Infrastructure classes for high-throughput SNP data

Robert Scharpf and Benilton Carvalho

July 14, 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  3 2
2  1 3
3  2 2
4  1 2
5  2 3
6  3 3
7  3 2
8  3 3
9  2 2
10 1 3
```

```
> confs(obj)
```

```
           1          2
1  0.3188686 0.19022593
2  0.4774319 0.18453763
3  0.2576987 0.32429589
4  0.3892087 0.15464617
5  0.6518076 0.88875274
```

```
6  0.5876983 0.64971214
7  0.0924440 0.59464822
8  0.8298373 0.01783897
9  0.1622202 0.28179483
10 0.7553679 0.09516258
```

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  0.65285058 0.20863818
2  0.09425729 0.94128147
3  0.99397598 0.89585801
4  0.85221548 0.86220663
5  0.91791500 0.95056022
6  0.07132831 0.98925162
7  0.57514188 0.22275526
8  0.24346010 0.07780631
9  0.35853458 0.57044264
10 0.07688365 0.67793439
```

```
> A(obj)
```

```
     1   2
1  366 508
2  198  71
3  643  45
4  913 130
5   10 887
6  417 221
7  590 311
8  178 424
9  429 992
10 447 139
```

```
> B(obj)
```

```
     1   2
1  769 908
2  402 394
3  680 358
4   79 970
5    8 696
6   85 582
7  749 929
8  767 435
9  728  21
10 344 205
```

```
> CA(obj)
```

```
             1          2
1   0.09763923  0.7451489
2   2.16017476 -0.3168747
3   1.76494762 -0.9359526
4  -0.13252238  0.6921249
5   1.01859800  1.8430487
```

```
6  -0.12303938 -0.8673966
7   2.01012557  1.7466194
8  -0.38111920  1.1632133
9   3.92954265  0.5046183
10 -0.32755537  1.6799286
```

```
> CB(obj)
```

```
            1           2
1  -0.2248659  2.28865399
2   0.4969427  2.59149180
3   2.3889220  1.96349414
4   0.8114921  0.83451404
5   2.1217974  2.57355114
6   1.1309536  1.59493715
7   2.7230521  0.67455462
8   2.0732825  0.22707415
9   1.1630574 -0.09253353
10  1.4193042  1.01571406
```

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), `x86_64-apple-darwin9.8.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/C/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, stats, tools, utils
- Other packages: Biobase 2.8.0, genomewidesnp6Crlmm 1.0.4, oligoClasses 1.11.3
- Loaded via a namespace (and not attached): affyio 1.15.2, Biostrings 2.16.0, IRanges 1.5.74