QTL Mapping using Diversity Outbred Mice

Daniel M. Gatti

08 October 2013

1 Introduction

Quantitative Trait Locus (QTL) mapping in DO mice is performed in several steps. First, we use the founder haplotype contributions to perform linkage mapping. In the mapping model, we adjust for kinship between DO mice using the R package QTLRel. Then, we perform permutations to determine and empirical signficance threshold. Next, we select chromosomes with QTL peaks above the signficance threshold, examine the founder allele effects and determine support intervals. Finally, we impute the founder SNPs onto the DO genomes to perform association mapping in the QTL intervals.

2 Mapping Models

2.1 Linkage Mapping

Linkage mapping involves the use of founder haplotype probabilities. We perform point mapping at each marker on the array. We fit an additive model that regresses the phenotype on the eight founder haplotype contributions and incorporates an adjustment for the kinship between samples.

$$y = X\alpha + H\beta + Zu + \varepsilon \tag{1}$$

where:

- n is the number of samples
- y is an n x 1 vector of phenotype values for each sample
- X is an n x p matrix of p fixed covariates (sex, diet, etc.)
- α is a p x 1 vector of fixed effects
- H is an n x 8 matrix of founder haplotype contributions (each row sums to 1)
- β is an 8 x 1 vector of founder haplotype effects
- Z is an n x n matrix of error covariances between samples
- u is an n x 1 vector of ???
- ε is an n x 1 vector of residual errors

2.2 Association Mapping

Between each pair of markers, we assign the genotype state with the highest probability to each DO sample. We then query the Sanger Mouse Genomes SNP file to obtain all of the founder SNPs in the interval.

For each Sanger SNP, we impute the Sanger SNPs onto DO genomes as follows:

$$a_j = \sum_{i=1}^{8} s_i h_{ij} \tag{2}$$

where:

- a is the allele call (coded as 0, 1 or 2) for sample j
- s is the Sanger founder allele call (coded as 0 or 1)
- h is the founder haplotype contribution of founder i for sample j

$$y = X\alpha + A\beta + Zu + \varepsilon \tag{3}$$

where:

- n is the number of samples
- y is an n x 1 vector of phenotype values for each sample
- X is an n x p matrix of p fixed covariates (sex, diet, etc.)
- α is a p x 1 vector of fixed effects
- A is an n x 3 matrix of imputed allele calls
- β is an 3 x 1 vector of allele effects
- Z is an n x n matrix of error covariances between samples
- u is an n x 1 vector of ???
- ε is an n x 1 vector of residual errors

3 QTL Mapping

We will use example data from Svenson et.al, *Genetics*, 2012. Breifly, 149 mice (75 F, 74 M) were placed on either a chow (n = 100) or a high fat diet (n = 49). A variety of clinical phenotypes were measured at two time points, roughly 14 weeks apart. In this example, we will map the hemoglobin distribution width (HDW) at the second time point. We will load this data from the Bioconductor data package `MUGAExampleData`.

```
> library(DOQTL)
> library(MUGAExampleData)
> data(pheno)
> data(model.probs)
```

QTL mapping requires phenotype and genotype data. Here, we have a `data.frame` of phenotypes called `pheno` and a 3D array of founder haplotype contributions (num.samples x 8 founders x num.markers) called `model.probs`. The sample IDs must be in rownames(pheno) and dimnames(model.probs)[[1]] and they must match each other. We will map the hemoglobin distribution width at time point 2 (HDW2).

First, we need to create a kinship matrix using the founder contributions.

```
> K = kinship.probs(model.probs)
```

Second, we need to create a matrix of additive covariates to run in the model. In this case, we will use sex, diet and CHOL1. Note that the sample IDs must be in `rownames(covar)`.

```
> covar = data.frame(sex = as.numeric(pheno$Sex == "M"), diet = as.numeric(pheno$Diet == "hf"))
> rownames(covar) = rownames(pheno)
```

Third, we need to get the marker locations on the array.

```
> load(url("ftp://ftp.jax.org/MUGA/muga_snps.Rdata"))
```

Fourth, we map the phenotype using `scanone`.

```
> qtl = scanone(pheno = pheno, pheno.col = "HDW2", probs = model.probs, K = K,
+               addcovar = covar, snps = muga_snps)

[1] "Mapping with 141 samples."
[1] "Mapping with 7654 markers."
[1] "HDW2"
```

Fifth, we run permutations to determine significane thresholds. We recommend running at least 1,000 permutations. In this demo, we run 100 permutations to save time.

```
> perms = scanone.perm(pheno = pheno, pheno.col = "HDW2", probs = model.probs,
+         addcovar = covar, snps = muga_snps, nperm = 100)
> thr = quantile(perms, probs = 0.95)
```

We then plot the LOD curve for the QTL.

```
> plot(qtl, sig.thr = thr, main = "HDW2")
```

The largest peak appears on Chr 9. The linkage mapping model (Eqn. 1) produces an estimate of the effect of each founder allele at each marker. We can plot these effects (model coefficients) on Chr 9 to see which founders contribute to a high HDW.

```
> coefplot(qtl, chr = 9)
```

Note that the DO mice with alleles from three strains, 129S1/SvImJ, NZO/HlLtJ and WSB/EiJ, have lower changes in cholesterol than the other five strains. Remember these strains because they will appear again below. We then determine the width of the QTL support interval using `bayesint`. Note that this function only provides reasonable support intervals if there is a single QTL on the chromosome.

```
> interval = bayesint(qtl, chr = 9)
> interval
```

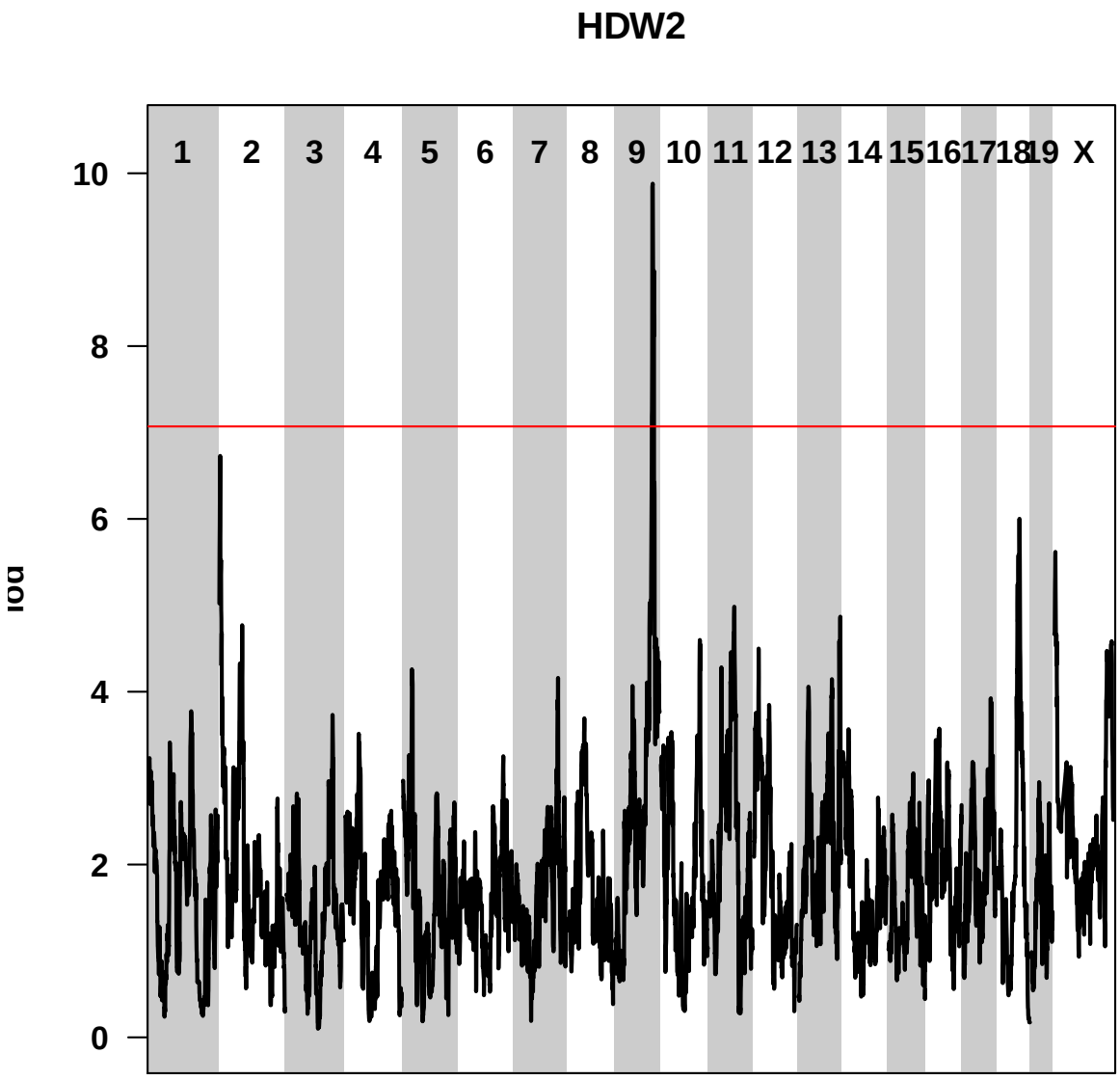

Figure 1: QTL plot of HDW2. The LOD of the mode in Eqn. 1 is plotted along the mouse genome. The red line is the $p < 0.05$ significance threshold.

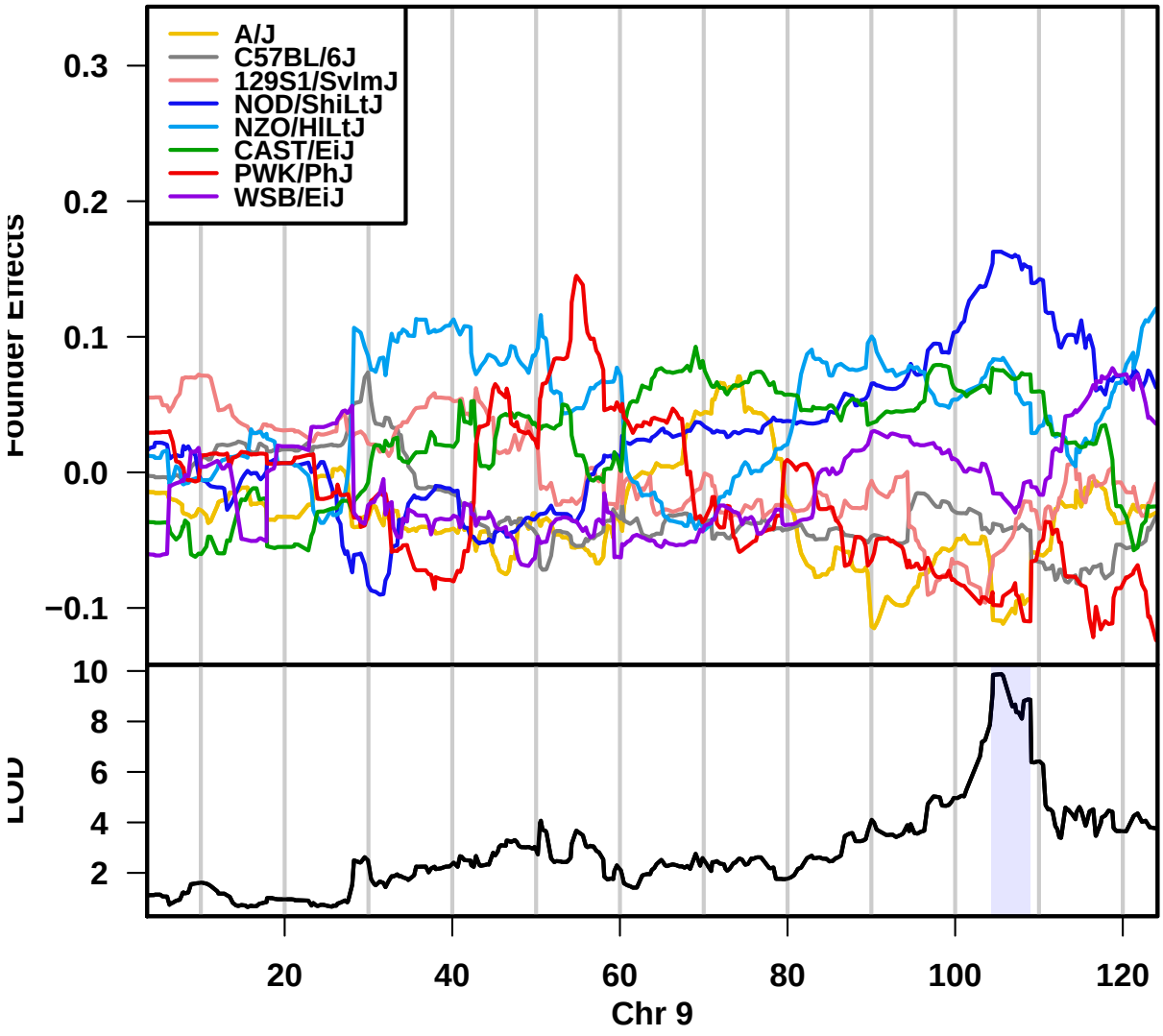

Figure 2: Coefficient plot of HDW2 on Chr 9. The top panel shows the 8 estimated founder allele effects along Chr 9. The NOD/ShiLtJ allele contributes to high values and the A/J and PWK/PhJ alleles contribute to low values. The bottom panel shows the LOD score.

```
           SNP_ID Chr Mb_NCBI38      cM perc.var      lrs      lod
1            <NA>   9  104.3100 56.6120 25.16321 39.16096 8.503694
4105 UNC091160886   9  105.5128 56.7432 28.60966 45.49603 9.879338
3            <NA>   9  108.9811 59.6310 25.69325 40.11901 8.711732
                        p      neg.log10.p
1                    <NA>                9
4105 1.09550863036677e-07 6.96038419673778
3                    <NA>                9
```

The QTL support interval is 4.7 Mb wide. Finally, we narrow the candidate gene list by imputing the founder SNPs onto the DO genomes. This idea is essentially assocation mapping in an outbred population.

```
> ma = assoc.map(pheno = pheno, pheno.col = "HDW2", probs = model.probs, K = K,
+                      addcovar = covar, snps = muga_snps, chr = interval[1,2],
+                      start = interval[1,3], end = interval[3,3])

[1] "Mapping with 135 samples."
[1] "Retrieving SNPs..."
found header lines for 3 'fixed' fields: ALT, QUAL, FILTER
found header lines for 4 'info' fields: INDEL, DP, DP4, CSQ
found header lines for 2 'geno' fields: GT, FI
[1] "Calculating mapping statistic..."
Warning: solution lies close to zero for some positive variance components, their standard errors may no

> tmp = assoc.plot(ma, thr = 4)
> unique(tmp$sdps)

NULL
```

We can get the genes in the QTL interval using the `get.mgi.features()` function.

```
> mgi = get.mgi.features(chr = interval[1,2], start = interval[1,3],
+       end = interval[3,3], type = "gene", source = "MGI")
> nrow(mgi)

[1] 168

> head(mgi)

    seqid source type     start      stop score strand phase              ID
1       9    MGI gene 104288240 104337728     .      -     . MGI:MGI:1928480
143     9    MGI gene 104426113 104426187     .      +     . MGI:MGI:4358922
150     9    MGI gene 104481368 104481510     .      -     . MGI:MGI:5455681
154     9    MGI gene 104547286 105034544     .      +     . MGI:MGI:1921270
313     9    MGI gene 104994916 104995019     .      +     . MGI:MGI:5453168
387     9    MGI gene 105053239 105079888     .      +     . MGI:MGI:2137204
       Name Parent
1      Acpp     NA
143 Mir2136     NA
150 Gm25904     NA
```

```
[1] "Mapping with 135 samples."
[1] "Retrieving SNPs..."
found header lines for 3 'fixed' fields: ALT, QUAL, FILTER
found header lines for 4 'info' fields: INDEL, DP, DP4, CSQ
found header lines for 2 'geno' fields: GT, FI
[1] "Calculating mapping statistic..."
Warning: solution lies close to zero for some positive variance components, their standard errors may no

NULL
```

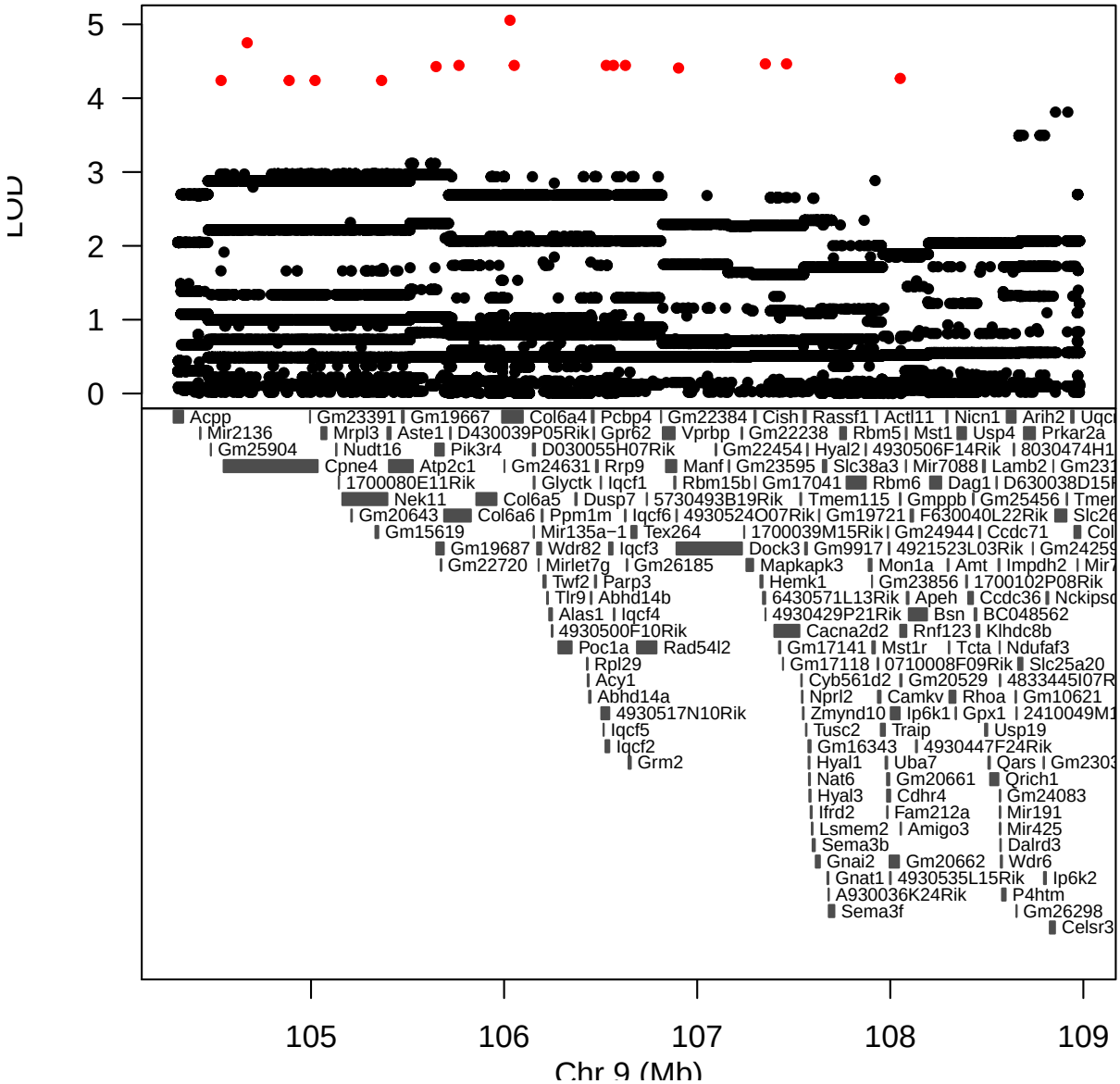

Figure 3: Association mapping plot of HDW2 in the Chr 9 support interval. The top panel shows the LOD score from assocation mapping (Eqn. 3) in the QTL support interval. The bottom panel shows the genes and non-coding RNAs from the Mouse Genome Informatics database.

```
154   Cpne4     NA
313 Gm23391     NA
387   Mrpl3     NA
                                                                      Dbxref
1   VEGA:OTTMUSG00000024988,NCBI_Gene:56318,ENSEMBL:ENSMUSG00000032561
143                     NCBI_Gene:100316725,ENSEMBL:ENSMUSG00000089406
150                                         ENSEMBL:ENSMUSG00000089116
154 VEGA:OTTMUSG00000023466,NCBI_Gene:74020,ENSEMBL:ENSMUSG00000032564
313                                         ENSEMBL:ENSMUSG00000088204
387 VEGA:OTTMUSG00000023521,NCBI_Gene:94062,ENSEMBL:ENSMUSG00000032563
                              mgiName               bioType
1         acid phosphatase%2c prostate protein coding gene\r
143                      microRNA 2136          miRNA gene\r
150             predicted gene%2c 25904         snoRNA gene\r
154                          copine IV protein coding gene\r
313             predicted gene%2c 23391          miRNA gene\r
387 mitochondrial ribosomal protein L3 protein coding gene\r
```

There are 169 genes in the QTL support interval. Several SNPs have LOD scores above 4. This is a somewhat arbitrary cutoff and an appropriate threshold will be supplied in future version of DOQTL. In this case, there may be more than one variant that influences the phenotype.

4 SessionInfo

```
> sessionInfo()

R version 3.2.1 (2015-06-18)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] MUGAExampleData_0.102.0             DOQTL_1.3.7
 [3] VariantAnnotation_1.14.13           Rsamtools_1.20.4
 [5] BSgenome.Rnorvegicus.UCSC.rn6_1.4.1 BSgenome.Mmusculus.UCSC.mm10_1.4.0
 [7] BSgenome_1.36.3                     rtracklayer_1.28.10
 [9] Biostrings_2.36.4                   XVector_0.8.0
[11] GenomicRanges_1.20.6                GenomeInfoDb_1.4.2
[13] IRanges_2.2.7                       S4Vectors_0.6.5
[15] BiocGenerics_0.14.0
```

```
loaded via a namespace (and not attached):
 [1] gtools_3.5.0            modeltools_0.2-21       kernlab_0.9-22
 [4] lattice_0.20-33         rhdf5_2.12.0            GenomicFeatures_1.20.4
 [7] XML_3.98-1.3            DBI_0.3.1               prabclus_2.2-6
[10] BiocParallel_1.2.21     lambda.r_1.1.7          fpc_2.1-10
[13] foreach_1.4.2           robustbase_0.92-5       zlibbioc_1.14.0
[16] futile.logger_1.4.1     hwriter_1.3.2           mvtnorm_1.0-3
[19] codetools_0.2-14        Biobase_2.28.0          biomaRt_2.24.0
[22] doParallel_1.0.8        RUnit_0.4.29            flexmix_2.3-13
[25] class_7.3-14            AnnotationDbi_1.30.1    DEoptimR_1.0-3
[28] trimcluster_0.1-2       xtable_1.7-4            corpcor_1.6.8
[31] diptest_0.75-7          gdata_2.17.0            annotate_1.46.1
[34] annotationTools_1.42.0  grid_3.2.1              tools_3.2.1
[37] bitops_1.0-6            regress_1.3-14          RCurl_1.95-4.7
[40] RSQLite_1.0.0           cluster_2.0.3           futile.options_1.0.0
[43] MASS_7.3-44             QTLRel_0.2-14           iterators_1.0.7
[46] mclust_5.0.2            GenomicAlignments_1.4.1 nnet_7.3-11
```