Files, available on the net

Here we describe in details the programs, used for model analysis of the complete phenotype *Y* model *MD1* and simpler one *MD2* in OpenBugs (Version 2.2.0). These files include the most important steps in hierarchical modeling of cQTLs and eQTLs, described in Sillanpää and Noykova (Heredity, 2008). With these examples we give some insight how the OpenBugs 2.2.0 program could be used for similar research purposes.
The modeling results are visualized in small program file, written in Matlab.

The expression eQTL and clinical cQTL models

The expression eQTL model, used as a part of clinical cQTL model, is given as:

$$E_{\mathrm{i,j}} \left| I_j, \mu_j, A_j, G_{i,j}, \alpha_j, \sigma_j^2 \sim N(\alpha_j + I_j \mu_j A_j G_{i,j}, \sigma_j^2) \right. . \qquad (1)$$

The complete clinical cQTL model (called also further phenotype *Y* model MD1) is presented as:

$$Y_i = a + \sum_{j=1}^{N_p} \left(I_j^M \beta_j^M G_{i,j} + I_j^E \beta_j^E E_{i,j} + I_j^{ME} \beta_j^{ME} G_{i,j} E_{i,j} \right) + e_i . \qquad (2)$$

The meaning and exact expressions of all parameters and variables in these models are fully described in Sillanpää and Noykova (Heredity, 2008).

Here we analyze one simulated cQTL data set using two cQTL-models, differing in the complexity of the missing data model for the missing values of expressions. The first model (*MD1*) involves the eQTL-model (1) as a missing data model, and the second one (*MD2*) uses a much simpler model to handle the missing expressions, $E_{i,j} \sim N\left(0, \sigma_0^2\right)$, where $p(E|I, \mu, A, G, \sigma_0^2)$ is replaced simply by $p(E|\sigma_0^2)$.

The simulated cQTL data

Marker *G* data are simulated using WinQTL Cartographer (Version 2.5). These data spanned 3 chromosomes (C1, C2 and C3) of length 99cM, so that there are 34 evenly spaced markers on every chromosome. The distance between every two neighboring markers is d=3 cM. The sample size is N=200 individuals. These data are presented schematically on Figure 1.

The complete expression *E* data are simulated using the corresponding eQTL model file (according Eq. (1)) in OpenBugs (Version 2.2.0), and already simulated marker *G* measurements as an input data file. We assume that the number of marker-gene pairs is Np =120.
Here we do not describe in details the simulation eQTL model file, because it is entirely involved in the cQTL model file (describing the model *MD1*), used later on during the simulation and model analysis of cQTL data.

Next the complete phenotype *Y* data are simulated in OpenBugs (Version 2.2.0). For this purpose we created the simulation model file in OpenBugs (Version 2.2.0). The simulation cQTL model corresponds to Eq.(2), where eQTL-model (1) is involved as a missing expression *E* data model. This simulation model code is very similar to the code, used in the analysis stage. Therefore we do not explain it separately.

As an input data file in OpenBugs we have used already simulated marker *G* and expression *E* complete data sets.
At this stage we have also calculated the joint heritability of the simulated phenotype data.

After that some marker *G* and expression *E* data are deleted at random in Matlab, so that the final simulated data file includes 5% missing *G* data and 50% missing *E* data.

For better insight at the end of this document we also attach the list of all files, used during the simulation stage, as well as the corresponding files.

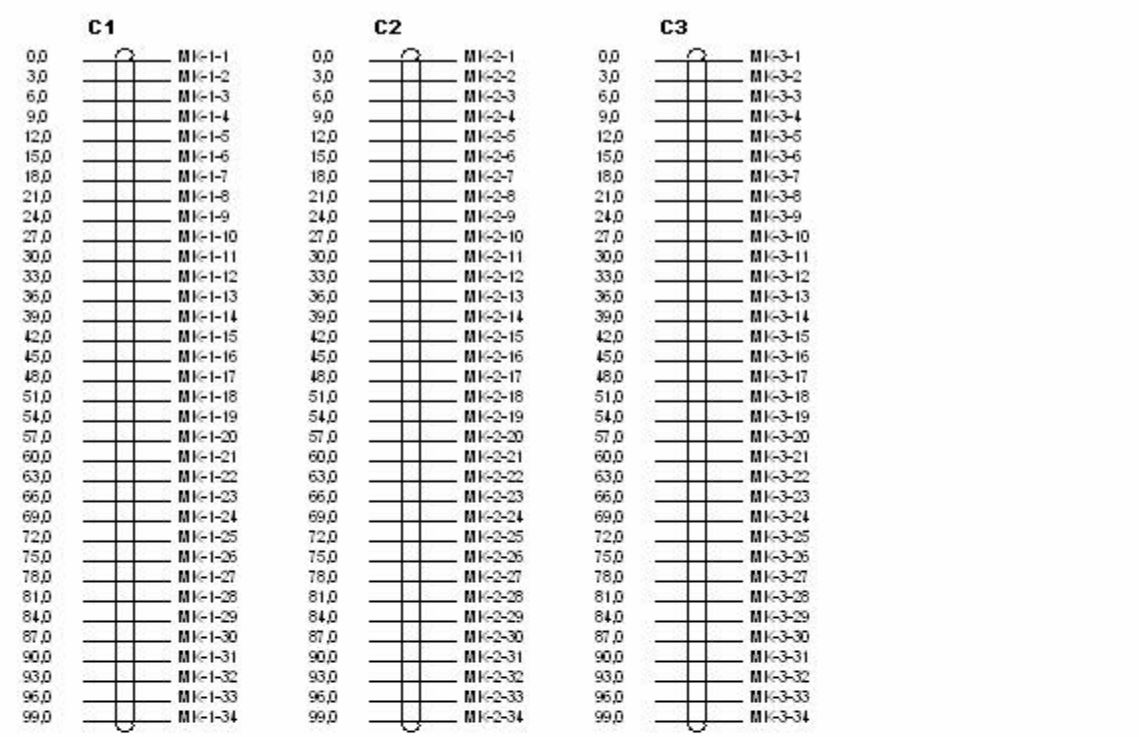

Figure 1. Schematic presentation of the markers, used for obtaining the simulated data. The data spanned 3 chromosomes (C1, C2 and C3) of length 99cM, so that there are 34 evenly spaced markers on every chromosome. The distance between every two neighbor markers is d=3 cM.

Analysis of the simulated cQTL data using the model *MD1*

Here we describe the OpenBUGS code of the two model files, used during the model analysis:
the file named ***ymodelEsigmaGmissing.odc***, which describes the model *MD1*, and the file named ***ymodelEwithoutGmissingbugslang.odc*** , which describes the model *MD2*.

Model file: ***ymodelEsigmaGmissing.odc*** (*MD1*):

```
model{

# Prior distributions:

# Priors of indicator, effect size and assignment variable, involved in the eQTL model

  for( j in 1 : Np ) {
    Im[j] ~ dbern(0.9)                    #The indicators prior of Im assumes that most of the
                                          #pairs j (90%) exhibit eQTL regulatory effect.

    mu[j] ~ dnorm(0,0.01)I(0,)            # Normally distributed weakly informative prior
                                          # of the effect size mu. Thus only moderate positive
                                          # values of mu are simulated.

     A1[j] ~ dbern(0.5)                   # Mutually independent prior of assignment variable A.
                                          # A1 – auxiliary variable. It can take values 0 (50%) or 1
                                          # (50%).
  }

# Priors of indicators, involved in the cQTL model:

  for( j in 1 : Np ) {
    IE[j] ~ dbern(bernprop)               # bernprop = 0. 0033
    IM[j] ~ dbern(bernprop)               # (bernprop =0.0294 during the data simulation).
    IEM[j] ~ dbern(bernprop)
      }

# Inverse Gamma priors of effect sizes, involved in the cQTL model

 for( j in 1 : Np ) {
    tauM[j] ~ dgamma(1,1)
    tauE[j] ~ dgamma(1,1)
    tauEM[j] ~ dgamma(1,1)
  }

   for( j in 1 : Np ) {
    thetaM[j] ~ dnorm(0.0,tauM[j])
    thetaE[j] ~ dnorm(0.0,tauE[j])
    thetaEM[j] ~ dnorm(0.0,tauEM[j])
  }
```

 # Modeling missing genotype data G in case of backcross, $p(G) \propto \prod_{i=1}^{N}\left[p(G_{i,1})\prod_{j=1}^{N_p} p(G_{i,j}\left|G_{i,j-1}\right.)\right]$

Prior transition probability *P[j - 1 , i]* from genotype *G[j-1,i]* at marker *j-1* to *G[j,i]* at marker *j*.
It takes values 0.97 if *G[j, i] =G[j - 1 , i]* (no recombination), and 0.0291 otherwise.

$P\left(G_{ij}\left|G_{ij-1}\right.\right)=\begin{cases}1-r_j, & if\ \ G_{ij}=G_{ij-1}\ (no\ recombination)\\ r_j & otherwise\end{cases}$

According Haldane's formula, $r_j=\frac{1}{2}\left(1-\exp\left(-2\left|d_j\right|\right)\right),\ d_j=0.03M$, the recombination fraction *r[j]*
for this particular example is *r[j]* =0.0291.
The genotype data *G* take values 0 or 1.

#(The model for missing genotype data *G* is not involved during the simulation stage because it is not needed for simulation of complete data sets).

```
for( i in 1 : N ) {
        P[1 , i] <- 0.5
  G[1 , i] ~ dbern(P[1 , i])                    # Probability of genotype G at marker 1, P(G_i1) = 1/2 .
  }

for( i in 1 : N ) {
     for( j in 3 : Np ) {
        P[j - 1 , i] <- G[j - 1 , i] * 0.97 + (1 - G[j - 1 , i]) * 0.0291
     }
  }


for( i in 1 : N ) {
     for( j in 2 : Np ) {
        G[j , i] ~ dbern(P[j - 1 , i])              # Probability, assigned to the missing G[j , i] values.
     }
  }
```

 # Priors of the regression parameters, involved in eQTL and cQTL model:

```
  a ~ dnorm(0,taua)                          # taua=0.0001 – variance of the regression
                                             # parameter a.
  tau0 ~ dgamma(1,1)                         # variance of the phenotype Y
                                             # (During the data simulation stage tau0 <- 0.065)

  tau ~ dgamma(1,1)I(0.0001,2)               #The limits 0.0001 and 2 are given randomly,
                                             #mostly because of the OpenBUGS restrictions.
  sigma <- 1 / sqrt(tau)
  sigma2 <- 1 / tau                          # variance of the expression E.
```

 # Model level II

```
 for( j in 1 : Np ) {
    Imu[j] <- Im[j] * mu[j]                    # The product Imu forms
                                               # the eQTL regulatory effect of pair j (Eq.(1)).

    A[j] <- 2 * A1[j] – 1                      # The prior values of A1 are 0 or 1.
                                               # Assignment variable A takes values -1 or 1.
  }

   for( j in 1 : Np ) {
    IMtheta[j] <- IM[j] * thetaM[j]            # These products form the cQTL regulatory effect
    IEtheta[j] <- IE[j] * thetaE[j]            # of pair j for the corresponding
    IEMtheta[j] <- IEM[j] * thetaEM[j]         # regression terms in Eq.(2).
  }
```

 # Model level III

 # Expression QTL model $E_{i,j} = \alpha_j + I_j \mu_j A_j G_{i,j} + \varepsilon_{i,j}$ **:**

```
 for( i in 1 : N ) {
    for( j in 1 : Np ) {
      meanE[j , i] <- (Imu[j] * A[j]) * G[j , i]    # G[j , i] are taken from the data file. Prior model values
                                                    # are assigned to the missing G[j , i] measurements.

      E[j , i] ~ dnorm(meanE[j , i],tau)            # Bimodal mixture distribution of the expression E[j , i].
}                                                   # It is assigned to the missing E[j , i] data.
  }
```

 # Model level IV

 # Clinical cQTL model $Y_i = a + \sum_{j=1}^{N_p} \left(I_j^M \beta_j^M G_{i,j} + I_j^E \beta_j^E E_{i,j} + I_j^{ME} \beta_j^{ME} G_{i,j} E_{i,j} \right) + e_i$ **:**

```
  for( i in 1 : N ) {
  for(j in 1:Np) {
        new[j,i]<-IMtheta[j] * (2 * G[j , i] - 1) + IEtheta[j] * E[j , i] + IEMtheta[j] * (2 * G[j , i] - 1)* E[j , i]
      }
        meanY[i] <- a+sum(new[1:Np,i])

      Y[i] ~ dnorm(meanY[i],tau0)                   # Normally distributed single trait phenotypes Y[i].

}
```

<u>Input data file:</u> ***ymodelmissingGEYdata.odc*** - 5% missing *G* data, 50% missing *E* data, and complete *Y* data.

Analysis of the simulated cQTL data using the model *MD2*

The difference comparing to the file ***ymodelEsigmaGmissing.odc*** (*MD1*), is in the **Model level III**, where the missing expression E values are modeled as $E_{i,j} \sim N\left(0, \sigma_0^2\right)$.

<u>Model file:</u> ***ymodelEwithoutGmissingbugslang.odc*** (*MD2*):

 # Model level III

 # Expression QTL model $E_{i,j} \sim N\left(0, \sigma_0^2\right)$**:**

```
for( j in 1 : Np ) {
    tauEmodel[j]~dgamma(1,1)
  }

 for( i in 1 : N ) {
     for( j in 1 : Np ) {
      meanE[j , i] <- (Imu[j] * A[j]) * (2 * G[j , i] - 1)
        E[j , i] ~ dnorm(0,tauEmodel[j])                    # Simpler eQTL model
                                                            # for missing E[j , i] measurements.
     }
  }
```

<u>Input data file:</u> ***ymodelmissingGEYdata.odc*** - 5% missing *G* data, 50% missing *E* data, and complete *Y* data.

Parameter estimation in OpenBugs (Version 2.2.0).

The parameter estimation for both model files (***ymodelEsigmaGmissing.odc***, (*MD1*), and ***ymodelEwithoutGmissingbugslang.odc*** (*MD2*)) is provided in the same way.

Running the model and data files:
Model *ymodelEsigmaGmissing.odc (ymodelEwithoutGmissingbugslang.odc*) - Specification tools – check model – load data *ymodelmissingGEYdata.odc*– compile – *gen inits*;
Inference – Samples: *IMtheta (IEtheta, IEMtheta, Im)*;
Model-Update:(*10 000 – burn in) 100 000*;
During the estimation first 10 000 MCMC iterations are omitted because of burn-in.
The posterior estimates are based on next 100 000 iterations:
To see the estimation results: Inference – Samples *IMtheta (IEtheta, IEMtheta, Im)* - Stats: *IMtheta (IEtheta, IEMtheta, Im)*.

Remark: The estimation programs in OpenBUGS are very slow because the investigated models are complicated.

Visualization of the results in Matlab

Suitable visualization of the estimation results, obtained in WinBugs, is provided in Matlab.
The name of the Matlab file, where all estimation results are saved, is ***Figyestimmissingsimdata.m***

List of the files, used during the analysis of simulated cQTL data

Models: ***ymodelEsigmaGmissing.odc*** (MD1)
ymodelEwithoutGmissingbugslang.odc (MD2)

Input data: ***ymodelmissingGEYdata.odc*** - 5% missing *G* data, 50% missing *E* data and complete *Y* data.

Visualization of the results
Figyestimmissingsimdata.m – Matlab file, where all estimation results are saved.

References:

1. Sillanpää, M., and N. Noykova, 2008, Hierarchical modeling of clinical and expression quantitative trait loci. Heredity (to appear).

2. Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn, 2005 *WinBUGS User Manual, Version 2.10.* MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

3. Thomas, A., R. B. O'Hara, U. Ligges, and S. Stuartz, 2006 Making BUGS open. R News **6**:1.

Attachments

List of the files, used during the simulation of cQTL data

Simulating genotypes

Results: ***Gijdataload.odc*** – this data file in OpenBUGS includes only marker data *G*

Simulating expressions

Model: ***Eijsimuldata.odc***
mu~dnorm(0, 0.01) I (0, no upper limit)
A1~dbern(0.5),
A[j] <- 2*A1[j]–1, j=1,…,102,
E[j , i] ~ dnorm(meanE[j , i],1)

Input data: ***Gijdataload.odc***
Results: ***ymodelGEdata.odc*** – includes the complete *G* (take values -1 or 1) and *E* data.

Simulating phenotypes

Model: ***ymodelbugslang.odc*** - the phenotype *Y* model *MD1*.
tau0 <- 0.065

bernprop = 0.0294. Thus nine components: 1 marker, 5 expressions and 3 genotype x expression interactions have been simulated. We assume that very small part of the pairs have cQTL regulatory effect.

Input data: earlier version of ***ymodelwithoutmissingdata.odc***, where cQTL data are not added, and the parameter bernprop =0.0294.

Results: ***ymodelwithoutmissingdata.odc -*** the complete *G* (take values 0 or 1) and *E* data + the new simulated cQTL data.
h21.m - the Y values are saved there

Calculation of the joint heritability of the simulated phenotype data.

Model: ***h21.m*** – calculates the joint heritability of the data (h21=0.68).

Simulating missing *G* and *E* data at random

Model: ***ymissingsimdata.m*** – simulate missing G and E values at random
Input data: ***ymodelGEdata.odc***
Results: ***ymodelmissingEdata.odc*** – complete *G* data and 50% missing *E* data;
ymodelmissingGEYdata.odc - 5% missing *G* data and 50% missing *E* data + complete *Y* data.