RJaCGH: A package for the analysis of CGH arrays through Reversible Jump MCMC.

Oscar M. Rueda[1] and Ramón Díaz-Uriarte[1]

February 23, 2008

1. Statistical Computing Team. Structural Computational Biology Group. Spanish National Cancer Center (CNIO), Madrid (SPAIN). omrueda@cnio.es, rdiaz@ligarto.org

Contents

1 Overview:

RJaCGH is an R package designed for the analysis of microarray CGH data. In this type of problems we have a collection of log-ratios that measure the ratio between the copy number of sequences of nucleotides between a test sample and a control sample for a number of probes. The main goal of the analysis is to detect which of those probes have a normal copy number, a loss copy number or a gained copy number.
This package basically fits a Non Homogeneous Hidden Markov Model through Reversible Jump Markov Chain Montecarlo. That is, we assume that there are k different groups (hidden states; different copy number ratios) within the data. Each of those groups follows a normal distribution with parameters μ_k and σ_k^2. The movements between those hidden states follow a Markov process whose transition probabilities depend on the distance between probes. The estimation of the parameters is made through a Markov Chain Monte Carlo (MCMC) algorithm. These techniques are based on the exploration of the parameter space through sampling. Instead of fitting several models and selecting just one, RJaCGH uses reversible jump [3] to jump between models and get the posterior probability for each of them. We can make birth/death moves (create

or delete a hidden state) and split/combine moves (separate or merge existing states). The inferences are then based on all models visited through Bayesian Model Averaging.

The package estimates the probability for every probe to have a normal copy number, gained or lost and computes probabilistic common regions. This vignette shows some of the package features with small examples. The references give full details about the statistical model and the parameterization it uses, plus further details of the algorithm.

Please note that our methods are computer intensive, so they may take a long time on a slow machine.

2 Data:

We use for the examples the public data set of Snijders et al. [5] with 15 human cells with known karyotypes, as found in the objects from package GLAD 1.6.0. [6].

3 Examples:

3.1 Same model for the whole genome

We'll analyze data cell gm13330 from [5]. First, we take out the missing values, because RJaCGH does not handle NA's. We'll need the log-2 ratios, the positions and the chromosome number:

```
> set.seed(1)
> library(RJaCGH)
> data(snijders)
> y <- gm13330$LogRatio[!is.na(gm13330$LogRatio)]
> Pos <- gm13330$PosBase[!is.na(gm13330$LogRatio)]
> Chrom <- gm13330$Chromosome[!is.na(gm13330$LogRatio)]
```

Now, we are going to fit the model through the function `RJaCGH()`. But first we must decide if we want to fit a model with equal variances for all the hidden states or with different variances. This can be set with the argument `var.equal=TRUE` (default) or `var.equal=FALSE` in the call to `RJaCGH()`. Besides, we can fit the same model to the whole genome or a different one for each chromosome. We can set this option with `model="genome"` or `model="Chrom"` in the call to `RJaCGH()`. In this section we will fit the same model for the whole genome.

We can also set the maximum number of hidden states that we want to fit. For example, we will fit HMMs with a maximum of four hidden states, so we'll set the parameter `k.max=4`.
Besides, we can set, if we wish to, the jumping parameters of the MCMC. They control the exploration of the probability distribution of the model via setting the jumps we make from a particular value of the parameters to a new one. There are two types of them:

- The standard deviation of the candidates of the jumps of the chain within a given model: `sigma.tau.mu`, `sigma.tau.sigma.2` and `sigma.tau.beta`.

They are vectors of length `k.max`. They are related to the dispersion within models.

- The standard deviation of the jumps between models in split/combine moves: `tau.split.mu` and `tau.split.beta`. They are scalars and are related to the dispersion between models.

We must remember that these are not parameters of the model, in the sense that different values produce different models. They are parameters of the algorithm that speed up or assure convergence.

We have to enclose them in a list. By some inspection of the data and/or trial/error we set them to the following values:

```
> jump.parameters <- list(sigma.tau.mu = rep(0.01, 4), sigma.tau.sigma.2 = rep(0.05,
+     4), sigma.tau.beta = rep(0.1, 4), tau.split.mu = 0.1, tau.split.beta = 0.1)
```

The arguments `burnin` and `TOT` control the number of iterations of the algorithm (the burn-in and the after burn-in).

We can also pass other arguments, such as the starting base and end base of the probes (`Start`, `End`), the distance between probes (`Dist`), the names of the probes (`probe.names`), the maximal distance between probes beyond which we consider them independent (`max.dist`)... See the help file for `RJaCGH()` for full reference.

```
> fit <- RJaCGH(y = y, Pos = Pos, Chrom = Chrom, model = "genome",
+     var.equal = TRUE, k.max = 4, burnin = 50000, TOT = 10000,
+     jump.parameters = jump.parameters)
```

```
Starting Reversible Jump
```

After the fit (it may take a little while), `RJaCGH()` returns an object with several interesting components. Its structure is a list with several lists nested inside of it; one for each model fitted. For example,

```
> fit[[4]]
```

is another list with the results of the fit of a model with 4 hidden states. There are several elements inside; for example, we can get the means and variances of the hidden states fitted:

```
> fit[[4]]$mu
> fit[[4]]$sigma.2
```

They are matrices with as many rows as samples have been drawn from a model with 4 hidden states and as many columns as hidden states (that is, four).

```
> apply(fit[[4]]$mu, 2, mean)
```

```
[1] -0.83938775 -0.07825253  0.03520531  0.52463888
```

This would be the mean of the posterior distribution of the means of the 4 hidden states.

In the case of the functions of transition probabilities:

```
> fit[[4]]$beta
```

is an array with the first and second dimensions the number of hidden states and the third the number of MCMC iterations in that model. So

```
> apply(fit[[4]]$beta, c(1, 2), mean)
```

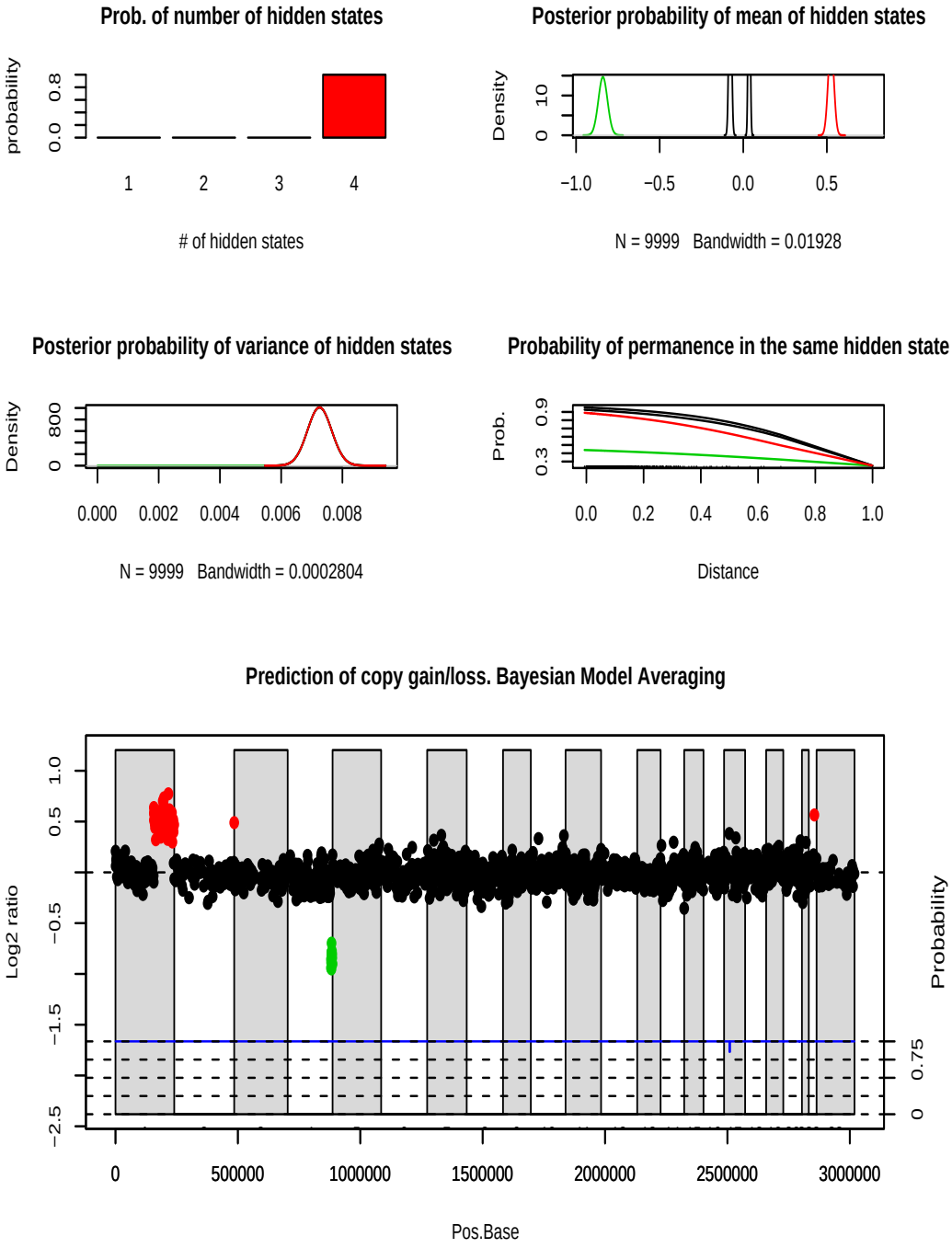
```
         [,1]      [,2]     [,3]     [,4]
[1,] 0.000000 0.2421245 2.001151 2.166395
[2,] 6.450002 0.0000000 2.562499 7.175590
[3,] 8.040229 3.1005899 0.000000 5.811885
[4,] 4.182116 4.0746709 2.493096 0.000000
```

would give the mean of the posterior distribution of `beta` (these are parameters of the transition matrix that depends itself on the distance between genes; see references for details on the model).
We can also summarize the fit and inspect these results. By default, `summary` returns the quantiles of the posterior distributions for the means and variances and the median of the parameters for the transition probabilities:

```
> summary.HMM <- summary(fit)
> summary.HMM

Distribution of the number of hidden states:

1 2 3 4
0 0 0 1

Model with  4  states:

Distribution of the posterior means of hidden states:
          10%    25%    50%    75%    90%
Loss-1 -0.864 -0.852 -0.839 -0.827 -0.814
Normal -0.085 -0.082 -0.078 -0.075 -0.071
Normal  0.030  0.032  0.035  0.038  0.040
Gain-1  0.509  0.517  0.525  0.533  0.541

Distribution of the posterior variances of hidden states:
         10%   25%   50%   75%   90%
Loss-1 0.007 0.007 0.007 0.007 0.008
Normal 0.007 0.007 0.007 0.007 0.008
Normal 0.007 0.007 0.007 0.007 0.008
Gain-1 0.007 0.007 0.007 0.007 0.008

Parameters of the transition functions:
       Loss-1 Normal Normal Gain-1
Loss-1  0.000  0.029  1.730  2.033
Normal  6.345  0.000  2.555  7.020
Normal  7.925  3.102  0.000  5.772
Gain-1  4.027  3.939  2.421  0.000
```

We can also plot the model with higher posterior probability and the classification of genes using information from all models visited: that is, through Bayesian Model Averaging:

```
> plot(fit, cex = 1.1)
```



The color 'green' is assigned to states of loss, the 'black' to normal states and the 'red' to gains. Note that two states have been labeled as 'Normal'. As statistical states do not always correspond to biological states (for example, a mixture of two normal distributions -two hidden states- might be needed to

fit the distribution of the normal copy numbers), RJaCGH does an automatic labeling based on the posterior means and variances of the hidden states and the arguments `normal.reference` (the reference value for the mean of the normal state -no change-), `normal.ref.percentile` (all states with credible intervals of this probability including the `normal.reference` are labeled as 'Normal') and `auto.label` (the minimum proportion of normal probes). The user can explore different thresholds with (not shown):

```
> plot(relabelStates(fit, normal.ref.percentile = 0.99))
> plot(relabelStates(fit, normal.ref.percentile = 0.05))
```

When a good labeling is found, we can update the fit:

```
> fit <- relabelStates(fit, normal.ref.percentile = 0.75)
```

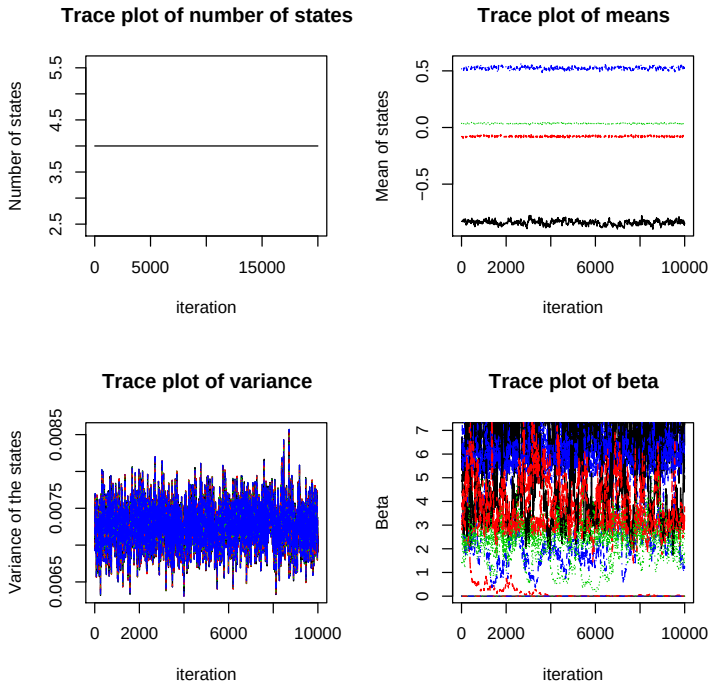
If the user wants to make his own relabeling of states, he has to define `fit[[k]]$state.labels`, for the model `k` of interest. It must be a vector of length `k` with elements 'Loss', 'Normal' or 'Gain'. For example, in our case:

```
> fit[[4]]$state.labels <- c("Loss", rep("Normal", 2), "Gain")
```

There are other methods to extract more information, as `states()`, `model.averaging()` or `smoothMeans()`. They will be introduced in the next section.

We can also inspect the convergence of the most visited model:

```
> trace.plot(fit)
```



If we don't see good mixing we can re-adjust the jumping parameters:

- If the lines are too straight for some parameters, we must reduce its corresponding jumping parameters and refit.
- If the lines oscilate too much, we should refit with greater jumping parameters.
- The parameters that rule the movements amongst states are `tau.split.mu` and `tau.split.beta`, and the parameters that rule the means, the variances and `beta` are `sigma.tau.mu`, `sigma.tau.sigma.2` and `sigma.tau.beta`.

We can also check the good mixing of the algorithm looking at the proportion of the different values sampled for μ, σ^2 and β: it should not be very low nor very high; some authors say that it should roughly be around 0.23. We'll do it for the model with highest posterior probability:

```
> maxK <- as.numeric(names(which.max(table(fit$k))))
> fit[[maxK]]$prob.mu

[1] 0.1628163

> fit[[maxK]]$prob.sigma.2

[1] 0.3794379

> fit[[maxK]]$prob.beta

[1] 0.2253225
```

And finally, we can check that the algorithm has made some jumps between models (birth, death, split and combine movements):

```
> fit$prob.b

[1] 3

> fit$prob.d

[1] 9

> fit$prob.s

[1] 9

> fit$prob.c

[1] 2
```

(Note that these numbers include the burn-in iterations, but the `trace.plot()` not.)

3.2 A different model for every chromosome

We can also fit a different model for every chromosome with the function `RJaCGH()` changing the parameter `model` to 'Chrom'. We'll fit a model to other cell line: 01524. If there is lot of difference in variance between chromosomes every chromosome should have its own set of jumping parameters, so we shouldn't specify them and let `RJaCGH` do a simple search to find 'good' ones. In this example there is no such different variances per chromosomes, but we'll let the program choose them as a demonstration:

```
> y2 <- gm01524$LogRatio[!is.na(gm01524$LogRatio)]
> Pos2 <- gm01524$PosBase[!is.na(gm01524$LogRatio)]
> Chrom2 <- gm01524$Chromosome[!is.na(gm01524$LogRatio)]
> fit.chrom <- RJaCGH(y = y2, Pos = Pos2, Chrom = Chrom2, model = "Chrom",
+     k.max = 4, burnin = 50000, TOT = 10000)
```

Again, the result of the fit are nested lists. We can access every chromosome in a simple way, because there is a list for every chromosome, and every chromosome is an object of the same class as explained in the former section. For example, to inspect the chromosome 6 we would do the following:

```
> summary(fit.chrom[[6]])

Distribution of the number of hidden states:

    1     2     3     4
0.000 0.993 0.007 0.000

Model with  2  states:

Distribution of the posterior means of hidden states:
          10%    25%   50%   75%   90%
Normal -0.009 -0.001 0.008 0.015 0.022
Gain-1  0.516  0.528 0.541 0.554 0.565

Distribution of the posterior variances of hidden states:
         10%   25%   50%  75%   90%
Normal 0.007 0.008 0.009 0.01 0.011
Gain-1 0.007 0.008 0.009 0.01 0.011

Parameters of the transition functions:
       Normal Gain-1
Normal  0.000  4.015
Gain-1  2.544  0.000
```

We can also see the sequence of hidden states, that is the copy number status for every probe. We can compute it conditionally to a particular model, (with the method `states`) or averaging through every model fit weighted by the posterior probability of that model (method `model.averaging`):

```
> sequence <- states(fit.chrom)
> sequence.averaged <- model.averaging(fit.chrom)
```

We can see the copy number of chromosome 6:

```
> head(sequence[[6]]$states)
```

```
[1] Normal Normal Normal Normal Normal Normal
Levels: Normal Gain-1
```

```
> head(sequence.averaged[[6]]$states)
```

```
[1] Normal Normal Normal Normal Normal Normal
Levels: Loss < Normal < Gain
```

And the probability of every state in that chromosome:

```
> head(sequence[[6]]$prob.states)
```

```
     Normal Gain-1
[1,]      1      0
[2,]      1      0
[3,]      1      0
[4,]      1      0
[5,]      1      0
[6,]      1      0
```

```
> head(sequence.averaged[[6]]$prob.states)
```

```
     Loss    Normal        Gain
[1,]    0 0.9987433 0.001256689
[2,]    0 0.9972173 0.002782669
[3,]    0 0.9972173 0.002782669
[4,]    0 0.9972173 0.002782669
[5,]    0 0.9972173 0.002782669
[6,]    0 0.9972173 0.002782669
```

We can also see the smoothed values for every probe (this method returns a vector, not a list with as many vectors as chromosomes):

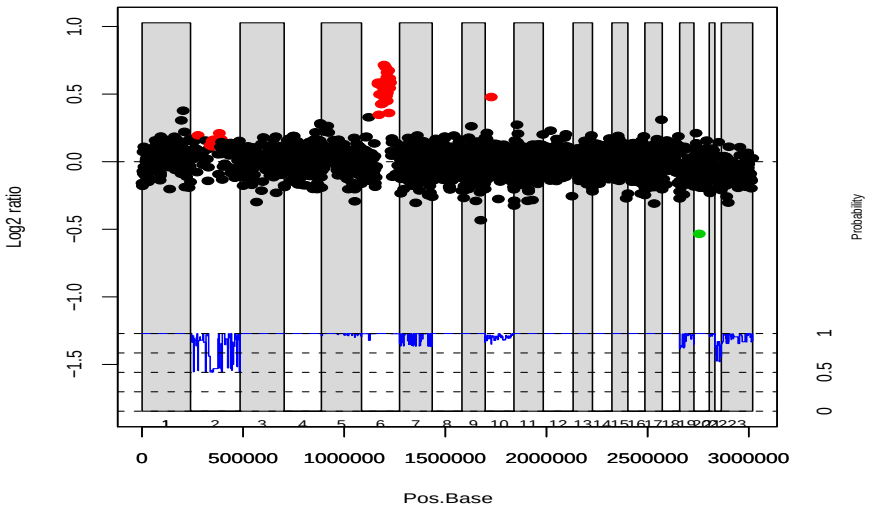
```
> s.means <- smoothMeans(fit.chrom)
> head(s.means)
```

```
[1] 0.004903643 0.004903643 0.006610641 0.022967068 0.012319315 0.026532209
```

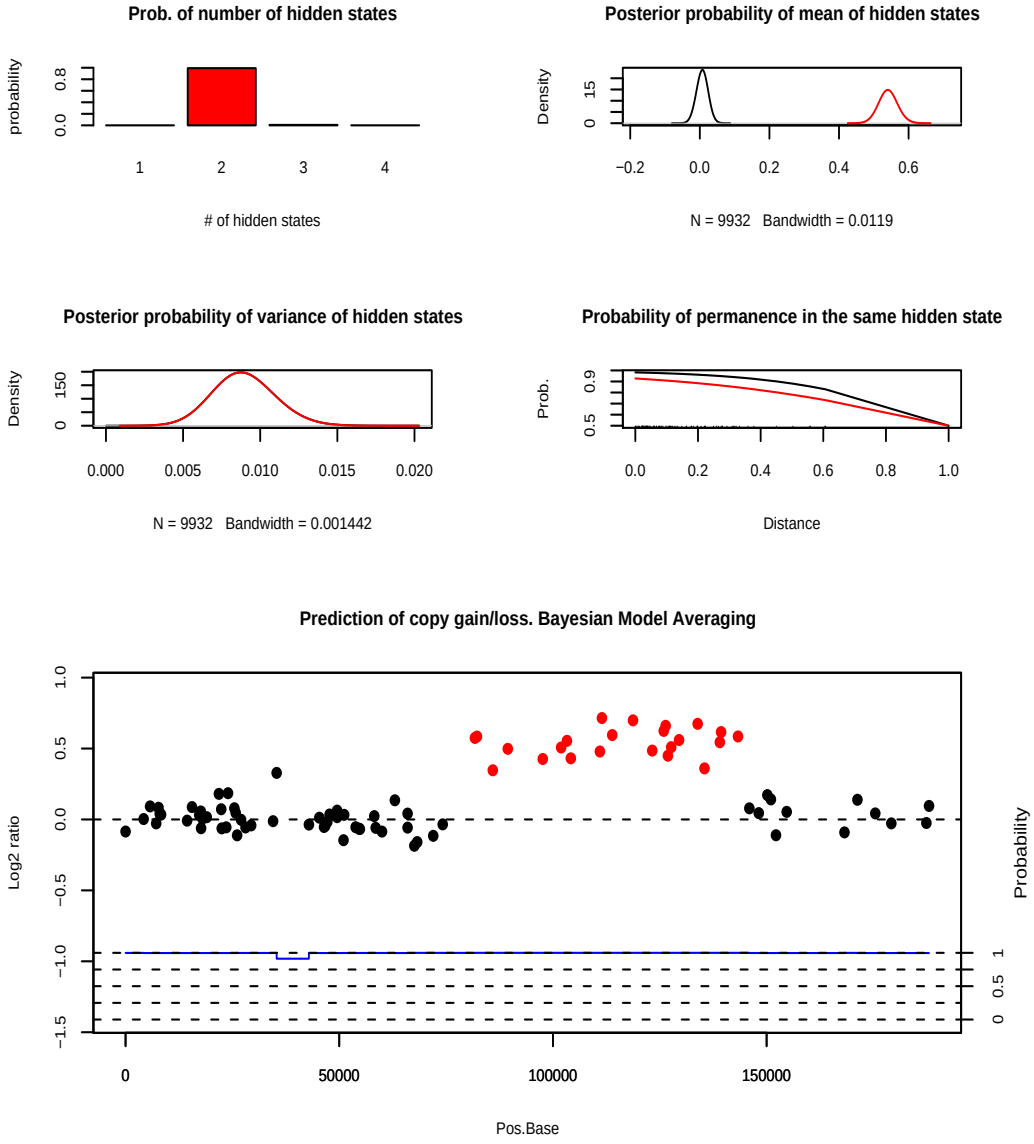
These methods can be also used on a fit with the same model on the whole genome, as the one we fit in the last section.

We can also plot the whole genome or just a chromosome:

```
> plot(fit.chrom)
```

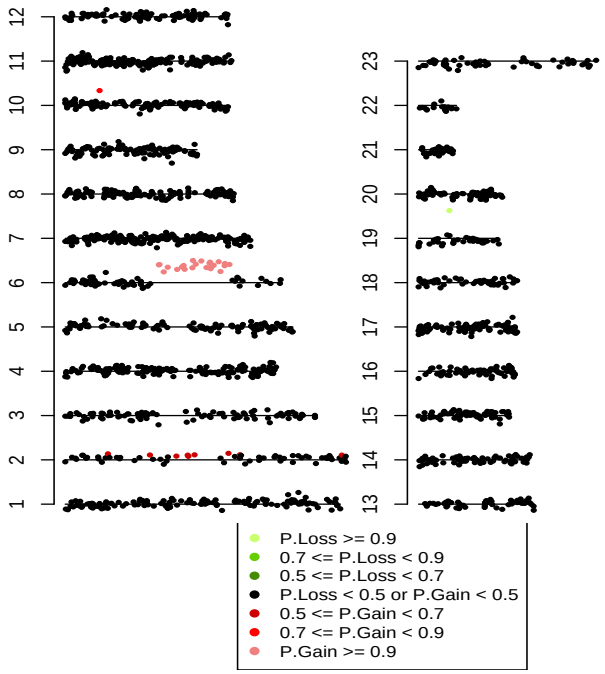



```
> plot(fit.chrom[[6]])
```



Finally, we can also see the probabilities of alteration in a graph chromosome by chromosome:

```
> genome.plot(fit.chrom)
```



3.3 Fitting several arrays

We can also fit at the same time several arrays (if they have the same probes spotted in the same positions). RJaCGH fits a different model to each of them:

```
> gm07081LR <- gm07081$LogRatio
> gm10315LR <- gm10315$LogRatio
> gm07408LR <- gm07408$LogRatio
> not.NA <- !is.na(gm07081LR) & !is.na(gm10315LR) & !is.na(gm07408LR)
> gm07081LR <- gm07081LR[not.NA]
> gm10315LR <- gm10315LR[not.NA]
> gm07408LR <- gm07408LR[not.NA]
> Pos3 <- gm07081$PosBase[not.NA]
> Chrom3 <- gm07081$Chromosome[not.NA]
> fit.arrays <- RJaCGH(y = cbind(gm07081LR, gm10315LR, gm07408LR),
+     Pos = Pos3, Chrom = Chrom3, model = "genome", k.max = 4,
+     burnin = 50000, TOT = 10000)
```

The returned object follows the same structure (nested lists); now every object is a list with the result of the fit to each array:

```
> summary(fit.arrays)

Summary for array  gm07081LR :

Distribution of the number of hidden states:
```

```
1 2 3 4
0 0 0 1

Model with  4  states:

Distribution of the posterior means of hidden states:
          10%    25%    50%    75%    90%
Normal -0.139 -0.139 -0.139 -0.120 -0.117
Normal -0.004 -0.003 -0.003 -0.002  0.000
Gain-1  0.218  0.226  0.226  0.245  0.268
Gain-2  0.479  0.482  0.499  0.499  0.504

Distribution of the posterior variances of hidden states:
         10%   25%   50%   75%   90%
Normal 0.005 0.005 0.005 0.005 0.005
Normal 0.005 0.005 0.005 0.005 0.005
Gain-1 0.005 0.005 0.005 0.005 0.005
Gain-2 0.005 0.005 0.005 0.005 0.005

Parameters of the transition functions:
       Normal Normal Gain-1 Gain-2
Normal  0.000  0.012  3.993  2.787
Normal  5.493  0.000  5.755  6.511
Gain-1  0.471  0.792  0.000  0.348
Gain-2  3.652  1.853  2.479  0.000

================================================

Summary for array  gm10315LR :

Distribution of the number of hidden states:

1 2 3 4
0 0 1 0

Model with  3  states:

Distribution of the posterior means of hidden states:
          10%    25%    50%    75%    90%
Normal -0.012 -0.011 -0.009 -0.007 -0.005
Normal  0.071  0.083  0.093  0.109  0.118
Gain-1  0.582  0.592  0.599  0.605  0.609

Distribution of the posterior variances of hidden states:
         10%   25%   50%   75%   90%
Normal 0.007 0.007 0.007 0.007 0.007
Normal 0.007 0.007 0.007 0.007 0.007
Gain-1 0.007 0.007 0.007 0.007 0.007
```

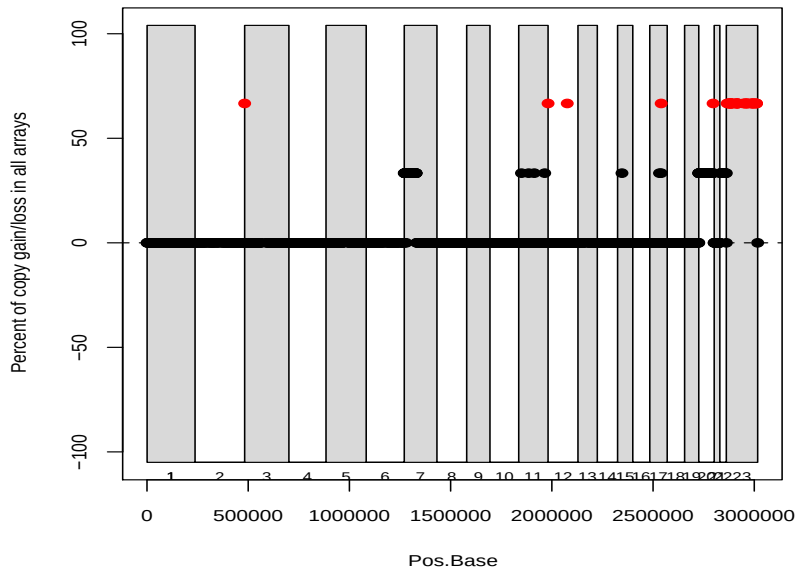
```
Parameters of the transition functions:
       Normal Normal Gain-1
Normal  0.000  5.880  8.948
Normal  1.507  0.000  3.271
Gain-1  6.419  3.056  0.000

================================================

Summary for array  gm07408LR :

Distribution of the number of hidden states:

1 2 3 4
0 0 0 1

Model with  4  states:

Distribution of the posterior means of hidden states:
          10%    25%    50%    75%    90%
Normal -0.007 -0.005 -0.004 -0.004 -0.003
Gain-1  0.437  0.449  0.454  0.454  0.462
Gain-2  0.606  0.625  0.635  0.658  0.661
Gain-3  0.877  0.909  0.918  0.957  0.990

Distribution of the posterior variances of hidden states:
         10%   25%   50%   75%   90%
Normal 0.004 0.004 0.004 0.004 0.004
Gain-1 0.004 0.004 0.004 0.004 0.004
Gain-2 0.004 0.004 0.004 0.004 0.004
Gain-3 0.004 0.004 0.004 0.004 0.004

Parameters of the transition functions:
       Normal Gain-1 Gain-2 Gain-3
Normal  0.000  6.093  8.480  8.312
Gain-1  2.736  0.000  1.985  3.507
Gain-2  2.008  0.001  0.000  0.732
Gain-3  0.238  0.023  0.069  0.000

================================================

> summary(fit.arrays[["gm07081LR"]])$mu

                10%          25%          50%          75%           90%
Normal -0.138791017 -0.138791017 -0.138791017 -0.119720784 -1.170062e-01
Normal -0.003748636 -0.003103135 -0.003103135 -0.001638606 -5.572882e-05
Gain-1  0.217765922  0.225574230  0.225574230  0.245239875  2.677448e-01
Gain-2  0.478619524  0.482270506  0.499055357  0.499055357  5.039175e-01
```

So we can apply the same methods used in previous sections to the whole set of arrays, to a given array (or to a given chromosome of a given array if we fit a different model to every chromosome for each array).

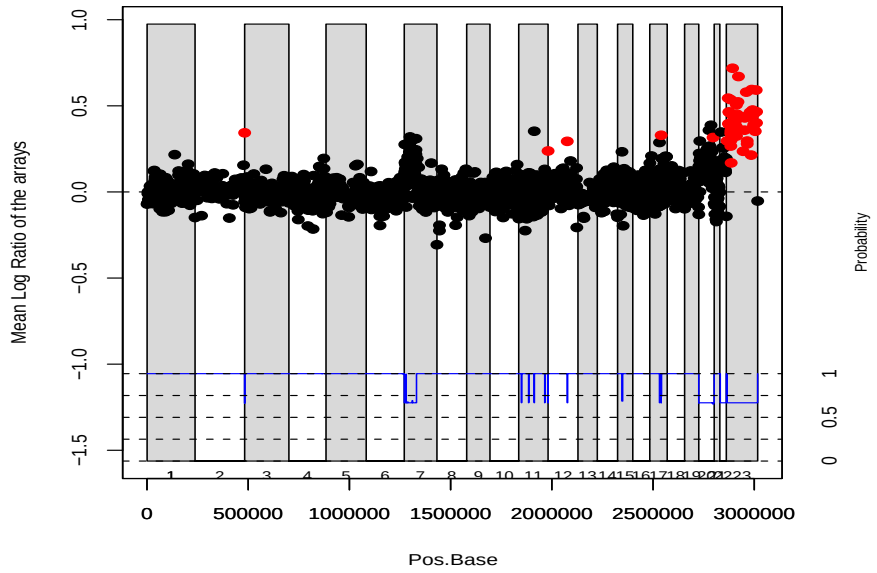
We may want to plot the fit for all the arrays. We can do it in two different ways:

- Plot for every probe the percentage of arrays which have that probe marginally altered.
- Plot for every probe the average probability on all arrays.

```
> plot(fit.arrays, show = "frequency")
```




```
> plot(fit.arrays, show = "average")
```



We can also compare the classification of genes with the true states of Snijders:

```
> seq.states <- model.averaging(fit.arrays[["gm07081LR"]])$states
> table(seq.states, gm07081$Statut[not.NA])
```

```
seq.states Normal Trisomy
    Loss        0       0
    Normal   1878       1
    Gain       11      67
```

3.4 Probabilistic Common Regions

RJaCGH can also compute probabilistic common regions. Note that these regions are different to other approaches, because they take into account the precision or variability inherent to the estimation of the true copy number for every probe on every array considered. There are two different methods:

- `pREC_A` returns regions common to the whole set of arrays with a joint probability of alteration as high as a given threshold.
- `pREC_S` returns regions shared by a subset of arrays (of size as high as a given threshold) with a joint probability within each array as high as a given threshold.

`pREC_A` detects regions common for most of the arrays. It has three arguments, `p` for the minimum probability to call a region altered, `alteration` for the type of alteration ('Gain' or 'Loss') and `array.weights` for the weight that we want to give to each array (by default, it is the same for all of them).

We can find common regions for the three arrays from the last section:

```
> Regions.Gain <- pREC_A(fit.arrays, p = 0.33, alteration = "Gain")
> Regions.Loss <- pREC_A(fit.arrays, p = 0.33, alteration = "Loss")
```

```
> Regions.Gain
```

```
   Chromosome  Start    End Probes Prob. Gain
1           2 245000 245000      1  0.6617995
2           7      0   6868     13  0.3333333
3           7   9696  17181      3  0.3333333
4           7  18019  38319     25  0.3333333
5           7  40773  57971     24  0.3333333
6          11  13646  13646      1  0.3333333
7          11  48923  48923      1  0.3333333
8          11  76848  76848      1  0.3333333
9          11 128440 128440      1  0.3333333
10         12      0      0      1  0.6635330
11         12  94805  94805      1  0.6666667
12         17  48088  48088      1  0.3333333
13         17  56276  56276      1  0.6666667
14         20      0  73000     85  0.3333333
15         22   3258  33000     14  0.3333333
16         23   4000 149342     45  0.6666667
```

```
> Regions.Loss
```

```
[1] "No common minimal regions found"
```

If we want to make this results into a data.frame we would do:

```
> RG <- as.data.frame(print(Regions.Gain))
```

`pREC_S` is useful to detect subset of arrays that share common alterations. It has the arguments `p` and `alteration` but also a `freq.array` that sets the minimum number of arrays that can form a region.

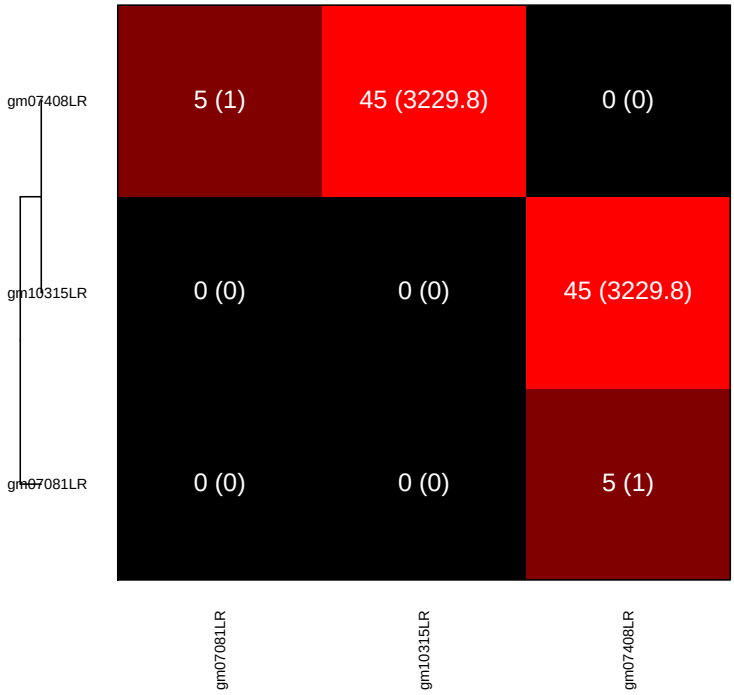
```
> Regions <- pREC_S(fit.arrays, p = 0.75, alteration = "Gain",
+     freq.array = 2)
```

```
> Regions
```

```
Common regions of Gain of at least 0.75 probability:
  Chromosome  Start    End Probes              Arrays
1          2 245000 245000      1 gm07081LR;gm07408LR
2         12      0      0      1 gm07081LR;gm07408LR
3         12  94805  94805      1 gm07081LR;gm07408LR
4         17  56276  56276      1 gm07081LR;gm07408LR
5         20  70647  70647      1 gm07081LR;gm07408LR
6         23   4000 149342     45 gm10315LR;gm07408LR
```

The result of plotting this object is an image plot that shows for each pair of arrays the number of alterations shared and their mean lengths. Besides, a hierarchical clustering based in that measure is performed and the arrays reordered:

```
> plot(Regions, cex.axis = 0.6)

$probes
           Var2
Var1        gm07081LR gm10315LR gm07408LR
  gm07081LR         0         0         5
  gm10315LR         0         0        45
  gm07408LR         5        45         0

$length
           Var2
Var1        gm07081LR gm10315LR gm07408LR
  gm07081LR         0     0.000     1.000
  gm10315LR         0     0.000  3229.844
  gm07408LR         1  3229.844     0.000
```

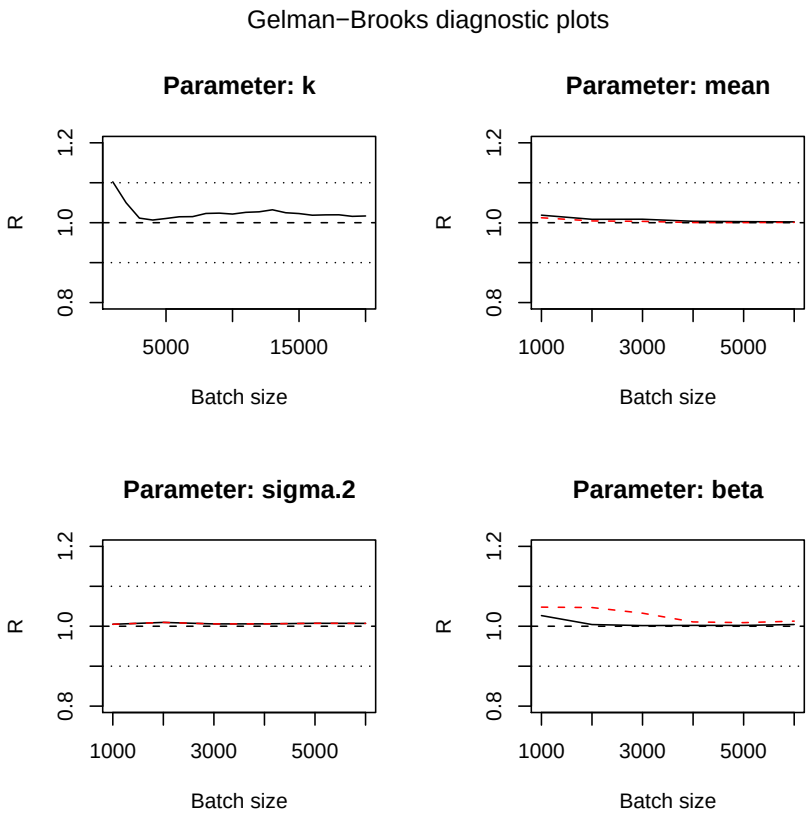


3.5 Checking convergence

We have seen the function `trace.plot` to check convergence in a given model, but the best way to be sure that RJaCGH has converged is to run several parallel

chains and draw the Gelman-Brooks convergence plot. In this example, we use data from cell line gm01524, but only from chromosome 1:

```
> fit <- list()
> for (i in 1:4) {
+     fit[[i]] <- RJaCGH(y = y2[Chrom2 == 1], Pos = Pos2[Chrom2 ==
+         1], k.max = 4, burnin = 50000, TOT = 10000, jump.parameters = jump.parameters)
+ }
> gelman.brooks.plot(fit)
```



We should check that the lines converge to zero, or at least that remain under 1.1. The values that return the function should be under 1.1, too. The results are satisfactory, so we can join the four chains into one:

```
> fit <- collapseChain(fit)
```

And use the former methods to the object `fit`.

Other useful function is `chainsSelect`, which deletes 'outliers' chains. See help file for details.

References

[1] Brooks, S.P. and Gelman, A. (1998). *"General Methods for Monitoring convergence of iterative simulations"*. Journal of Computational and Graphical Statistics. p434-455.

[2] Cappé, Moulines and Rydén. (2005) *"Inference in Hidden Markov Models"*. Springer.

[3] Green, P.J. (1995) *"Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination"*. Biometrika, 82, 711-732.

[4] Rueda OM, Diaz-Uriarte R. (2007) *"Flexible and Accurate Detection of Genomic Copy-Number Changes from aCGH"*. PLoS Comput Biol, 3(6):e122

[5] Snijders, M. J. et al. (2001) *"Assembly of microarrays for genome-wide measurement of DNA copy number"*. Nature Genetics 29, pp 263 - 264.

[6] Huppé, P. (2005). *"GLAD: Gain and Loss Analysis of DNA"*. R package version 1.6.0. http://bioinfo.curie.fr