VanillaICE: Hidden Markov Models for the Assessment of Chromosomal Alterations using High-throughput SNP Arrays

Robert Scharpf

March 11, 2012

Abstract

This package provides an implementation of a hidden Markov Model for high throughput SNP arrays. Users of this package should already have available locus-level estimates of copy number and B allele frequencies or genotype calls. Copy number estimates can be relative or absolute.

1 Overview

This vignette requires that you have

- an absolute estimate of the *total* copy number organized such that rows correspond to loci and columns correspond to samples

  and / or

- a matrix of B allele frequencies or genotype calls (1=AA, 2 = AB, 3= BB): rows correspond to loci and columns correspond to samples

If Affymetrix CEL files or Illumina Idat files were processed with the R package *crlmm*, see the vignette `crlmmDownstream` included with this package.
Other HMM implementations for the joint analysis of copy number and genotype include QuantiSNP [1] and PennCNV [3].

Data considerations. The HMM implemented in this package is most relevant for heritable diseases for which integer copy numbers are expected. For somatic cell diseases such as cancer, we suggest circular binary segmentation, as implemented in the R package DNAcopy [2].

Citing this software. Robert B Scharpf, Giovanni Parmigiani, Jonathan Pevsner, and Ingo Ruczinski. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Annals of Applied Statistics*, 2(2):687–713, 2008.

2 Organizing the locus-level data

This package includes simulated genotype and copy number data for 9165 SNPs.

```
> library(VanillaICE)
> library(SNPchip)
> data(locusLevelData)
```

The copy number estimates in the `locusLevelData` object were multiplied by 100 and saved as an integer. An object of class `oligoSnpSet` is instantiated from the simulated data in the next code chunk. Note that the assay data elements for the `oligoSnpSet` object should be integers. The `integerMatrix` scales the copy number matrix by a factor of 100 (by default) and returns a matrix of integers. When available, B allele frequencies should be scaled by a factor of 1000. For example,
integerMatrix(bAlleleFreq, scale=1000) \verb

```
> cn <- log2(locusLevelData[["copynumber"]]/100)
> oligoSet <- new("oligoSnpSet",
+                 copyNumber=integerMatrix(cn),
+                 call=locusLevelData[["genotypes"]],
+                 callProbability=locusLevelData[["crlmmConfidence"]],
+                 annotation=locusLevelData[["platform"]])
> oligoSet <- oligoSet[!is.na(chromosome(oligoSet)), ]
> oligoSet <- oligoSet[chromosome(oligoSet) < 3, ]
```

3 Fitting the HMM

When jointly modeling the copy number and genotypes/allele frequencies, we assume that the genotype and copy number estimates are independent conditional on the underlying hidden state. The method `hmm` is defined for several classes, including objects of class `oligoSnpSet` and `BeadStudioSet`.
The viterbi algorithm is used to obtain the most likely sequence of hidden states given the observed data. The value returned is an object of class `RangedDataHMM` with genomic coordinates of the normal and altered regions. We also return the log-likelihood ratio (LLR) of the predicted sequence in an interval versus the likelihood of diploid copy number. For intervals inferred to have diploid copy number, the LLR is zero.

```
> oligoSet <- oligoSet[chromosome(oligoSet) <= 2, ]
> fit.van <- hmm(oligoSet)
```

Figure 1 plots the data for chromosome 1 and the predictions from the hidden markov model:

To find which markers are included in each genomic interval returned by the `hmm` method, one can use the `findOverlaps` method in the *oligoClasses*. For example, the second interval in the `RangedDataHMM` object `fit.van` contains 102 markers.

```
> fit.van[2, ]
```

```
RangedDataHMM with 1 row and 7 value columns across 1 space
     space               ranges |     chrom  num.mark          id
  <factor>            <IRanges> | <integer> <integer> <character>
1        1 [49825223, 54637167] |         1       102     NA06993
  start.index end.index     state       LLR
    <integer> <integer> <integer> <numeric>
1         997      1098         4  14.84152
```

To find the names of the 102 markers that are included in this genomic interval, one could do the following

```
> mm <- as.matrix(findOverlaps(fit.van, oligoSet))
> markersInRange <- featureNames(oligoSet)[mm[,1]==2]
```

Multipanel displays can be useful for visualizing the low-level data for copy number alterations. We extend the `xyplot` method in the R package lattice for two common use-cases: by-locus and by-sample.

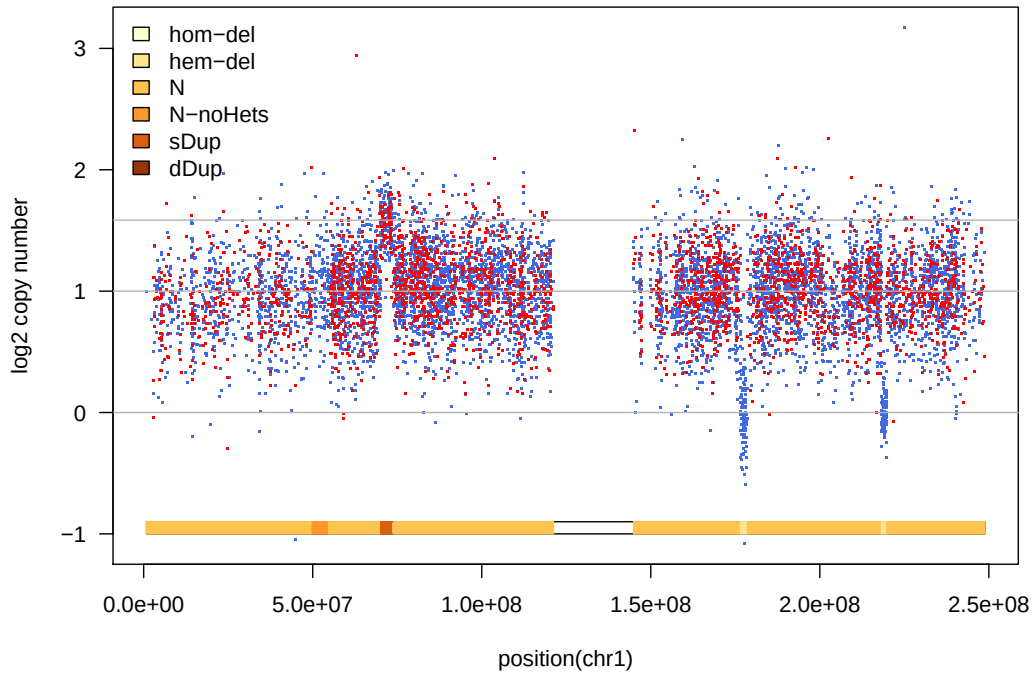


Figure 1: Plot of artificial data for one chromosome.

By locus. To plot the genomic data for a set of ranges at a given locus, we provide an unevaluated code chunk below (our example contains only a single sample):

```
> xyplot(cn ~ x | id, oligoSet, range=RangedDataObject, ylim=c(-0.5,4), panel=xypanel)
```

Note that the default in the above command is to use a common x- and y-scale for each panel. To allow the x-axes to change for each panel, one could set the x-scales to 'free':

```
> xyplot(cn ~ x | id, oligoSet, range=RangedDataObject, ylim=c(-0.5,4), scales=list(x="free"),
+        panel=xypanel)
```

The function `xypanel` provides default colors for annotating the plotting symbols by genotype and by whether the markers are polymorphic. The `RangedDataHMM` must be passed by the name *range* to the `xyplot` method.

By sample. To plot the low-level data for multiple alterations occurring in a single sample, one can again pass a `RangedDataHMM` object with name *range* to the `xyplot`. For example, see Figure 2. The code for producing Figure 2 is in the following code chunk. Note that the formula in the example below conditions on *range* instead of *id*. The conditioning variable for displaying multiple panels of a single sample must be called 'range'. We plot a 2 Mb window surrounding each of the alterations in the simulated data by specifying `frame=2e6`.

```
> ranges.altered <- fit.van[!state(fit.van) %in% c(3,4) & coverage2(fit.van) > 5, ]
> xy.example <- xyplot(cn ~ x | range, oligoSet, range=ranges.altered, frame=2e6,
```

```
+                       scales=list(x="free"), ylim=c(-1,3),
+                       panel=xypanel, cex.pch=0.4, pch=21, border="grey",
+                       ylab=expression(log[2]("copy number")))
```

See `?xyplot` for additional details.

Note also that `scales` must be set to *free* in the above call to `xyplot`.

3.1 ICE HMM

This will likely be phased out in favor of using B allele frequencies instead of genotype calls / call probabilities.

To compute emission probabilities that incorporate the *crlmm* genotype confidence scores, we set `ICE` to `TRUE`. This option is limited too a few platforms and the Affy 100k platforms is not one of them.

```
> VanillaICE:::icePlatforms()

[1] "pd.genomewidesnp.6"
[2] "genomewidesnp6"
[3] "genomewidesnp6Crlmm"
[4] "pd.mapping250k.nsp"
[5] "pd.mapping250k.sty"
[6] "pd.mapping250k.nsp, pd.mapping250k.sty"
```

For illustration, we assign 'genomewidesnp6' to the annotation slot since this platform is supported.

```
> ann <- annotation(oligoSet)
> annotation(oligoSet) <- "genomewidesnp6"
> fit.ice <- hmm(oligoSet, ICE=TRUE)
> fit.ice

RangedDataHMM with 17 rows and 7 value columns across 1 space
      space                 ranges |     chrom  num.mark          id
   <factor>              <IRanges> | <integer> <integer> <character>
1         1 [   846864,  49772452] |         1       996     NA06993
2         1 [ 49825223,  54637167] |         1       102     NA06993
3         1 [ 54726362,  70065480] |         1       902     NA06993
4         1 [ 70081878,  73401801] |         1       204     NA06993
5         1 [ 73804712, 121226982] |         1      2497     NA06993
6         1 [144908589, 176547669] |         1      1294     NA06993
7         1 [176548473, 178437444] |         1        98     NA06993
8         1 [178533776, 218173294] |         1      1901     NA06993
9         1 [218219379, 219806187] |         1        99     NA06993
10        1 [219825554, 248794371] |         1      1065     NA06993
11        1 [   170616,   8409113] |         2         4     NA06993
12        1 [ 15224340,  68466618] |         2        29     NA06993
13        1 [ 75654059,  79296379] |         2         3     NA06993
14        1 [ 81048974,  88467934] |         2         4     NA06993
15        1 [102998279, 105880487] |         2         2     NA06993
16        1 [114471782, 115837414] |         2         2     NA06993
17        1 [123126479, 240767758] |         2        56     NA06993
   start.index end.index     state         LLR
     <integer> <integer> <integer>   <numeric>
1            1       996         3   0.0000000
```

```
2          997      1098        4  14.8415228
3         1099      2000        3   0.0000000
4         2001      2204        5 195.1615962
5         2205      4701        3   0.0000000
6            1      1294        3   0.0000000
7         1295      1392        2 166.9786379
8         1393      3293        3   0.0000000
9         3294      3392        2 205.2171294
10        3393      4457        3   0.0000000
11           1         4        5  -0.5191900
12           5        33        3   0.0000000
13          34        36        5  -2.4122910
14          37        40        2  -0.4449246
15          41        42        5   1.3314298
16          43        44        4   0.4059383
17           1        56        3   0.0000000
```

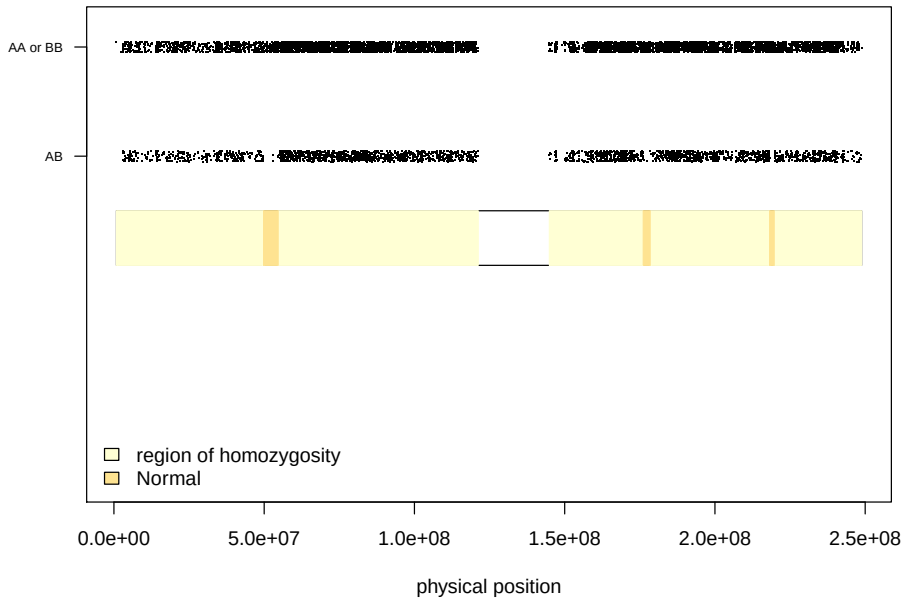
```
> annotation(oligoSet) <- ann
```

3.2 Other options

Regions of homozygosity. A HMM for genotype-only data can be used to find long stretches of homozygosity. Note that hemizygous deletions are also identified as 'ROH' when copy number is ignored (as the biallelic genotypte call in a hemizygous deletions tends to be all homozygous calls).

```
> snpSet <- as(oligoSet, "SnpSet")
> fit.gt <- hmm(snpSet, prGtHom=c(0.7, 0.99), normalIndex=1L, S=2L)
```

A suggested visualization:



4 Session Information

The version number of R and packages loaded for generating the vignette were:

- R version 2.15.0 alpha (2012-03-10 r58666), `x86_64-apple-darwin11.3.0`
- Locale:
  `en_US.US-ASCII/en_US.US-ASCII/en_US.US-ASCII/C/en_US.US-ASCII/en_US.US-ASCII`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: Biobase 2.15.4, BiocGenerics 0.1.14, codetools 0.2-8, DBI 0.2-5, foreach 1.3.2, IRanges 1.13.30, iterators 1.0.5, oligo 1.19.14, oligoClasses 1.17.30, pd.mapping50k.hind240 1.4.0, pd.mapping50k.xba240 1.4.0, RColorBrewer 1.0-5, RSQLite 0.11.1, SNPchip 2.0.1, VanillaICE 1.17.20
- Loaded via a namespace (and not attached): affxparser 1.27.4, affyio 1.23.2, annotate 1.33.3, AnnotationDbi 1.17.27, Biostrings 2.23.6, bit 1.1-8, compiler 2.15.0, crlmm 1.13.10, ellipse 0.3-5, ff 2.2-5, genefilter 1.37.0, grid 2.15.0, lattice 0.20-6, msm 1.1, mvtnorm 0.9-9992, preprocessCore 1.17.5, splines 2.15.0, stats4 2.15.0, survival 2.36-12, tools 2.15.0, xtable 1.7-0, zlibbioc 1.1.1

References

[1] Stefano Colella, Christopher Yau, Jennifer M Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S Bassett, Anneke Seller, Christopher C Holmes, and Jiannis Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*, 35(6):2013–2025, 2007.

[2] Adam B Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–72, Oct 2004.

[3] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F A Grant, Hakon Hakonarson, and Maja Bucan. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Res*, 17(11):1665–1674, Nov 2007.

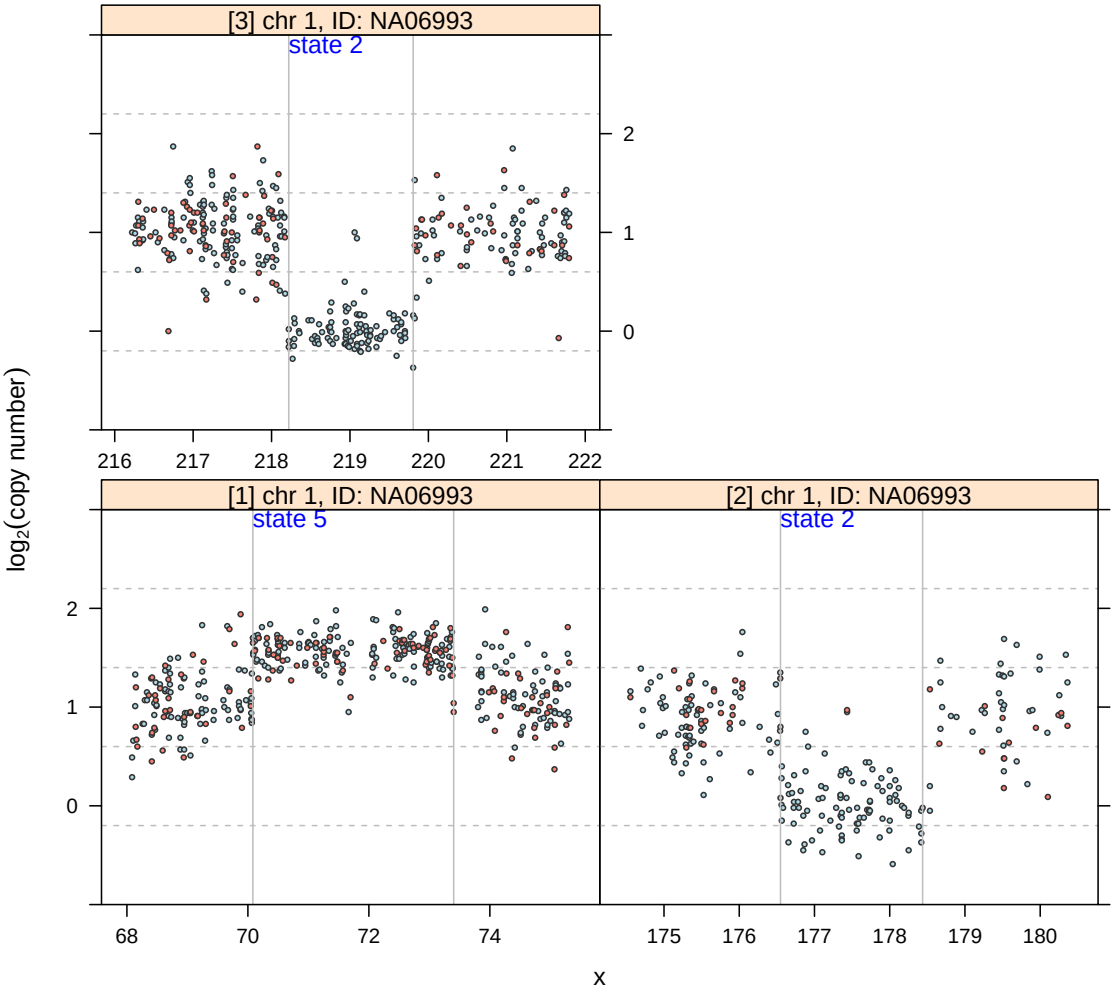


Figure 2: The method `xyplot` is used to create a multi-panel display of alterations in a single sample. Each panel displays a single copy number alteration detected by the HMM that is boxed by a rectangle. The alteration is framed by specifying the number of basepairs to plot upstream and downstream of the alteration. Here, we used a frame of 2 Mb. Homozygous SNPs with diallelic geotypes 'AA' and 'BB' are shaded blue; SNPs with diallelic genotype call 'AB' are shaded in red.