

Infrastructure classes for high-throughput SNP data

Robert Scharpf and Benilton Carvalho

April 22, 2010

This document describes some of the infrastructure classes used for high-throughput genomic data. For the classes used to organize SNP data, we provide examples for initialization and illustrate some of the accessors. We should add a diagram showing the relationships of these classes here.

[Insert diagram of classes here]

1 Feature-level classes

2 Locus-level classes

The examples below are completely simulated and are not meant to convey any biological plausibility.

2.1 SnpSet

2.1.1 Initialization

```
> theCalls <- matrix(sample(1:3, 20, rep = TRUE), nc = 2)
> p <- matrix(runif(20), nc = 2)
> theConfs <- round(-1000 * log2(1 - p))
> obj <- new("SnpSet", call = theCalls, callProbability = theConfs)
```

2.1.2 Accessors

```
> calls(obj)
```

```
      1 2
1    1 1
2    1 1
3    1 1
4    3 1
5    1 3
6    3 2
7    2 2
8    2 1
9    2 1
10   2 2
```

```
> confs(obj)
```

```
      1      2
1 0.74614702 0.80051176
2 0.37624649 0.03342850
3 0.12278223 0.57172938
4 0.45827674 0.99617042
5 0.85942301 0.99880785
```

```

6 0.94900359 0.37249272
7 0.77909002 0.35143951
8 0.09425729 0.42993221
9 0.08515443 0.69970802
10 0.85397829 0.44622677

```

2.1.3 Annotating

```

> if (require("genomewidesnp6Crlmm")) {
+   ids <- c("SNP_A-2131660", "SNP_A-1967418", "SNP_A-1969580",
+           "SNP_A-4263484", "SNP_A-1978185", "SNP_A-4264431",
+           "SNP_A-1980898", "SNP_A-1983139", "SNP_A-4265735",
+           "SNP_A-1995832")
+   rownames(theCalls) <- rownames(p) <- rownames(theConfs) <- ids
+   obj <- new("SnpSet", call = theCalls, callProbability = theConfs,
+             annotation = "genomewidesnp6")
+   featureData(obj) <- addFeatureAnnotation(obj)
+   fvarLabels(obj)
+   isSnp(obj)
+   position(obj)
+   chromosome(obj)
+ }

[1] 1 1 1 1 1 1 1 1 1 1

```

2.2 CopyNumberSet

2.2.1 Initialization

2.2.2 Accessors

2.2.3 Annotating

2.3 CNSet

2.3.1 Initialization

```

> theCalls <- matrix(2, nc = 2, nrow = 10)
> A <- matrix(sample(1:1000, 20), 10, 2)
> B <- matrix(sample(1:1000, 20), 10, 2)
> CA <- matrix(rnorm(20, 1), nrow = 10)
> CB <- matrix(rnorm(20, 1), nrow = 10)
> p <- matrix(runif(20), nc = 2)
> theConfs <- round(-1000 * log2(1 - p))
> obj <- new("CNSet", alleleA = A, alleleB = B, call = theCalls,
+   callProbability = theConfs, CA = CA, CB = CB)

```

2.3.2 Accessors

```

> calls(obj)

  1 2
1  2 2
2  2 2
3  2 2
4  2 2
5  2 2

```

```

6 2 2
7 2 2
8 2 2
9 2 2
10 2 2

```

```
> confs(obj)
```

```

      1      2
1 0.98425134 0.2643494
2 0.28107627 0.9866067
3 0.98847227 0.4872670
4 0.48982425 0.4753375
5 0.24873738 0.8019033
6 0.76236039 0.8529526
7 0.60662818 0.8914994
8 0.39830223 0.9942812
9 0.07503557 0.6451287
10 0.58188552 0.4456727

```

```
> A(obj)
```

```

      1      2
1 318 65
2 421 23
3 978 385
4 216 965
5 280 7
6 565 487
7 109 86
8 221 856
9 132 858
10 101 687

```

```
> B(obj)
```

```

      1      2
1 776 510
2 558 420
3 593 884
4 397 427
5 98 597
6 244 507
7 497 781
8 315 594
9 456 825
10 565 458

```

```
> CA(obj)
```

```

      1      2
1 0.2417622 2.2871349
2 -0.8927851 2.4248798
3 0.7122214 0.8530986
4 0.1155456 0.5946828
5 1.7244929 0.5157260

```

```

6  2.0705214 1.7154788
7  1.6601794 0.3433505
8  -1.3795081 0.4317529
9  0.4289145 0.0787818
10 1.9752856 1.4900867

```

```
> CB(obj)
```

```

      1      2
1  1.6187216 -0.03486392
2  0.9606666  1.00103480
3  0.9730937  1.25903846
4  0.8758988 -0.20674122
5  1.4018587  1.68203691
6  0.5002312  0.81031770
7  0.1581807  1.29988917
8  1.0960045  1.28153208
9  0.3079477 -1.06330902
10 1.8038336  0.69873342

```

2.3.3 Annotating

Annotating with chromosome and physical position:

```

> if (require("genomewidesnp6Crlmm")) {
+   ids <- c("SNP_A-2131660", "SNP_A-1967418", "SNP_A-1969580",
+           "SNP_A-4263484", "SNP_A-1978185", "SNP_A-4264431",
+           "SNP_A-1980898", "SNP_A-1983139", "SNP_A-4265735",
+           "SNP_A-1995832")
+   rownames(theCalls) <- rownames(p) <- rownames(theConfs) <- ids
+   rownames(A) <- rownames(B) <- rownames(CA) <- rownames(CB) <- ids
+   obj2 <- new("CNSet", alleleA = A, alleleB = B, call = theCalls,
+               callProbability = theConfs, CA = CA, CB = CB, annotation = "genomewidesnp6")
+   fvarLabels(obj2)
+   isSnp(obj2)
+   chromosome(obj2)
+   position(obj2)
+ }

```

3 Session Information

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

- R version 2.11.0 (2010-04-22), i386-pc-mingw32
- Locale: LC_COLLATE=English_United States.1252, LC_CTYPE=English_United States.1252, LC_MONETARY=English_United States.1252, LC_NUMERIC=C, LC_TIME=English_United States.1252
- Base packages: base, datasets, graphics, grDevices, methods, stats, tools, utils
- Other packages: Biobase 2.8.0, genomewidesnp6Crlmm 1.0.2, oligoClasses 1.10.0
- Loaded via a namespace (and not attached): affyio 1.16.0, Biostrings 2.16.0, IRanges 1.6.0