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

```
> confs(obj)
```

```
           1           2
1  0.1918439 0.006975557
2  0.8960661 0.438980716
3  0.7913297 0.735522739
4  0.7866881 0.137568885
5  0.8218269 0.552017003
```

```
6  0.5488701 0.019801327
7  0.5905744 0.702100769
8  0.8046573 0.234326929
9  0.1358423 0.928852493
10 0.2803570 0.418997373
```

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  0.91310024 0.68462725
2  0.76636612 0.83017724
3  0.06105653 0.05161999
4  0.41315820 0.99901906
5  0.99838916 0.54159399
6  0.39649442 0.56699240
7  0.99986857 0.37249272
8  0.96669341 0.14101172
9  0.99844612 0.99832677
10 0.23814574 0.83220294
```

```
> A(obj)
```

```
     1   2
1  506 785
2  261 351
3  607 943
4  577 475
5  382 925
6  257 181
7  996 532
8  299 138
9  847 395
10 120 827
```

```
> B(obj)
```

```
     1   2
1  331 356
2   99 736
3  323 860
4  710 267
5  984 547
6  713 879
7   93 965
8  951 890
9  353 810
10 239 533
```

```
> CA(obj)
```

```
            1         2
1   1.0092238 0.8405214
2   1.1472936 0.9309621
3  -0.8306025 1.6433172
4   0.8327707 1.9346413
5   0.9084145 0.6331778
```

```
6   0.4990710 1.0097688
7   2.5562772 1.0578405
8  -0.3692148 0.9893496
9   1.5056520 2.3006948
10  0.1377170 1.2682668
```

```
> CB(obj)
```

```
            1          2
1  -0.3627261  0.1281610
2   0.2725810 -0.6505326
3   0.9611718  1.5212885
4   1.6760230  0.8739650
5   1.4101528  1.8772932
6  -1.8130730 -1.4804315
7   1.1357816  2.7162109
8   2.1790865  2.0984553
9   1.6323722  1.4273124
10  2.0166956  0.9195163
```

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-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US`, `LC_NUMERIC=C`, `LC_TIME=en_US`, `LC_COLLATE=en_US`, `LC_MONETARY=C`, `LC_MESSAGES=en_US`, `LC_PAPER=en_US`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US`, `LC_IDENTIFICATION=C`
- 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