

Genotyping with the `crlmm` Package

Benilton Carvalho

March, 2009

1 Quick intro to `crlmm`

The `crlmm` package contains a new implementation for the CRLMM algorithm (Carvalho et. al. 2007). Our focus is on efficient genotyping of SNP 5.0 and 6.0 Affymetrix arrays, although extensions of the method are under development for similar platforms.

This implementation, compared to the previous one (in `oligo`), offers improved confidence scores, quality scores for SNP's and batches, higher accuracy on different datasets and better performance.

Additionally, this package does not use the `pd.genomewidesnp` packages created via `pdInfoBuilder` for `oligo`. Instead, it uses different annotation packages (`genomewidesnp.5` and `genomewidesnp.6`), which use simple R objects to store only the information needed for genotyping. This allowed us to improve the speed of the method, as SQL queries are no longer performed here.

It is also our priority to make the package simple to use. Below we demonstrate how to get genotype calls with the 'new' CRLMM. We use 3 samples on SNP 5.0 made available via the `hapmapsnp5` package.

```
R> require(oligoClasses)
R> library(crlmm)
R> library(hapmapsnp6)
R> path <- system.file("celFiles", package="hapmapsnp6")
R> celFiles <- list.celfiles(path, full.names=TRUE)
R> system.time(crlmmResult <- crlmm(celFiles, verbose=FALSE))

   user  system elapsed
104.592   0.551  105.744
```

The `crlmmResult` is a *SnpSet* (see Biobase) object.

- **calls**: genotype calls (1 - AA; 2 - AB; 3 - BB);
- **confs**: confidence scores, which can be translated to probabilities by using:

$$1 - 2^{-(\text{confs}/1000)},$$

although we prefer this representation as it saves a significant amount of memory;

- SNPQC: SNP quality score;
- SNR: Signal-to-noise ratio.

```
R> calls(crlmmResult)[1:10,]
```

	NA06985_GW6_C.CEL	NA06991_GW6_C.CEL
SNP_A-2131660	2	2
SNP_A-1967418	3	3
SNP_A-1969580	3	3
SNP_A-4263484	2	1
SNP_A-1978185	1	1
SNP_A-4264431	1	1
SNP_A-1980898	3	3
SNP_A-1983139	1	1
SNP_A-4265735	2	2
SNP_A-1995832	2	3
	NA06993_GW6_C.CEL	
SNP_A-2131660	3	
SNP_A-1967418	3	
SNP_A-1969580	3	
SNP_A-4263484	1	
SNP_A-1978185	1	
SNP_A-4264431	1	
SNP_A-1980898	3	
SNP_A-1983139	1	
SNP_A-4265735	1	
SNP_A-1995832	3	

```
R> confs(crlmmResult)[1:10,]
```

	NA06985_GW6_C.CEL	NA06991_GW6_C.CEL
SNP_A-2131660	0.9999964	0.9999996
SNP_A-1967418	0.9999969	0.9999997
SNP_A-1969580	0.9995187	0.9995139
SNP_A-4263484	0.9999999	1.0000000
SNP_A-1978185	1.0000000	1.0000000
SNP_A-4264431	1.0000000	1.0000000
SNP_A-1980898	0.9995192	0.9995206
SNP_A-1983139	1.0000000	0.9999877
SNP_A-4265735	0.9999822	0.9999862
SNP_A-1995832	0.9999758	1.0000000
	NA06993_GW6_C.CEL	
SNP_A-2131660	0.9999998	
SNP_A-1967418	0.9999969	
SNP_A-1969580	0.9995124	
SNP_A-4263484	1.0000000	

```
SNP_A-1978185      1.0000000
SNP_A-4264431      1.0000000
SNP_A-1980898      0.9995134
SNP_A-1983139      1.0000000
SNP_A-4265735      0.9999998
SNP_A-1995832      0.9999999
```

```
R> crlmmResult[["SNR"]]
```

```
[1] 8.578197 8.368455 7.265255
```

2 Details

This document was written using:

```
R> sessionInfo()
```

```
R version 3.2.0 (2015-04-16)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: Ubuntu 14.04.2 LTS
```

locale:

```
[1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8      LC_COLLATE=C
[5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8     LC_NAME=C
[9] LC_ADDRESS=C             LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

attached base packages:

```
[1] stats      graphics  grDevices  utils      datasets
[6] methods    base
```

other attached packages:

```
[1] genomewidesnp6Crlmm_1.0.7 hapmapsnp6_1.9.1
[3] crlmm_1.26.0                preprocessCore_1.30.0
[5] oligoClasses_1.30.0
```

loaded via a namespace (and not attached):

```
[1] Rcpp_0.11.5      XVector_0.8.0
[3] splines_3.2.0    GenomicRanges_1.20.0
[5] BiocGenerics_0.14.0 zlibbioc_1.14.0
[7] IRanges_2.2.0    bit_1.1-12
[9] ellipse_0.3-8    lattice_0.20-31
[11] foreach_1.4.2    GenomeInfoDb_1.4.0
[13] tools_3.2.0      base64_1.1
```

[15] grid_3.2.0 parallel_3.2.0
[17] Biobase_2.28.0 ff_2.2-13
[19] DBI_0.3.1 matrixStats_0.14.0
[21] iterators_1.0.7 RcppEigen_0.3.2.4.0
[23] affyio_1.36.0 Matrix_1.2-0
[25] S4Vectors_0.6.0 codetools_0.2-11
[27] VGAM_0.9-7 RSQLite_1.0.0
[29] BiocInstaller_1.18.1 Biostrings_2.36.0
[31] stats4_3.2.0 mvtnorm_1.0-2
[33] illuminaio_0.10.0