :Bardie
     NUMBER OF PHENOTYPED PROGENY              :   236
     MEANS                                     :    7.169+-   0.650
     MINIMUM                                   :    6.048
     MAXIMUM                                   :   9.668
     NUMBER OF MISSING PHENOTYPES              :     0
     NUMBER OF CENSORED PHENOTYPES             :    0
 WITHOUT MODEL for fixed effects and covariables
```

Box 16.4: Example of main output file information, first section (simple statistics about performance traits data)

- o The **second section** informs about preliminary steps of the process :

- Description of parental phase reconstruction

This information about parental phases is fully given in the file specified after the heading "parental phases". In this file figure the most probable phases of the sires, and of the dams if available from the analysis, built from available marker and pedigree information .

Remember that parameters to control the minimal sire and dam phase probability can be reset by the user with the keys *opt_minsirephaseproba* and *opt_mindamphaseproba* in the **parameter file**.

```
     ***************** PARENTAL PHASES *****************
                      FILE :./OUTPUT/phases
    ************************************************
```

Box 16.5: Example of main output file information, second section (parental phase information)

- Description of data structure

An header edited for each trait, with the number of QTL effects to be estimated.

```
        ***********************************
        *                                 *
        *  Analysis of trait      Bardiere *
        *                                 *
        *                                 *
        ***********************************

 LRT profile on the linkage group :
  position, test statistic  ,
     4 sire QTL effects ,
    16 dam QTL effects
```

Box 16.6: Example of main output file information, second section (quality of parameters estimations)

As the design may be unbalanced, leading to strong colinearity between QTL effects and some other effects in the model: a warning is provided if this situation occurs. The confusion is measured by the correlation between the columns of the incidence matrix in an equivalent fully linear model at the starting position of the scan (a warning is edited if this correlation exceeds *opt_eps_confusion*).
A second test of confusion between the QTL effects and the estimable effects finally kept in the model is edited.

```
Test of confusion between QTL and other effects in the initial full model
(test based on the correlation between columns of the incidence matrix)

*****************************************************************************
Confusion between QTL and other effects (final constained model)

 No confusion detected
 the highest correlation is :   0.257
```

```
*****************************************************************************
```

Box 16.7: Example of main output file information, second section (quality of parameters estimations)

- The **third section**, provides results of H0 hypothesis analyses for each trait.

Parameters estimation, with estimability information and precision indicators are listed for:

- ✓ Within sire standard deviation (global standard deviation for models 3, 25 and 27)
- ✓ Sire polygenic efects
- ✓ Dam polygenic effects (if their family size is over OPT_NDMIN)
- ✓

```
-------------------------------------------------------------
 Estimation of parameters under H0
 -------------------------------------------------------------

 Within sire standard deviation
 sire 910001  s.d. :     0.551
 sire 910045  s.d. :     0.578
 sire 910081  s.d. :     0.658
 sire 910088  s.d. :     0.654

  parameter                      estimable ?    value     precision

 General Mean                          yes      7.539     0.033

 Sire polygenic effects

          Sire 910001                  yes     -0.667     0.067
          Sire 910045                  yes     -0.448     0.058
          Sire 910081                  yes     -0.264     0.065
          Sire 910088                  no

 Dam polygenic effects

          Dam 910014 [Sire 910001]     yes      0.062     0.069
          Dam 910002 [Sire 910081]     yes     -0.052     0.073
          Dam 910010 [Sire 910081]     yes     -0.128     0.068
          Dam 910074 [Sire 910088]     yes     -0.220     0.075

  NOTE: known allelic origin means QTL effect =  maternal - paternal allele effects
                        ***
  The mean of absolute value of substitution effect WQ (in std unit) =
  -------------------------
```

Box 16.8: Example of main output file information, third section (Analysis under H0)

- The **fourth section**, provides results of H1 hypothesis analyses for each trait.

Details depend on tests and models, as detailed in the following table.

Section \ calcul	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Possible confusions between QTL and fixed effects or polygenic effects		x	x	x				
Residual variances and estimation of the main effects (polygenic,QTL)	x	x	x	x	x	x		x
LRT for the nuisance effects		x	x	x			x	

Risk Factor estimation							x	
Precision of the parameter estimation		x	x	x				
General mean estimation		x	x	x				
Fixed effect and covariate estimations		x	x	x			x	
Interactions between QTL and fixed effects		x	x	x			x	
Residual correlations between traits					x			

- Maximum of the test statistics

The value of the LRTmax and the maximum likelihood estimation of the QTL under the H1 hypothesis, with identification of its flanking markers are given.

```
    ** chromosome  7 **

          I---------------I---------------I---------------I---------------I-
    52.0 -                           oo
    50.6 I                          oo o
    49.2 I                          o  o
    47.9 I                         o    oo
    46.5 I                              oo
    45.2 I                        oo       o
    43.8 I                        o        o
    42.4 I                        o         o
    41.1 -                       o          o
    39.7 I                      o            o
    38.4 I                      o            o
    37.0 I                      o             o
    35.6 I                     o
    34.3 I                   ooo             o
    32.9 I                  oo               o
    31.6 I                 oo
    30.2 I                oo                  o
    28.8 -                o                   o
    27.5 I                o                    o
    26.1 I               o
    24.7 I              o                     o
    23.4 I              o                     o
    22.0 I             o                       o
    20.7 I             o                       o    ooo
    19.3 -             o                        o  oo oo
    17.9 I            o                         o o    oo
    16.6 I          oo                          o o     oooooo
    15.2 I          o                            o            oo
    13.9 I          o                                           o
    12.5 I         o                                             oo
    11.1 I       o                                               ooo
     9.8 I  ooooooooo                                              ooo
     8.4 I    o                                                      ooo
     7.1 I                                                             ooo
     5.7 -                                                               ooo

           I---------------I---------------I---------------I---------------I-
         0.01            0.43            0.86            1.28            1.70

 Maximum likelihood ratio test :

Test H0 / H1 :  51.96564

 The maximum is reached at position(s)    0.6800(Chr :7   )
 flanking marker (qtl:    1) ,SLA                ,S0102
```

Box 16.9: Example of main output file information, fourth section (Analysis under H1)

- Parameters estimation

Within sire residual variance estimations are printed under all tested hypotheses (global standard deviation for `--calcul=` 3, 25 and 27).
The maximum likelihood solutions for the parameters are given, with an indication about their precision (available only for `--calcul` =2, 3, 4), estimated from the diagonal element of the incidence matrix inverse in an equivalent fully linear model (the lower the better):
Parameters estimation, with estimability information and precision indicators are listed for:

- ✓ General mean
- ✓ Nuisance fixed and covariates factors (not for `--calcul=` 1, 5 and 6)
- ✓ Sire QTL effects
- ✓ Dam QTL effects (if their family size is over OPT_NDMIN)
- ✓ Sire polygenic effects
- ✓ Dam polygenic effects (if their family size is over OPT_NDMIN)

The mean of the absolute values of QTL effect obtained at the maximum LRT is edited.

```
-----------------------------------------------------------------
 Estimation of parameters under H1
 -----------------------------------------------------------------

 Within sire standard deviation
 sire 910001  s.d. :     0.492
 sire 910045  s.d. :     0.564
 sire 910081  s.d. :     0.568
 sire 910088  s.d. :     0.582

  parameter            estimable ?    value     precision

 General Mean               yes      7.524     0.033

 Sire QTL effects
          Sire 910001 1     yes     -0.164     0.021
          Sire 910045 1     yes     -0.140     0.028
          Sire 910088 1     yes     -0.226     0.021

 Dam QTL effects
          Dam 910014 1      yes     -0.300     0.049
          Dam 910002 1      yes     -0.482     0.050
          Dam 910010 1      yes     -0.214     0.055

 Sire polygenic effects
          Sire 910001       yes     -0.676     0.067
          Sire 910045       yes     -0.440     0.058
          Sire 910088       no

 Dam polygenic effects
          Dam 910014 [Sire 910001] yes      0.049     0.069
          Dam 910002 [Sire 910081] yes     -0.164     0.075
          Dam 910010 [Sire 910081] yes     -0.177     0.072

  NOTE: known allelic origin means QTL effect =  maternal - paternal allele effects
                        ***
  The mean of absolute value of substitution effect WQ (in std unit) =
  --------------------------
 | qtl     1 | wq :     0.669 |
  --------------------------
```

Box 16.10: Example of main output file information, fourth section (Analysis under H1)

- Parameters tests

For each of the nuisance effect (fixed effects and covariates), a LRT is reported with the value and significance level of the likelihood ratio when comparing a model with or without this effect. The significance level is the probability for the likelihood under the alternative hypothesis to be higher than the likelihood under the null (no effect). When this LRT exceeds the threshold of a Chi^2 test with (p-1) degrees of freedom (p being the number of levels for a fixed effect, 1 for a covariate) corresponding a 5%, a 1% or a 0.1 % type-I error, the effect can be declared significant.

```
*************************************************************************
Testing model effects


Tested effect     df.    Likelihood     p-value
                         ratio

f1          (direct effect)    23    100.823   1.000
f2          (direct effect)    10    121.576   1.000
sex         (direct effect)     2     11.146   1.000
...
```

Box 16.11: Example of main output file information, fourth section (Analysis under H1)

- Description of data structure

Warning about possible **confusions** between traits effects estimations are given.

```
*************************************************************************
Test of confusion between QTL and other effects in the final constained model
(test based on the correlation between columns of the incidence matrix)

*************************************************************************
Confusion between QTL and other effects (final constained model)

 No confusion detected
 the highest correlation is :   0.257
*************************************************************************
```

Box 16.12: Example of main output file information, fourth section (Analysis under H1)

- Interactions between QTL and fixed effects

When interactions between the QTL and *m* fixed effects are considered in the model, the dam (if needed) and sire QTL effects are estimated for each level of the composite interacting fixed effect (if n_1, n_2.. n_m are the number of levels for effects 1, 2...m, a maximum of $n_1 \times n_2 \times ... \times n_m$ QTL effects is estimated for each parent, as all levels of the interaction might not be represented in the progeny).

```
********************************************************************************
testing model effects

  Direct effects
Tested effect     df.    Likelihood     p-value
                         ratio

          sexe    1        120.35       0.000
         modal    1         10.91       0.001
           lot   13        126.57       0.000

  Intra qtl effects
Tested effect     df.    Likelihood     p-value
                         ratio
          sexe    4          1.72       0.787
 When this probability exceeds the standard threshold corresponding to the 5, 1 or 0.1
Pent level, you might consider removing this effect from the model
```

Box 16.13: Example of main output file information, fourth section (Analysis under H1)

- Risks factor estimation
- Residual correlations between traits
- Informativity of markers at the QTL position

After a description of the maximum likelihood QTL location (position on the chromosome, flanking markers), each family is considered in turn :

- ✓ Informativity (2 times mean absolute deviation to 0.5 of the transmission probability in the offspring). Closer to 0 lower the informativity.
- ✓ Haplotypes surrounding the QTL location, the haplotype length being defined by the *opt_longhap* key in the **parameter file**
- ✓ Grand parental origin of alleles at the QTL

```
Marker informativity at the maximum likelihood estimation

  0 < informativity < 1

  -------------------------------------
 QTL   1 position= 0.6700

Chromosome number tested =   1 [Chromosome=7]
Position number tested =  67/169


Left marker = SLA position =  0.6200
Right marker = S0102 position =  0.7400
                    ***
                    Sire  910001
  - Informativity = 0.990
  - Haplotype     = [3121/4432]
                    ***
                    Sire  910045
  - Informativity = 0.894
  - Haplotype     = [3651/5432]
                    ***
                    Sire  910081
  - Informativity = 0.940
  - Haplotype     = [21522/61243]
```

```
                    ***
                    Sire  910088
  - Informativity = 0.918
  - Haplotype     = [2821/4444]


 Allelic origin for 910001
Chromosome 7    : known
 Allelic origin for 910045
Chromosome 7    : known
 Allelic origin for 910081
Chromosome 7    : known
 Allelic origin for 910088
Chromosome 7    : known

 NOTE: known allelic origin means QTL effect =  maternal - paternal allele effects
```

Box 16.14: Example of main output file information, fourth section (Analysis under H1)

- Confidence intervals

For each detected QTL, confidence interval estimated within sire family or globally, by Boostrap, Drop Off or Hendge and Li methods are reported:

- ✓ QTL ID
- ✓ Position
- ✓ method
- ✓ Length of confidence interval
- ✓ Left bound
- ✓ Right bound
- ✓ left flanking marker
- ✓ position of left flanking marker
- ✓ Right flanking marker
- ✓ position of right flanking marker

```
  === Confidence Intervals ===
--------------------------- QTL = 1------------------------------
  Trait [Bardiere]
Name   Position     Method    Average   Pos Left  Pos Right Left flank mark  Pos    Right flank mark
Pos
                      ==  H 1/0 ==
QTL_1    0.670      Drop off     90.0%,    0.600     0.740     SW1369       0.520           S0102
0.740
QTL_1    0.670      Drop off     95.0%,    0.580     0.750     SW1369       0.520           SW352
1.010
QTL_1    0.670      Drop off     98.0%,    0.560     0.780     SW1369       0.520           SW352
1.010
                      ***
QTL_1    0.670     Hengde Li     90.0%,    0.590     0.743     SW1369       0.520           SW352
1.010
QTL_1    0.670     Hengde Li     95.0%,    0.575     0.762     SW1369       0.520           SW352
1.010
QTL_1    0.670     Hengde Li     98.0%,    0.557     0.785     SW1369       0.520           SW352
1.010
                      ***
```

Box 16.15: Example of main output file information, fourth section (Analysis under H1)

10.2. Output for eQTL analyses

A special output presents the analysis for each gene expression (depends the dynamic flag `--data-transcriptomic`). Only single trait analyses provide this output format.

For each hypothesis, the report gives:

- ✓ First line: a header indicating the content of the columns
- ✓ Next lines:
  - o first column: gene name
  - o others column: estimation of each parameter as indicated in the header

Note: Value 0.0 for the estimation means that the parameter is not estimable.

Under the null hypothesis "no QTL segregating", columns are the gene position on the expression array (as indicated in the eQTL performance file), the standard deviation of the distribution, the mean, the sire familial polygenic effects. The standard deviations and the polygenic means are given for each sire successively on the same line.

```
 Hypothesis :0
 Given parameters are respectively :
 Gene position on the array, [ *std dev *GMB11940,GMB11945][General Mean][ *Sire polygen-
ic effects*Sire GMB11940,Sire GMB11945]

 note : 0.0 means not estimable

         A_87_P019257  0.367  0.319  0.000 -0.000  0.000
         A_87_P012666  0.527  0.377 -0.060  0.060  0.000
         A_87_P011908  0.177  0.209 -0.003  0.003  0.000
         A_87_P009548  0.293  0.340  0.051 -0.051  0.000
         A_87_P021977  1.604  0.818 -0.044  0.271  0.000
         A_87_P041357  2.555  2.596  0.114 -0.114  0.000
         A_87_P003477  0.297  0.280 -0.019  0.019  0.000
         A_87_P003455  0.450  0.386 -0.042  0.042  0.000
         A_87_P008466  0.367  0.407  0.013 -0.013  0.000
         A_87_P014237  0.366  0.376 -0.005  0.005  0.000
         A_87_P014292  0.538  0.486 -0.050  0.050  0.000
         A_87_P018347  0.176  0.180  0.007 -0.007  0.000
         A_87_P022233  0.908  1.043 -0.086  0.028  0.000
         A_87_P016179  0.511  0.254  0.025 -0.092  0.000
         A_87_P004630  0.347  0.343  0.008 -0.008  0.000
        seq_RIGG19575  0.260  0.364 -0.022  0.022  0.000
        seq_RIGG14618  0.367  0.440  0.019 -0.019  0.000
            26_ACADSB  0.214  0.251 -0.006  0.006  0.000
```

Box 17.1: eQTL report under the null hypothesis

Under the alternative hypothesis "1 QTL segregating on the linkage group", columns are the gene position on the expression array (as indicated in the eQTL performance file), the chromosome where the QTL is detected, the QTL Position, the LRT for the test H0/H1, the standard deviation of the distribution, the mean, the sire QTL effect for each sire, the sire familial polygenic effects.
As the missing data may vary from one expression trait to another, information are pooled in "profile" sections, the missing data structure being homogeneous within section. This pooling facilitates the comparison of LRT to rejection thresholds which have to be computed independantly for each profile.

```
Profile    :          1
 Hypothesis :1
 Given parameters are respectively :
 Gene position on the array, Chromosome 1, QTL Position 1,H0/H1,[ *std dev
*GMB11940,GMB11945][General Mean][ *Sire QTL effects [1]*Sire GMB11940 l1[TT1T/AC2C],Sire
GMB11945 l1[AC2C/TT1T]][ *Sire polygenic effects*Sire GMB11940,Sire GMB11945]

 note : 0.0 means not estimable

A_87_P019257 5       0.739  4.033   0.354   0.316  -0.005   0.304  -0.040  -0.267   0.000
A_87_P021977 5       0.739 48.620   0.932   0.812  -0.030  -4.071   0.108   3.944   0.000
A_87_P016179 5       0.739 45.185   0.306   0.254   0.025   1.276   0.007  -1.249   0.000

 Profile    :          2
 Hypothesis :1
 Given parameters are respectively :
 Gene position on the array, Chromosome 1, QTL Position 1,H0/H1,[ *std dev
*GMB11940,GMB11945][General Mean][ *Sire QTL effects [1]*Sire GMB11940 l1[GCCC/ACCC],Sire
GMB11945 l1[ACTC/ACCC]][ *Sire polygenic effects*Sire GMB11940,Sire GMB11945]

 note : 0.0 means not estimable

A_87_P012666 5       0.139  6.622   0.491   0.376  -0.063  -1.430   0.021  -1.191   0.000
A_87_P011908 5       1.169  5.555   0.177   0.199  -0.027   0.001   0.086   0.027   0.000
```

Box 17.2: eQTL report under the alternative hypothesis of one QTL

Under the alternative hypothesis "2 QTL segregating on the linkage group(s)", columns are the gene position on the expression array (as indicated in the eQTL performance file), the chromosome where the first QTL is detected, the position of the first QTL, the chromosome where the second QTL is detected, the position of the second QTL, the LRT for the test H0/H2, the LRT for the test H1/H2, the standard deviation of the distribution, the mean, the sire QTL1 effect for each sire, the sire QTL2 effect for each sire, the sire familial polygenic effects.

```
Profile    :          1
 Hypothesis :2
 Given parameters are respectively :
 Gene position on the array, Chromosome 1, QTL Position 1,Chromosome 2, QTL Position
2,H0/H2,H1/H2,[ *std dev *GMB11940,GMB11945][General Mean][ *Sire QTL effects [1]*Sire
GMB11940 l1[TTGA/CAAA],Sire GMB11945 l1[TTGA/CAAG]][ *Sire QTL effects [2]*Sire GMB11940
l1[CGC4/TGC4],Sire GMB11945 l1[CGG4/CGC4]][ *Sire polygenic effects*Sire GMB11940,Sire
GMB11945]

 note : 0.0 means not estimable

A_87_P019257 5       0.339 5       0.429 11.772   7.739   0.367   0.285  -0.056   0.029
0.298  -0.023  -0.331   0.050   0.000
A_87_P012666 5       1.089 5       1.169 11.985   5.363   0.513   0.344  -0.076  -0.249
-0.325   0.104   0.304  -0.027   0.000
A_87_P011908 5       1.109 5       1.169 12.197   6.642   0.177   0.186  -0.028   0.013
-0.137  -0.009   0.215   0.030   0.000

 Profile    :          2
 Hypothesis :2
 Given parameters are respectively :
 Gene position on the array, Chromosome 1, QTL Position 1,Chromosome 2, QTL Position
2,H0/H2,H1/H2,[ *std dev *GMB11940,GMB11945][General Mean][ *Sire QTL effects [1]*Sire
GMB11940 l1[12CAC/12TTC],Sire GMB11945 l1[12TTC/12CAC]][ *Sire QTL effects [2]*Sire
GMB11940 l1[AGCC/GTTC],Sire GMB11945 l1[ATTC/AGCC]][ *Sire polygenic effects*Sire
GMB11940,Sire GMB11945]

A_87_P009548 5       0.799 5       1.049 11.366   6.904   0.281   0.316   0.057  -0.137
0.148  -0.105   0.155  -0.009   0.000
A_87_P021977 5       0.739 5       0.759 87.274  38.654   0.601   0.811  -0.034  -9.187
0.028   4.226   0.087   5.030   0.000
```

Box 17.3: eQTL report under the alternative hypothesis of two QTL

10.3. Analysis summary

In the file SUMMARY (**parameter file** key *out_summary*), several sections are given summarising the analysis.
For each hypothesis (H0: 0 qtl, H1: 1 QTL, H2: QTL, ...)
For each analysed variable (by line) :

- ✓ Number of genotyped progeny with phenotypes for the trait
- ✓ Maximum likelihood ratio
- ✓ QTL most likely position(s) (chromosome of each QTL, position of each QTL on the chromosome)
- ✓ for each sire
  - o Estimation of the QTL effect(s)
  - o Within sire family standard deviation
  - o Significance of each QTL effect(s) (based on a Student test). 'sign' = significant; 'ns'= not significant; 'na'=not available (too limited offspring size).

```
*******************************************************************************
 Summary 0 QTL versus 1 QTL
Variable  N    Max Lik     Pos (M)  Sire          910001              910045
               0/1QTL  Chr 1  Pos1         eff1     SD  sig1     eff1     SD  sig1
bardiere 236   45.2      1    0.7        -0.089  0.511  sign   -0.118  0.560  sign
     imf 236   43.7      1    0.7         0.156  0.338  sign    0.187  0.426  sign

 *******************************************************************************
 Summary 0 QTL versus 2 QTL,1 QTL versus 2 QTL
Variable  N     Max Lik           Pos (M)      Sire        910001                 910045
           0/2QTL 1/2QTL Chr 1 Pos1 Chr 2 Pos2 eff1 eff2    SD  sig1 sig2       eff1  eff2   SD   sig1 sig2
bardiere 236  57.0 11.9     1  0.7  1  1.1    -0.148  0.082 0.481 sign sign   -0.226  0.160 0.543 sign sign
     imf 236  49.3  5.6     1  0.9  1  1.0     0.405 -0.245 0.335 sign sign    0.415 -0.227 0.427 sign sign

 *******************************************************************************
 Summary 0 QTL versus 3 QTL,1 QTL versus 3 QTL,2 QTL versus 3 QTL
Variable  N        Max Lik        Pos (M)    Sire    910001        910045
               0/3QTL 1/3QTL 2/3QTL   Chr 1 Pos1   Chr 2 Pos2   Chr 3  Pos3            eff1 eff2 eff3      SD
sig1 sig2 sig3 eff1 eff2 eff3 SD  sig1 sig2 sig3
bardiere 236   63.9   18.8    6.9      1     0.7    1    0.8     1    1.1        -0.340  0.266  0.006  0.480  sign
sign  ns   0.211 -0.528  0.271  0.533  sign sign sign
imf      236   60.6   16.9   11.3      1     0.1    1    0.3     1    0.7        -0.123  0.092  0.132  0.324  sign
sign sign -0.439  0.540  0.072  0.408  sign sign  ns
```

Box 18: Summary with --qtl=3 option

10.4. Output of the LRT

The following key should be defined in the **parameter file** to output the LRT values for each tested position along the linkage group under hypothesis of one QTL segregating : *out_lrtsires, out_lrtdam,* and/or a grid output for the likelihood ratio test under hypothesis of 2 QTL: *out_grid2qtl*.

- o **LRT files: general test and sire family contributions** (*out_lrtsires)*

For each tested position, the file contains

- ✓ For the H1 : Chromosome, tested position, global LRT, Sire 1 LRT contribution, Sire 2 LRT contribution...
- ✓ For the H2 : Chromosome1, Chromosome2, tested position 1, tested position 2, global LRT, Sire 1 LRT contribution, Sire 2 LRT contribution...

```
Chr     Pos  GlobalLRT  910001 910045  910081 910088
   1   0.010    8.63    4.93    0.91    2.47    0.33
   1   0.020    8.62    4.82    1.03    2.47    0.30
   1   0.030    8.56    4.66    1.14    2.45    0.31
   1   0.040    8.45    4.47    1.23    2.41    0.35
   1   0.050    8.29    4.24    1.28    2.34    0.42
   1   0.060    8.35    4.21    1.35    2.31    0.48
...
  Chr1  Chr2   Pos1     Pos2 GlobalLRT 910001  910045 910081  910088
...
     1     1    0.02    0.65    3.78    2.72   -0.15   -1.11    2.32
     1     1    0.02    0.66    4.70    3.05    0.12   -0.38    1.92
     1     1    0.02    0.67    5.38    3.31    0.40    0.26    1.41
     1     1    0.02    0.68    5.80    3.51    0.70    0.79    0.80
     1     1    0.02    0.69    5.96    3.65    1.01    1.19    0.11
     1     1    0.02    0.70    5.86    3.71    1.32    1.46   -0.63
...
```

Box 19: LRT file: general test and sire family contributions

- **LRT files: dam family contributions** (*out_lrtdam)*

For each tested position, the file contains
Chromosome, Position, Dam 1 LRT contribution, Dam 2 LRT contribution…. (as in *out_lrtsires*)
Note: when the offspring size of a dam is below the threshold nd_min, the LRT is printed as 0.000 (see *opt_ndmin* option).

- **LRT files: grid 2 QTL** (*out_grid2qtl)*

The first part concerns the test of thypothesis 1 QTL *versus* the hypothesis 2 QTL,

- ✓ The fist line gives the tested position for the 1st QTL
- ✓ The following lines give the tested position for the 2nd QTL, followed by the LRT (1 *vs.*2 QTL) for each couple of positions.

```
++++++++++++++++ TEST 1QTL / 2QTL ++++++++++++++++++++

         .01     .02     .03     .04     .05     .06
  .01     .00    3.67    8.42   10.30   11.66   12.80
  .02     .00     .00    3.74    8.43   10.30   11.68
  .03     .00     .00     .00    3.81    8.43   10.31
  .04     .00     .00     .00     .00    3.87    8.44
  .05     .00     .00     .00     .00     .00    3.91
[…]
```

Box 20.1: LRT file: grid 2 QTL

The second part test of thypothesis 0 QTL *versus* the hypothesis 2 QTL,

- ✓ The fist line gives the tested position for the 1st QTL
- ✓ The following lines give the tested position for the 2nd QTL, followed by the LRT (0 *vs.*2 QTL) for each couple of positions.

```
  +++++++++++++++++ TEST 0QTL / 2QTL ++++++++++++++++++

         .01     .02     .03     .04     .05     .06
  .01     .00   27.46   32.21   34.09   35.45   36.59
  .02     .00     .00   27.53   32.22   34.09   35.47
  .03     .00     .00     .00   27.60   32.22   34.10
  .04     .00     .00     .00     .00   27.66   32.23
  .05     .00     .00     .00     .00     .00   27.70
...
```

Box 20.2: LRT file: grid 2 QTL

10.5. QTL effect estimations output

The following key should be defined in the **parameter file** to output the QTL effect estimations along the linkage group under hypothesis of one QTL segregating: *out_pateff, out_mateff.*

- **Sire QTL effects** (*out_pateff)*

For each tested position, the file contains for each tested position on all tested chromosomes: The chromosome, tested position, the sire 1 QTL effect estimation, the sire 2 QTL effect estimation …

```
  ****************************************
  This file is unvalide if interaction qtl case
  ****************************************
   Chr  Pos     910001   910045   910081   910088
   1   0.010   -0.24   -0.14   -0.13    0.02
   1   0.020   -0.24   -0.15   -0.14    0.01
   1   0.030   -0.24   -0.15   -0.14   -0.01
   1   0.040   -0.23   -0.16   -0.15   -0.03
   1   0.050   -0.22   -0.16   -0.15   -0.05
   1   0.060   -0.23   -0.16   -0.15   -0.06
   1   0.070   -0.23   -0.17   -0.16   -0.08
   1   0.080   -0.23   -0.17   -0.16   -0.09
...
Chr1    Chr2   Pos1     Pos2     910001/Qtl[1]    910001/Qtl[2]    910045/Qtl[1]    910045/Qtl[2]
910081/Qtl[1]    910081/Qtl[2]    910088/Qtl[1]    910088/Qtl[2]
     1     1    0.010    0.020    0.57    0.04    0.57    0.04    0.57    0.04    0.57    0.04
     1     1    0.010    0.030    0.24    0.04    0.24    0.04    0.24    0.04    0.24    0.04
     1     1    0.010    0.040    0.17    0.04    0.17    0.04    0.17    0.04    0.17    0.04
     1     1    0.010    0.050    0.14    0.04    0.14    0.04    0.14    0.04    0.14    0.04
     1     1    0.010    0.060    0.14    0.04    0.14    0.04    0.14    0.04    0.14    0.04
     1     1    0.010    0.070    0.14    0.03    0.14    0.03    0.14    0.03    0.14    0.03
     1     1    0.010    0.080    0.13    0.03    0.13    0.03    0.13    0.03    0.13    0.03
     1     1    0.010    0.090    0.12    0.02    0.12    0.02    0.12    0.02    0.12    0.02
...
```

Box 21: Sire QTL effect file

- **Dam QTL effects** (*out_mateff)*

For each tested position, the file contains for each tested position on all tested chromosomes: The chromosome, tested position, the dam 1 QTL effect estimation, the dam 2 QTL effect estimation …

Note: the QTL effect are given only for dams with offspring size larger than the threshold given by *opt_ndmin.*

10.6. Parental phase output

For each sire and dam (if the dam had more than NDMIN progeny) , the chromosomal phases are displayed on two lines, first for the paternal gamete and secod for the maternal gamete.

```
***************** SIRE PARENTAL PHASES *****************
                      CHROMOSOME :7

 910001 s 2 2 4 3 1 2 1 12 2 3
 910001 d 10 1 5 4 4 3 2 13 2 2
 910045 s 3 6 3 3 6 5 1 8 6 6
 910045 d 9 4 8 5 4 3 2 13 2 2
 910081 s 3 2 9 2 15 2 2 6 2 4
 910081 d 16 1 8 6 12 4 3 13 5 2
 910088 s 7 6 1 2 8 2 1 6 2 6
 910088 d 9 3 5 4 4 4 4 13 5 2

     ***************** DAM PARENTAL PHASES *****************
                      CHROMOSOME :7


None of the females had more than the minimum number of progeny needed to estimate its
possible phases
```

Box 22: Sire QTL effect file

10.7. Offspring phases

The progeny phases are output when the key *out_phases_offspring* is given a value in the **parameter file**.

For each progeny, two lines are printed, one for each phased chromosome. The sire chromosome is printed first, the dam chromosome second. For each marker, the transmitted chromosome of the parent is printed (1 for the first phase, 2 for the second phase as printed in the parental phase output). When known with certainty, only '1' and '2' are printed. When known with high probability (between 0.90 and 1), '1' and '2' are followed with a 'p'. When unknown (probability of the transmitted phase at the position is lower than 0.90), a '-' replaces the marker origin.

By default: progeny chromosomes phases are edited for all markers. The output can be reduced to a particular chromosomal region (around a QTL position for example), using the following parameters keys:

opt_phases_offspring_marker_start= First marker of the printed progeny phases

opt_phases_offspring_marker_end= Last marker of the printed progeny phases

The phases are then output for all the markers located between these bounds.

10.8. Marginal probabilities of the parental chromosome transmission

Each line gives for a tested QTL position:

- ✓ The tested position (cM)

- ✓ The sire ID
- ✓ The dam ID
- ✓ The dam phase number (as given in the main output and the parental phases output) when multiple phases are available for the dam
- ✓ The progeny ID
- ✓ The probability that the progeny inherited the 2nd sire chromosome (as given in the main output and the parental phases output) at position x given the dam phase
- ✓ The probability that the progeny inherited the 2nd dam chromosome (as given in the main output and the parental phases output) at position x given the dam phase (0.5 if the dam transmission probabilities are not considered in the analysis)

```
Position Sire    Dam  Dam_Phase Animal p(2nd sire allele) p(2nd dam allele)
    1.   910001 910014     1    944217        1.000               0.000
    2.   910001 910014     1    944217        0.999               0.001
    3.   910001 910014     1    944217        0.999               0.001
    4.   910001 910014     1    944217        0.999               0.001
    5.   910001 910014     1    944217        0.999               0.001
...
```

Box 23: Marginal probabilities of the parental chromosome transmission

10.9. Joint probabilities of the parental chromosome transmission

Each line gives for a tested QTL position x

- ✓ The tested position (cM)
- ✓ The sire ID
- ✓ The dam ID
- ✓ The dam phase number (as given in the main output and the parental phases output) when multiple phases are available for the dam
- ✓ The progeny ID
- ✓ The probability that the progeny inherited the 1st sire and 1st dam chromosome (as given in the main output and the parental phases output) at position x given the dam phase
- ✓ The probability that the progeny inherited the 1st sire and 2nd dam chromosome (as given in the main output and the parental phases output) at position x given the dam phase
- ✓ The probability that the progeny inherited the 2nd sire and 1st dam chromosome (as given in the main output and the parental phases output) at position x given the dam phase
- ✓ The probability that the progeny inherited the 2nd sire and 2nd dam chromosome (as given in the main output and the parental phases output) at position x given the dam phase

```
Position  Sire      Dam  Dam_Phase Animal  p(Hs1/Hd1) p(Hs1/Hd2) p(Hs2/Hd1) p(Hs2/Hd2)
    1.  910001   910014     1      944217     0.000      0.000      1.000      0.000
    2.  910001   910014     1      944217     0.001      0.000      0.999      0.001
    3.  910001   910014     1      944217     0.001      0.000      0.998      0.001
    4.  910001   910014     1      944217     0.001      0.001      0.998      0.001
    5.  910001   910014     1      944217     0.000      0.001      0.999      0.000
    6.  910001   910014     1      944217     0.001      0.001      0.941      0.056
    7.  910001   910014     1      944217     0.003      0.001      0.884      0.112
...
```

Box 24: Joint probabilities of parental segment transmission

10.10. Outputs for simulations

When data are simulated, a reduced output is given in the main result file. For each simulated trait, general parameters describing the distribution of the test statistic corresponding to the simulated data are printed (mean, standard deviation, minimum, maximum, skewness, kurtosis), together with a table containing the thresholds computed for 10%, 5%, 1%, 0.5%, 0.27%, 0.1%, 0.05%, 0.01% type-I errors. It should be mentioned that accurate thresholds are obtained only if a sufficient number of simulations are carried out. Typically, at least 1000 simulations should be run to compute thresholds corresponding to 5% type-I errors.

The printed thresholds correspond to type-I errors at the level of the linkage group for which data are simulated (typically, chromosome-wide thresholds). In case of multiple linkage groups/chromosomes analyses, it is recommended, using an approximate Bonferroni correction, to either adjust the type-I error for the number of independant linkage groups analysed, or adjust the type-I error as a proportion of the genome covered by each linkage group (when linkage groups have large different sizes, as in chicken).

```
       Variable traitsimul1
*--------------------------------------*
          Test 0vs1Q
*--------------------------------------*
  Test statistic distribution :
      Number of simulations:    100
      Mean                  :     14.24685
      Standard deviation    :      4.07168
      Skewness              :      0.70693
      Kurtosis              :      1.05302
      Minimum               :      6.62047
      Maximum               :     28.64581
 *--------------------------------------*
 | chromosome | genome     |  Threshold |
 |          level          |            |
 |-------------------------------------|
 |   0.1000   |            |     19.39  |
 |   0.0500   |            |     21.39  |
 |   0.0100   | chrom_level|     27.40  |
 |   0.0050   |      *     |     28.18  |
 |   0.0027   |   nb_chrom |     28.44  |
 |   0.0010   |            |     28.58  |
 |   0.0005   |            |     28.61  |
 |   0.0001   |            |     28.64  |
 *--------------------------------------*
```

Box 25: Output file from simulations

In addition to the main output, a summary output is provided. For each analysed variable, a

line is given with the empirical thresholds at 5%, 1% and 0.1 % at the chromosome and the genome level. The calculation of the genome-wide level corresponds to a genome scan of 18 autosomes (as in pigs). For other species, the genome-wide level can easily be obtained multiplying the chromosome-wide level by the number of chromosomes, or express it in proportion of the genome represented by the linkage group as explained in the previous paragraph.

```
    0vs1Q                              p_value at
Trait              chromosome level                    genome level
              5%         1%         0.1%          5%         1%         0.1%

traitsim     21.39      27.40      28.58         28.44      28.61      28.64
```

Box 26: Summary file from simulations

10.11. Detailed output of the LRT for simulations

The file corresponding to the key `out_maxlrt` of the **parameter file** contains the maximum LRT, the corresponding position and linkage group for each simulation/permutation.

For each analysed variable:

- ✓ a header
- ✓ for each simulation:
  - o the maximum likelihood ratio test
  - o the position and linkage group of the first QTL
  - o the position and linkage group of the second QTL (if 2 QTL hypothesis)
  - o ...

```
# Trait [traitsimul1]
LRTMAX H0/H1 CHR  Position
   12.7928    1    0.4100
   18.5180    1    0.1100
   17.0331    1    1.2100
 # Trait [traitsimul2]
LRTMAX H0/H1 CHR  Position
    8.9628    1    0.7100
    9.3228    1    1.0000
   16.6090    1    0.7100
```

Box 27.1: Detailed output of the LRT for simulations when comparing H0 "no QTL is segregating" versus H1 "1 QTL is segregating on the linkage group"

```
# Trait [traitsimul1]
LRTMAX H0/H1 CHR  Position  LRTMAX H1/H2   CHR1  Position1 CHR2 Position2
   12.7928    1    0.4100    9.6459          1    0.4100    1    1.2100
   18.5180    1    0.1100   14.2922          1    0.1100    1    1.0100
   17.0331    1    1.2100   15.4039          1    0.3100    1    1.2100
 # Trait [traitsimul2]
LRTMAX H0/H1 CHR  Position  LRTMAX H1/H2   CHR1  Position1 CHR2 Position2
    8.9628    1    0.7100   12.8711    1    1.5100    1    1.6100
    9.3228    1    1.0000    8.4281    1    0.0100    1    0.3100
```

```
   16.6090    1    0.7100    9.5829    1    0.3100    1    0.4100
```

Box 27.2: Detailed output of the LRT for simulations when comparing H0 "no QTL is segregating" versus H2 "2 QTL are segregating on the linkage group", and H1 "1 QTL is segregating on the linkage group"versus H2"2 QTL are segregating on the linkage group

11. References

Elsen JM, Filangi O, Gilbert H, Le Roy P, Moreno C, 2009. A fast algorithm for estimating transmission probabilities in QTL detection designs with dense maps. Genet Sel Evol., 41:50.

Elsen JM, Filangi O, Gilbert H, Le Roy P, Moreno C., 2009. A fast algorithm for estimating transmission probabilities in QTL detection designs with dense maps. Genetics Selection Evolution,41:50.

Elsen JM, Mangin B, Goffinet B, Boichard D, Le Roy P, 1999. Alternative models for QTL detection in livestock. I. General introduction. Genet. Sel. Evol., 31, 213-224

Gilbert H, Le Roy P, Moreno C, Robelin D, Elsen JM, 2008. QTLMAP, a software for QTL detection in outbred population. Annals of Human Genetics, 72(5): 694.

Gilbert H, Le Roy P., 2007. Methods for the detection of multiple linked QTL applied to a mixture of full and half sib families. Genet Sel Evol., 39(2):139-58.

Goffinet B, Le Roy P, Boichard D, Elsen JM, Mangin B, 1999. Alternative models for QTL detection in livestock. III. Heteroskedastic model and models corresponding to several distributions of the QTL effect.. Genet. Sel. Evol., 31, 341-350.

Knott S, Elsen JM, Haley C, 1996. Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. Theoretical and Applied Genetics, 93(1-2):71–806.

Larrosa J, Schiex T, 2004. Solving weighted CSP by maintaining arc consistency. Artificial Intelligence,159(1-2):1–26.

Legarra A, Fernando RL, 2009. Linear models for joint association and linkage QTL mapping. Genet Sel Evol., 41:43.

Mangin B, Goffinet B, Le Roy P, Boichard D, Elsen JM, 1999. Alternative models for QTL detection in livestock. II. Likelihood approximations and sire marker genotype estimations. Genet. Sel. Evol., 31, 225-237.

Moreno CR, Elsen JM, Le Roy P, Ducrocq V, 2005. Interval mapping methods for detecting QTL affecting survival and time–to–event phenotypes. Genet. Res. Camb., 85: 139-149.