Infrastructure classes for high-throughput SNP data

Robert Scharpf and Benilton Carvalho

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

```
> confs(obj)
```

```
           1         2
1  0.02371429 0.8664127
2  0.87729841 0.6740463
3  0.98242012 0.1367060
4  0.72909850 0.8777882
5  0.47638573 0.5909836
```

```
6  0.69394792 0.9692234
7  0.66040447 0.8684015
8  0.98569291 0.3824529
9  0.81194109 0.4668753
10 0.33701307 0.7146389
```

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  0.7023985 0.3553192
2  0.9837392 0.8654743
3  0.8216487 0.1918439
4  0.6782563 0.8303470
5  0.7290985 0.8241282
6  0.6753475 0.1175031
7  0.0942573 0.9788052
8  0.9044399 0.9730561
9  0.9737477 0.3436099
10 0.6476041 0.9956518
```

```
> A(obj)
```

```
     1   2
1  384 593
2   14 358
3  564   5
4  805 541
5  567 894
6  116 458
7   24 692
8  939 882
9  348 114
10 734 312
```

```
> B(obj)
```

```
     1   2
1  628 290
2   39 623
3   11 647
4  946   3
5  763 589
6  687 199
7  253 363
8   49  84
9  899 733
10 430 617
```

```
> CA(obj)
```

```
            1         2
1   1.1651930 0.3232036
2   0.7717648 0.6783353
3   0.3990063 1.0989849
4   2.0373101 1.3480760
5   2.2598620 0.1827224
```

```
6   1.2796330 1.4056871
7   2.2517126 2.0694474
8  -2.3133200 0.2755148
9   1.2312328 2.2691824
10  1.3488303 0.3345079
```

```
> CB(obj)
```

```
            1          2
1   0.7632525 -0.26869039
2   0.8358506  2.91226036
3   0.5829480  0.86731722
4   0.3704712 -0.07043006
5  -1.1909143  0.94082018
6   1.5940368  1.69730917
7  -0.6725159  0.37714892
8   1.0096008  1.97106542
9   1.6326186  2.36423243
10  3.1473249  2.46284922
```

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: `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, utils
- Other packages: Biobase 2.7.4, genomewidesnp6Crlmm 1.0.4, oligoClasses 1.9.50
- Loaded via a namespace (and not attached): affyio 1.15.2, Biostrings 2.15.11, IRanges 1.5.21