A hidden Markov model for SNP arrays processed with *crlmm*

Robert Scharpf

May 12, 2011

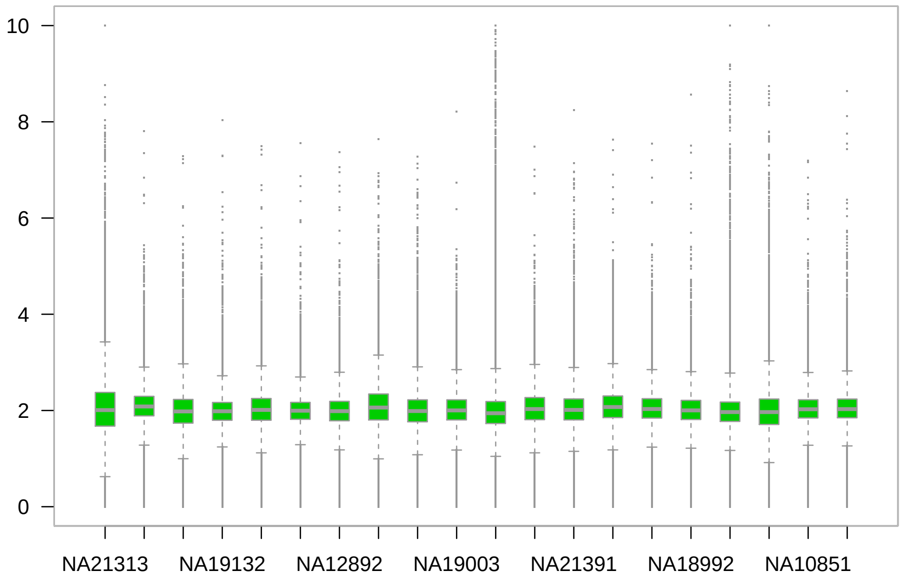
```
> library(VanillaICE)
> library(HapmapCrlmmAffySet)
> library(RColorBrewer)
> if (!exists("hapmapSet")) data(hapmapSet)
> class(hapmapSet)

[1] "CNSet"
attr(,"package")
[1] "oligoClasses"

> dim(hapmapSet)

Features  Samples
   96875      172

> NA
```

1 Small to moderate size datasets

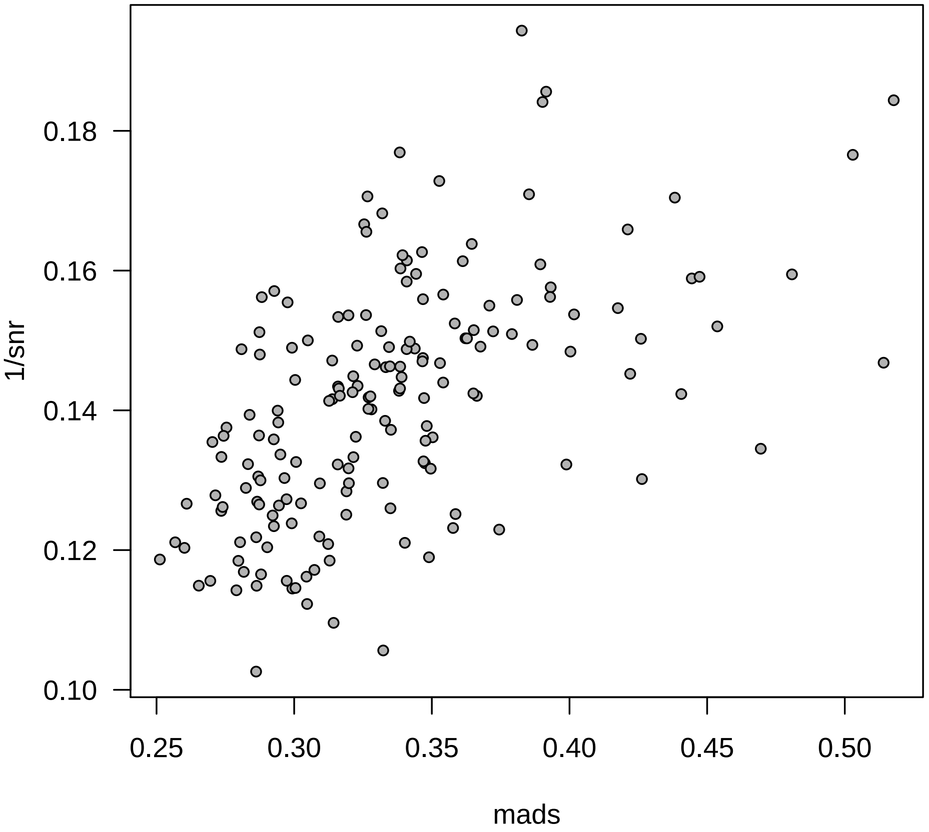
For smaller datasets (e.g, fewer than 100 samples), it may be preferable to coerce the object of class `CNSet` to a `oligoSnpSet` prior to fitting the HMM. The predictions from the HMM can then be visualized along side the marker-level estimates of copy number from *crlmm*.

Coercion of a `CNSet` object to an `oligoSnpSet` object is illustrated in the following code chunk. This coercion is not instantaneous as it may involve reading data from disk (if the `assayData` elements of the `CNSet` object are `ff`-derived objects), and computing the allele-specific estimates of copy number.

```
> oligoSet <- as(hapmapSet, "oligoSnpSet")
> NA

          used (Mb) gc trigger  (Mb)  max used   (Mb)
Ncells 1620497 86.6    2732238 146.0   2732238  146.0
Vcells  899429  6.9   90254485 688.6 275756432 2103.9

> sample.index <- sample(1:ncol(oligoSet), 20)
> par(las = 1)
> boxplot(data.frame(copyNumber(oligoSet)[, sample.index]), boxwex = 0.5,
+     col = "green3", pch = ".", border = "grey60", xaxt = "n")
> par(las = 0)
> axis(1, at = seq_along(sample.index), labels = sampleNames(oligoSet)[sample.index])
> box(col = "grey")
> NA
```



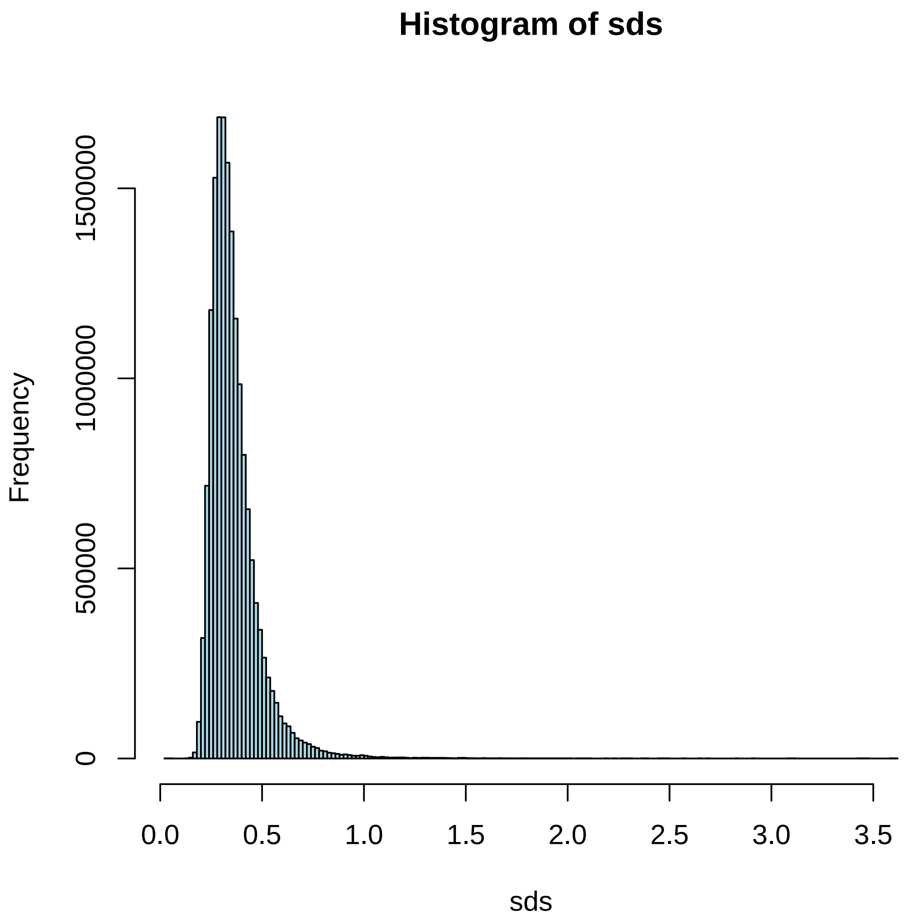
```
> snr <- oligoSet$SNR[]
> mads <- apply(copyNumber(oligoSet), 2, mad, na.rm = TRUE)
> par(las = 1)
> graphics:::plot(mads, 1/snr, pch = 21, cex = 0.8, bg = "grey70")
> abline(h = 1/5, lty = 2, col = "grey")
> NA
```



Note that 1/SNR is proportional to the sample-level MAD. Calculate the MAD for each marker and shrink to the sample-level MAD.

```
> sds <- VanillaICE:::robustSds2(copyNumber(oligoSet), DF.PRIOR = 10)
> NA
```

```
> hist(sds, col = "lightblue", breaks = 200)
> NA
```

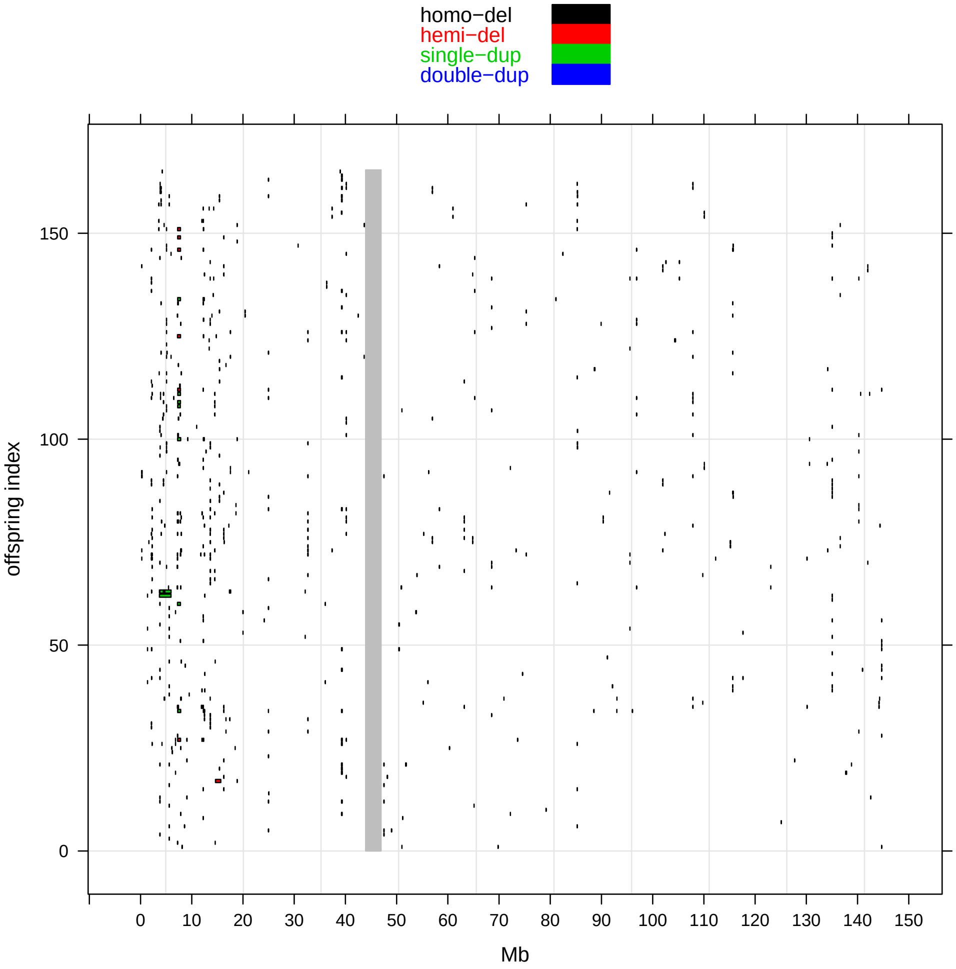


```
> cnConfidence(oligoSet) <- 1/sds
> NA

> suppressWarnings(rm(sds, mads, snr, sample.index))
> gc()

           used  (Mb) gc trigger  (Mb)  max used   (Mb)
Ncells  1701822  90.9    2732238 146.0   2732238  146.0
Vcells 51680471 394.3  127648481 973.9 275756432 2103.9

> NA

> oligoSet <- VanillaICE:::centerCopyNumber(oligoSet, at = 2)
> NA

> hmmOpts <- VanillaICE:::newHmmOptionList(object = oligoSet, verbose = 1L)
> NA

> fit <- hmm2(oligoSet, hmmOpts)
> NA

> fit <- as(fit, "RangedDataCn")
> NA
```

Visualizing the results:

```
> fig <- plot(fit, hmmOpts, show.coverage = FALSE)
> fig$x.scales[["tick.number"]] <- 20
> NA
```

```
> print(fig)
> NA
```



TODO: show how the color scheme could be modified.
TODO: plot method for low level data that uses locus zoom

2 Large datasets

Idea: define hmm2 method for CNSet. Coersion to oligoSnpSet inside of for loop. Center and compute sds as before. Extend to allow parallelization.

```
          used (Mb) gc trigger  (Mb)  max used   (Mb)
Ncells 1610051 86.0    2732238 146.0   2732238  146.0
Vcells  878051  6.7  102118784 779.2 275756432 2103.9
```

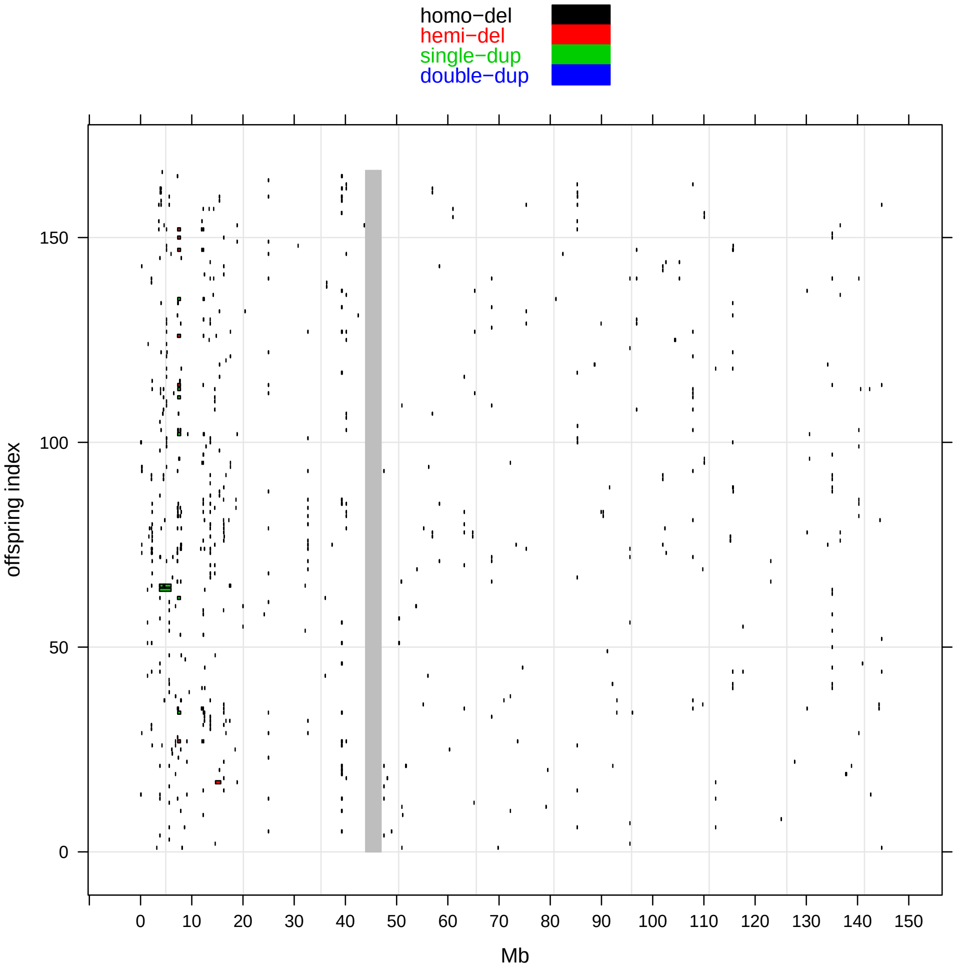
```
> if (!exists("hapmapSet")) data(hapmapSet)
> NA
> fit2 <- hmm2(hapmapSet, hmmOpts)
> NA
```

TODO: test that it works for ff objects.

```
> fit2 <- as(fit2, "RangedDataCn")
> NA
```

```
> fig <- plot(fit2, hmmOpts, show.coverage = FALSE)
> fig$x.scales[["tick.number"]] <- 20
> NA
```

```
> print(fig)
> NA
```



3 Session Information

```
> toLatex(sessionInfo())
```

- R version 2.14.0 Under development (unstable) (2011-05-12 r55861), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.iso885915`, `LC_NUMERIC=C`, `LC_TIME=en_US.iso885915`, `LC_COLLATE=en_US.iso885915`, `LC_MONETARY=en_US.iso885915`, `LC_MESSAGES=en_US.iso885915`,

`LC_PAPER=C`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.iso885915`, `LC_IDENTIFICATION=C`

- Base packages: base, datasets, graphics, grDevices, methods, stats, tools, utils
- Other packages: Biobase 2.13.2, cacheSweave 0.4-5, crlmm 1.11.2, filehash 2.1-1, HapmapCrlmmAffySet 0.0.1, IRanges 1.11.1, oligoClasses 1.15.5, RColorBrewer 1.0-2, SNPchip 1.17.0, stashR 0.3-3, VanillaICE 1.15.4
- Loaded via a namespace (and not attached): affyio 1.21.1, annotate 1.31.0, AnnotationDbi 1.15.2, Biostrings 2.21.1, bit 1.1-7, DBI 0.2-5, digest 0.4.2, ellipse 0.3-5, ff 2.2-2, genefilter 1.35.0, grid 2.14.0, lattice 0.19-26, mvtnorm 0.9-999, preprocessCore 1.15.0, RSQLite 0.9-4, splines 2.14.0, survival 2.36-9, xtable 1.5-6

```
> NA
```