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

Robert Scharpf and Ingo Ruczinski

October 1, 2007

1 Introduction

Chromosomal DNA is characterized by variation between individuals at the level of entire chromosomes (e.g. aneuploidy in which the chromosome copy number is altered), segmental changes (including insertions, deletions, inversions, and translocations), and changes to small genomic regions (including single nucleotide polymorphisms). A variety of alterations that occur in chromosomal DNA, many of which can be detected using high density single nucleotide polymorphism (SNP) microarrays, are linked to normal variation as well as disease and therefore of particular interest. These include changes in copy number (deletions and duplications) and genotype (e.g. the occurrence of regions of homozygosity). Hidden Markov models (HMM) are particularly useful for detecting such abnormalities, modeling the spatial dependence between neighboring SNPs. Here, we extend previous approaches that utilize HMM frameworks for inference in high throughput SNP arrays by integrating copy number, genotype calls, and the corresponding measures of uncertainty when available. Using simulated and real data, we demonstrate how confidence scores control smoothing in a probabilistic framework. The goal of this vignette is to provide a simple interface for fitting HMMs and plotting functions to help visualize the predicted states alongside the experimental data.

2 Simple Usage

```
> library(VanillaICE)
> data(chromosome1)
> class(chromosome1)
> annotation(chromosome1)
```

See the documentation pages in the R package *VanillaICE* for more information about the `chromosome1` example dataset. The vanilla HMM can be fit to both arms of this simulated chromosome as follows:

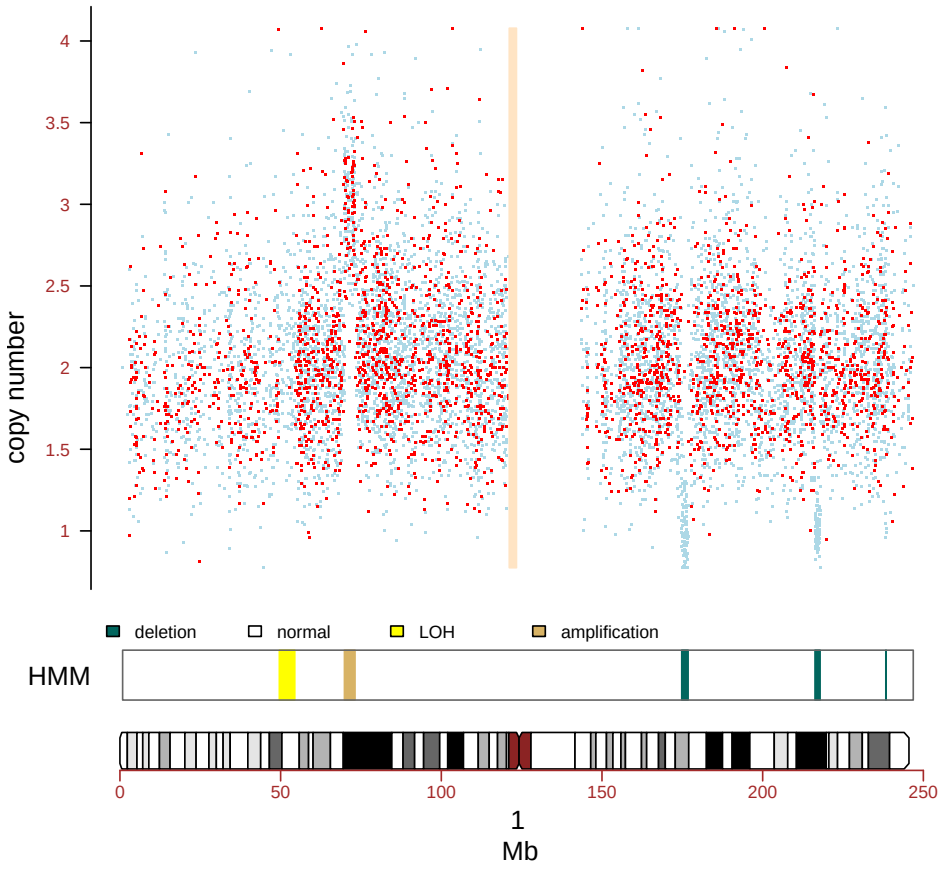
```
> copyNumberIce(chromosome1) <- FALSE
> callsIce(chromosome1) <- FALSE
> fit1 <- hmm(chromosome1)

[1] "Fitting HMM to chromosome 1"
[1] "   processing  p  arm of sample NA06993  ..."
[1] "   processing  q  arm of sample NA06993  ..."
```

Graphical parameters for plotting the results are easily obtained and manipulated:

```
> graph.par <- getPar(fit1)
> class(graph.par)

[1] "ParHmmSnpSet"
attr(,"package")
[1] "VanillaICE"
```

The predicted states are plotted beneath the observed data by

```
> plotSnp(graph.par, fit1)

$rect
$rect$w
[1] 91180098

$rect$h
[1] 0.5862942

$rect$left
[1] -9004244

$rect$top
[1] 1.8


$text
$text$x
[1] -2037711 20446308 42930326 65414345

$text$y
[1] 1.506853 1.506853 1.506853 1.506853
```



See [2] for a more complete description of the simulated dataset and the features detected by this HMM.

3 Integrating Confidence Estimates (ICE)

In this section, we illustrate how one may fit an HMM that incorporates confidence esimates of the SNP-level summaries for genotype calls and copy number.

3.1 Confidence scores for genotype calls

We suggest using the `CRLMM` algorithm [1] for genotype calls. `CRLMM` (in the R package *oligo*) provides confidence scores ($S_{\widehat{GT}}$) of the genotype estimates ($\widehat{GT}$). From 269 HapMap samples assayed on the Affymetrix 50k Xba and Hind chips, we have a gold standard of the true genotype defined by the consensus of the HapMap centers. We use kernal based density estimates to obtain

$$f\left\{ S_{\widehat{HOM}} \mid \widehat{HOM}, HOM \right\},\ f\left\{ S_{\widehat{HOM}} \mid \widehat{HOM}, HET \right\},\ f\left\{ S_{\widehat{HET}} \mid \widehat{HET}, HOM \right\},\quad \text{and } f\left\{ S_{\widehat{HET}} \mid \widehat{HET}, HET \right\} \tag{1}$$

separately for the Xba and Hind 50k chips. The first term in (1), for example, denotes the density of the scores when the genotype is correctly called homozygous ($\widehat{HOM}$) and the true genotype is homozygous

(HOM). See [2] for a more complete description of the methods. The data needed to estimate these densities is stored in the experiment data package *callsConfidence*. *callsConfidence* is available from the author's website.

3.2 Confidence scores for copy number estimates

To illustrate how standard errors of the copy number estimate could be integrated in the HMM, the R object `chromosome1` contains standard errors simulated from a shifted Gamma: $\text{Gamma}(1, 2) + 0.3$, where 1 is the shape parameter and 2 is the rate parameter. To ascertain the effect of qualitatively high confidence scores on the ICE HMM, we scaled a robust estimate of the copy number standard deviation by $\frac{1}{2}$. Similarly, to simulate less precise $\widehat{\text{CN}}$ we scaled ϵ by 2. For more detailed information about how the data in the `chromosome1` was generated, see the documentation for this object in the R package *VanillaICE*.

3.3 Fitting the ICE HMM

The results from fitting the ICE HMM to the `chromosome1` data have been stored in the assay data element `predictions` of `chromosome1`. To see a table of the predicted underlying states (excluding normal segments) sorted by size (MB):

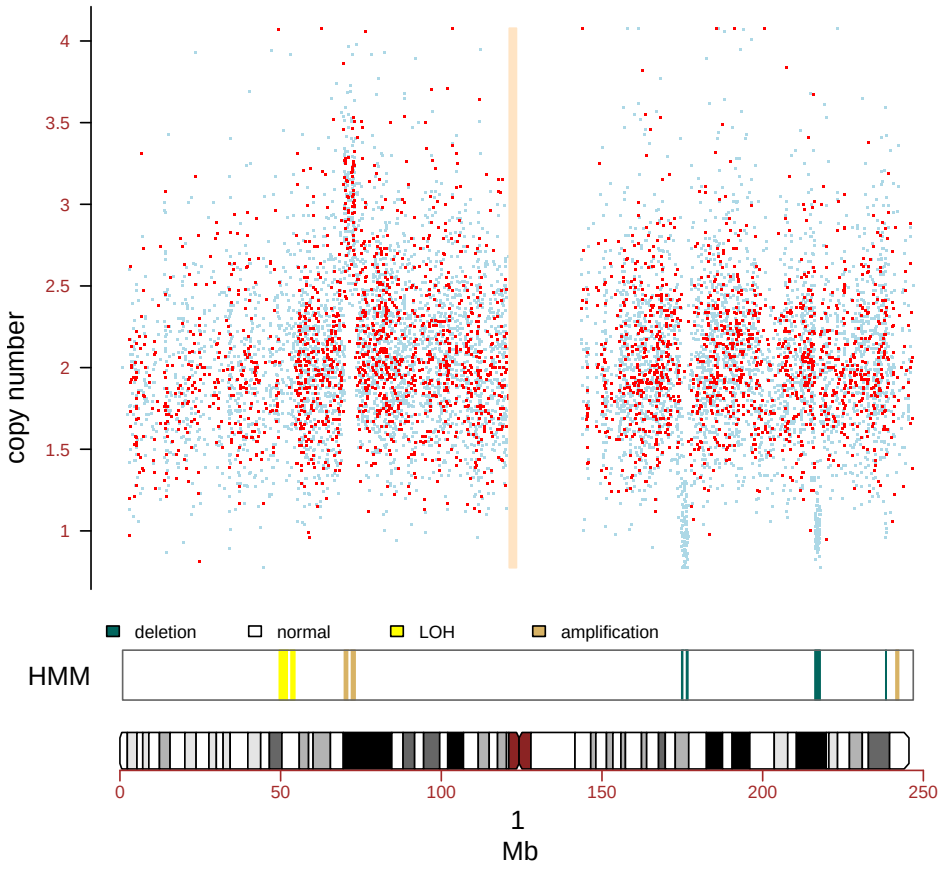
```
> breakpoints(chromosome1)
```

```
   sampleId chromosome previous_SNP     start      last  next_SNP
2   NA06993          1     49545039  49597810  52590776  52669983
4   NA06993          1     52728046  52820181  54409755  54498950
14  NA06993          1    216239917 216286002 217872810 217892177
6   NA06993          1     69838068  69854466  71342708  71406383
8   NA06993          1     71474047  71826917  73174389  73577300
12  NA06993          1    175700199 175726310 176704067 176800399
18  NA06993          1    241076215 241483148 242295322 242374397
10  NA06993          1    174815096 174828535 175630040 175683520
16  NA06993          1    238302668 238319943 238417864 238429864
         MB   hiddenState   cytoband numberSnps CHET is.na
2  2.992966           LOH p33, p32.3         51    0     0
4  1.589574           LOH      p32.3         47    0     0
14 1.586808      deletion        q41         99    0     0
6  1.488242 amplification      p31.1         97   29     0
8  1.347472 amplification      p31.1        103   35     0
12 0.977757      deletion      q25.2         47    0     0
18 0.812174 amplification        q44          7    2     0
10 0.801505      deletion      q25.2         46    0     0
16 0.097921      deletion        q43          5    0     0
```

To reproduce these results, one must install the Experiment Data package *callsConfidence* and run the command:

```
> library(callsConfidence)
> copyNumberIce(chromosome1) <- TRUE
> callsIce(chromosome1) <- TRUE
> chromosome1 <- hmm(chromosome1)

> graph.par <- getPar(chromosome1)
```

```
> plotSnp(graph.par, chromosome1)

$rect
$rect$w
[1] 91180098

$rect$h
[1] 0.5862942

$rect$left
[1] -9004244

$rect$top
[1] 1.8


$text
$text$x
[1] -2037711 20446308 42930326 65414345

$text$y
[1] 1.506853 1.506853 1.506853 1.506853
```



Note that the ICE HMM correctly identifies the simulated normal segments in features B and C. Additionally, the ICE HMM detects the micro-amplification in region E.

4 HMM for copy number alterations

The method `hmm` has a different set of underlying hidden states depending on whether copy number estimates, genotype calls, or both are available. When only copy number estimates are available, the hidden states (for autosomes) are hemizygous or homozygous deletion (one or fewer copies), normal (two copies), and amplification (three or more copies). The corresponding class is `HmmSnpCopyNumberSet`. To illustrate, we convert the `chromosome1` example to an object of this class and fit the HMM.

```
> chr1.cn <- as(chromosome1, "HmmSnpCopyNumberSet")
> locationCopyNumber(chr1.cn) <- c(log2(1), log2(2), log2(3))
> initialStateProbability(chr1.cn) <- c((1 - 0.99)/2,
+     0.99, (1 - 0.99)/2)
> scaleCopyNumber(chr1.cn) <- NULL
> stateNames(chr1.cn) <- c("deletion", "normal", "amplification")
> predictions(chr1.cn) <- matrix(NA, nrow(chr1.cn), ncol(chr1.cn))
> chr1.cn <- hmm(chr1.cn, verbose = TRUE)
```

```
[1] "Fitting HMM to chromosome 1"
[1] "   processing  p  arm of sample NA06993  ..."
[1] "   processing  q  arm of sample NA06993  ..."
```

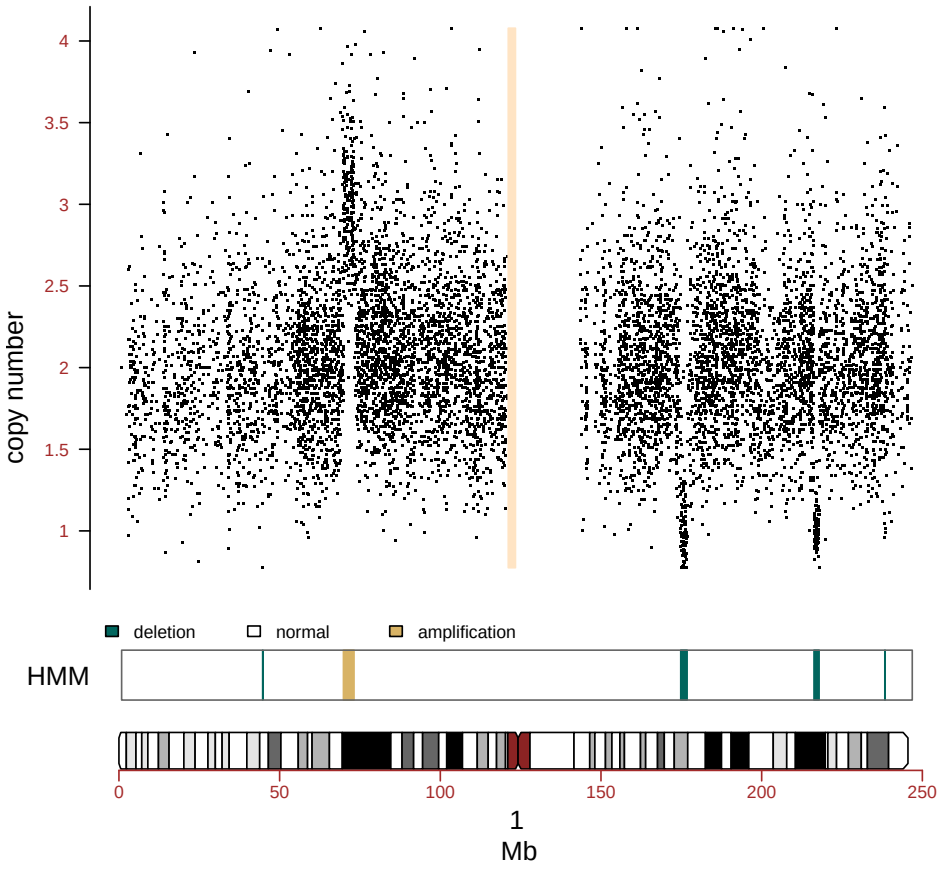
```
> breakpoints(chr1.cn)
```

```
   sampleId chromosome previous_SNP     start      last  next_SNP
4   NA06993          1     69838068  69854466  73153900  73174070
6   NA06993          1    174814292 174815096 176800844 176926556
8   NA06993          1    216239917 216286002 217872810 217892177
10  NA06993          1    238302668 238319943 238417864 238429864
2   NA06993          1     44702452  44722116  44762242  44779351
         MB   hiddenState cytoband
4  3.299434 amplification    p31.1
6  1.985748      deletion    q25.2
8  1.586808      deletion      q41
10 0.097921      deletion      q43
2  0.040126      deletion    p34.1
```

```
> graph.par <- getPar(chr1.cn)
```

```
> plotSnp(graph.par, chr1.cn)
```

```
$rect
$rect$w
[1] 68696079

$rect$h
[1] 0.5862942

$rect$left
[1] -9004244

$rect$top
[1] 1.8


$text
$text$x
[1] -2037711 20446308 42930326

$text$y
[1] 1.506853 1.506853 1.506853
```



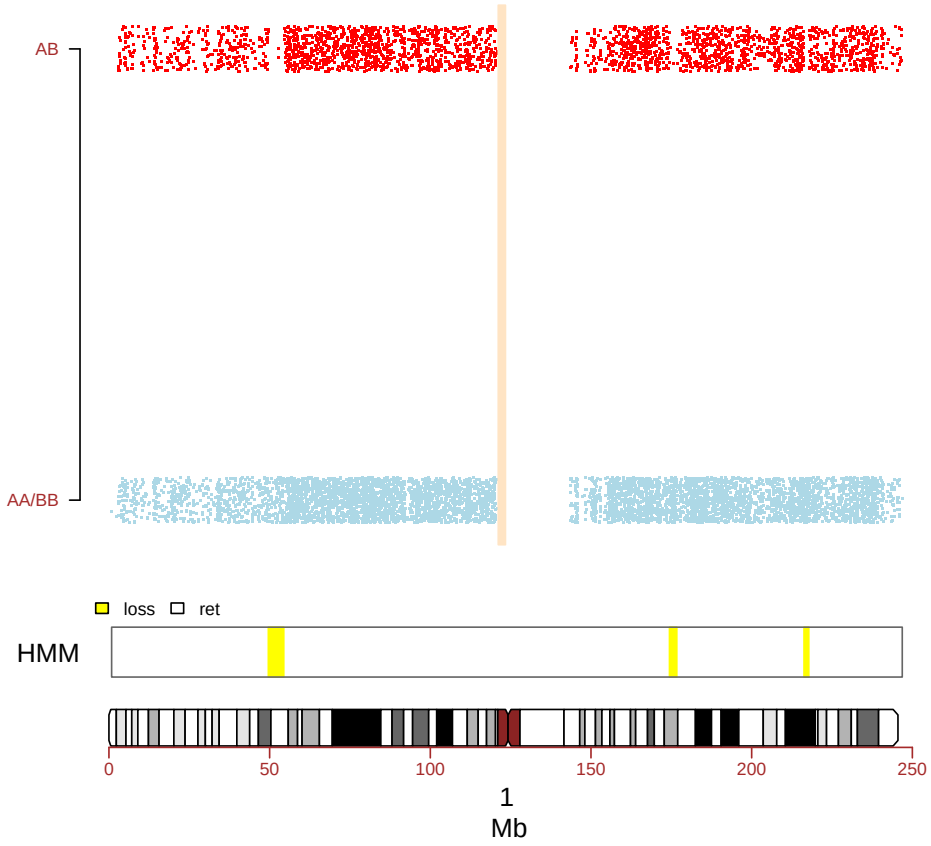
5 HMM for loss of heterozygosity

When only genotype calls are available, the hidden states are loss and retention (ret) of heterozygosity. We define *loss* to be a sequence of homozygous SNPs longer than what we would expect to observe by chance. Note that many long stretches of homozygosity may occur as a result of a population sharing a common underlying haplotype structure; loss predictions from an HMM fit to an indvidual do not necessarily reflect the 'loss' of an allele in that individual. For illustration, we convert the `chromosome1` example to an object of class `HmmSnpCallSet` and refit the HMM.

```
> chr1.calls <- as(chromosome1, "HmmSnpCallSet")
> stateNames(chr1.calls) <- c("loss", "ret")
> pCHOM(chr1.calls) <- c(0.999, 0.7)
> initialStateProbability(chr1.calls) <- c(1 - 0.99, 0.99)
> predictions(chr1.calls) <- matrix(NA, nrow(chr1.calls),
+     ncol(chr1.calls))
> chr1.calls <- hmm(chr1.calls)
```

```
[1] "Fitting HMM to chromosome 1"
[1] "   processing  p  arm of sample NA06993  ..."
[1] "   processing  q  arm of sample NA06993  ..."
```

```
> breakpoints(chr1.calls)

  sampleId chromosome previous_SNP     start      last  next_SNP
2  NA06993          1     49545039  49597810  54409755  54498950
4  NA06993          1    174309008 174418117 176704067 176800399
6  NA06993          1    216239917 216286002 217771828 217872013
        MB hiddenState   cytoband
2 4.811945        loss p33, p32.3
4 2.285950        loss      q25.2
6 1.485826        loss        q41

> graph.par <- getPar(chr1.calls)

> plotSnp(graph.par, chr1.calls)

$rect
$rect$w
[1] 25106940

$rect$h
[1] 0.5862942

$rect$left
[1] -9004244

$rect$top
[1] 1.8


$text
$text$x
[1] -2037711  9893748

$text$y
[1] 1.506853 1.506853
```



6 Classes for HMMs

More documentation about the classes can be found in the documentation for the R package *VanillaICE*. Note that the minimal SNP-level annotation required for the HMMs described here are physical position of the SNP (stored as an integer) and the chromosome (stored as a character string). The motivation for requiring that this information be stored in the `featureData` is primarily for convenience when subsetting instances of these classes, but may be relaxed in future releases. See, for instance, the `featureData` in the `chromosome1` object:

```
> fvarLabels(chromosome1)

[1] "dbsnp_rs_id"     "chrom"           "physical_pos"
[4] "strand"          "allele_a"        "allele_b"
[7] "fragment_length" "enzyme"
```

In addition to the class `HmmSnpSet` that contains SNP-level summaries of genotype and copy number, we define classes for genotype-only (`HmmSnpCallSet`) and copy number-only datasets (`HmmSnpCopyNumberSet`). The parameters of the model, including the latent states, transition probabilities, and emission probabilities, are dependent on the class of data. See [2] additional details.

7 Session Information

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

- R version 2.6.0 alpha (2007-09-05 r42788), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8;LC_NUMERIC=C;LC_TIME=en_US.UTF-8;LC_COLLATE=en_US.UTF-8;LC_MONETARY=en_US.UTF-8;LC_MESSAGES=en_US.UTF-8;LC_PAPER=en_US.UTF-8;LC_NAME=C;LC_ADDRESS=C;LC_T8;LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, stats, tools, utils
- Other packages: Biobase 1.15.33, oligoClasses 0.99.3, RColorBrewer 1.0-1, SNPchip 1.1.32, VanillaICE 0.99.7

References

[1] Benilton Carvalho, Henrik Bengtsson, Terence P Speed, and Rafael A Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, 8(2):485–499, Apr 2007.

[2] Robert B Scharpf, Giovanni Parmigiani, Jonathan Pevsner, and Ingo Ruczinski. A hidden Markov model for joint estimation of genotype and copy number in high-throughput SNP chips. Technical Report Working Paper 136, Johns Hopkins University, February 2007.