

Integration with the *crlmm* package for copy number inference

Robert Scharpf

May 1, 2024

```
> library(oligoClasses)
> library(VanillaICE)
> library(crlmm)
> library(IRanges)
> library(foreach)
```

We load a portion of chromosome 8 from 2 HapMap samples that were processed using the *crlmm* package.

```
> data(cnSetExample, package="crlmm")
```

The data `cnSetExample` is an object of class `CNSet`. We coerce the `CNSet` object to a `SnpArrayExperiment` that contains information on copy number (log R ratios) and B allele frequencies.

```
> se <- as(cnSetExample, "SnpArrayExperiment")
```

Wave correction

To correct for genomic waves that correlate with GC content [refs], we use the R package *ArrayTV* – an approach adapted from the wave correction methods proposed by Benjamini and Speed for next generation sequencing platforms [1]. In the following code-chunk, we select a subset of the samples in the study to evaluate the genomic window for wave correction. See the *ArrayTV* vignette for details. For large datasets, one could randomly select 20 or 25 samples to compute the window, and then use a pre-selected window for wave correction on the remaining samples.

```
> library(ArrayTV)
> i <- seq_len(ncol(se))
> increms <- c(10,1000,100e3)
> wins <- c(100,10e3,1e6)
> res <- gcCorrect(lrr(se),
+                 increms=increms,
+                 maxwins=wins,
+                 returnOnlyTV=FALSE,
+                 verbose=TRUE,
+                 build="hg18",
+                 chr=chromosome(se),
+                 starts=start(se))
> se2 <- se
> assays(se2)[["cn"]] <- res$correctedVals

> ## TODO: correct for GC bias by loess
> se2 <- se
```

HMM

To identify CNVs, we fit a 6-state hidden markov model from estimates of the B allele frequency and log R ratios. A `hmm` method is defined for the `BafLrrSetList` class, and we apply the method directly with a few parameters that change the arguments from their default values. For example, the `TAUP` parameter scales the transition probability matrix. Larger values of `TAUP` makes it more expensive to transition from the normal copy number state to states with altered copy number.

```
> res <- hmm2(se2)
```

The object `res` can be filtered and putative CNVs can be visually inspected with the low-level summaries. Further details on such post-hoc analyses are provided in the section 'Inspecting, Filtering, and plotting HMM results' in the `VanillaICE` vignette.

Session Information

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

- R version 4.4.0 beta (2024-04-15 r86425), x86_64-pc-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_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Time zone: America/New_York
- TZcode source: system (glibc)
- Running under: Ubuntu 22.04.4 LTS
- Matrix products: default
- BLAS: /home/biocbuild/bbs-3.19-bioc/R/lib/libRblas.so
- LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
- Base packages: base, datasets, graphics, grDevices, methods, stats, stats4, utils
- Other packages: Biobase 2.64.0, BiocGenerics 0.50.0, crlmm 1.62.0, foreach 1.5.2, GenomeInfoDb 1.40.0, GenomicRanges 1.56.0, IRanges 2.38.0, MatrixGenerics 1.16.0, matrixStats 1.3.0, oligoClasses 1.66.0, preprocessCore 1.66.0, S4Vectors 0.42.0, SummarizedExperiment 1.34.0, VanillaICE 1.66.0
- Loaded via a namespace (and not attached): abind 1.4-5, affyio 1.74.0, askpass 1.2.0, base64 2.0.1, beanplot 1.3.1, BiocManager 1.30.22, Biostrings 2.72.0, bit 4.0.5, codetools 0.2-20, compiler 4.4.0, crayon 1.5.2, data.table 1.15.4, DBI 1.2.2, DelayedArray 0.30.0, ellipse 0.5.0, ff 4.0.12, GenomeInfoDbData 1.2.12, grid 4.4.0, httr 1.4.7, illuminaio 0.46.0, iterators 1.0.14, jsonlite 1.8.8, lattice 0.22-6, limma 3.60.0, Matrix 1.7-0, mvtnorm 1.2-4, openssl 2.1.2, parallel 4.4.0, R6 2.5.1, Rcpp 1.0.12, RcppEigen 0.3.4.0.0, S4Arrays 1.4.0, SparseArray 1.4.0, splines 4.4.0, statmod 1.5.0, tools 4.4.0, UCSC.utils 1.0.0, VGAM 1.1-10, XVector 0.44.0, zlibbioc 1.50.0

References

- [1] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*, 40(10):e72, May 2012.