Infrastructure classes for high-throughput SNP data

Robert Scharpf and Benilton Carvalho

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

```
> confs(obj)
```

```
           1         2
1  0.91370641 0.8774211
2  0.95899249 0.8706198
3  0.65630478 0.5600086
4  0.71916838 0.9944336
5  0.86330457 0.9998688
```

```
6  0.36045609 0.8474099
7  0.77122127 0.9702224
8  0.80387440 0.6426359
9  0.22430820 0.5397567
10 0.07964885 0.8387824
```

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  0.58934425 0.6694509
2  0.96787112 0.6843117
3  0.64227831 0.0713283
4  0.80387440 0.7642539
5  0.03439458 0.5742913
6  0.55558657 0.4384194
7  0.46097860 0.9089184
8  0.23585674 0.4361686
9  0.72251822 0.5243153
10 0.89923861 0.6486597
```

```
> A(obj)
```

```
     1   2
1  224 977
2  387 452
3    1 958
4  491 697
5  749 272
6   52 905
7  570 429
8  998 717
9  289 488
10 626 813
```

```
> B(obj)
```

```
     1   2
1  997 806
2  586 854
3  494 395
4  719 431
5  410 330
6  812 351
7  698 565
8  630 165
9  388 516
10 529 658
```

```
> CA(obj)
```

```
               1            2
1  -0.0001551915  0.015529101
2   0.0217232092  0.013800745
3   0.0095727172 -0.001077235
4   0.0078306166  0.014987873
5  -0.0068994153  0.013478595
```

```
6   0.0080598788  0.019281828
7  -0.0012928982  0.014932916
8   0.0129150067  0.020472545
9   0.0327498594  0.010126858
10  0.0198007926  0.010725786
```

```
> CB(obj)
```

```
               1             2
1   0.0301343109  0.0040923267
2   0.0070854587  0.0141486523
3   0.0014621552  0.0198082749
4  -0.0003529107  0.0008205164
5   0.0076772459 -0.0053354811
6   0.0131603924 -0.0025952547
7   0.0137181660 -0.0021015574
8  -0.0115776809  0.0157947858
9   0.0099467861  0.0038608726
10  0.0083275436  0.0163106459
```

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) (2009-11-22 r50541), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.iso885915`, `LC_NUMERIC=C`, `LC_TIME=en_US.iso885915`, `LC_COLLATE=en_US.iso885915`, `LC_MONETARY=C`, `LC_MESSAGES=en_US.iso885915`, `LC_PAPER=en_US.iso885915`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.iso885915`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, stats, tools, utils
- Other packages: Biobase 2.7.3, genomewidesnp6Crlmm 1.0.4, oligoClasses 1.9.41
- Loaded via a namespace (and not attached): affyio 1.15.2, Biostrings 2.15.11, IRanges 1.5.34