Infrastructure classes for high-throughput SNP data

Robert Scharpf and Benilton Carvalho

February 20, 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  2 1
2  3 2
3  2 2
4  3 2
5  3 2
6  1 2
7  2 1
8  3 2
9  1 3
10 2 2
```

```
> confs(obj)
```

```
          1         2
1  0.7703043 0.5127607
2  0.7913297 0.6733938
3  0.1227822 0.4473332
4  0.4964136 0.8861644
5  0.4695343 0.6212959
```

```
6  0.9447562 0.8859365
7  0.1820876 0.9313682
8  0.4417788 0.8328728
9  0.1845376 0.8323707
10 0.3928625 0.2621391
```

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")
+     obj2 <- annotate(obj)
+     fvarLabels(obj2)
+     isSnp(obj2)
+     position(obj2)
+     chromosome(obj2)
+ }
```

```
 [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  0.39830223 0.1077420
2  0.11750310 0.7489241
3  0.84260552 0.7399809
4  0.85499688 0.3724927
5  0.06573953 0.9957888
6  0.71263440 0.1512580
7  0.83699917 0.8828148
8  0.98110768 0.3336898
9  0.13151069 0.3885976
10 0.59016474 0.5200148
```

```
> A(obj)
```

```
     1   2
1  474 271
2  147 270
3  785 236
4  465 461
5  342  70
6  695 890
7   83 561
8  791 451
9  960 492
10 892 591
```

```
> B(obj)
```

```
     1   2
1  970 405
2  509 146
3  245 293
4   69 101
5  961 303
6  236 917
7  395 211
8  222  47
9  763 656
10  20 609
```

```
> CA(obj)
```

```
             1            2
1  0.008983685  0.015524212
2  0.001100247  0.001647221
3  0.013217960  0.007248202
4  0.005388631 -0.011092939
5  0.011176663  0.021521456
```

```
6  0.018182685  0.008248945
7  0.006076325  0.011227146
8  0.005625125  0.011521493
9  0.007165264 -0.003701394
10 0.006025052  0.013518060
```

```
> CB(obj)
```

```
               1            2
1   0.0188421270 -0.001422206
2   0.0318297695 -0.002306198
3   0.0354458776  0.011794732
4   0.0225196750  0.010542550
5   0.0153654657  0.035661676
6   0.0087355890  0.012281596
7   0.0240341197  0.009970778
8   0.0016935304  0.025440471
9  -0.0009821043  0.023683896
10  0.0144125793  0.006976135
```

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 Under development (unstable) (2009-11-20 r50517), `x86_64-apple-darwin9.8.0`
- Locale: `C`
- Base packages: base, datasets, grDevices, graphics, methods, stats, tools, utils
- Other packages: Biobase 2.7.4, genomewidesnp6Crlmm 1.0.4, oligoClasses 1.9.30
- Loaded via a namespace (and not attached): Biostrings 2.15.11, IRanges 1.5.21, affyio 1.15.2