Infrastructure classes for high-throughput SNP data

Robert Scharpf and Benilton Carvalho

April 21, 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 3
2  1 1
3  2 2
4  1 3
5  3 1
6  2 3
7  3 3
8  2 1
9  2 1
10 2 2
```

```
> confs(obj)
```

```
           1          2
1  0.7323299 0.05446086
2  0.6617601 0.71888741
3  0.6620982 0.03535971
4  0.5443362 0.63578102
5  0.8713938 0.07596008
```

```
6  0.1005754 0.49942608
7  0.8438597 0.09244399
8  0.9638472 0.60505154
9  0.8881951 0.88210961
10 0.7371049 0.71634597
```

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  0.01882064 0.5247907
2  0.87267342 0.2931947
3  0.67889914 0.3970976
4  0.25769866 0.9356231
5  0.42821944 0.9878933
6  0.45118836 0.3290093
7  0.14955880 0.5998837
8  0.69516968 0.9861850
9  0.46953431 0.3585346
10 0.83650944 0.9124898
```

```
> A(obj)
```

```
     1   2
1  832 362
2  356 795
3  301 637
4  230  11
5  706 399
6  373 376
7  524 598
8  848 458
9  177  95
10  42 141
```

```
> B(obj)
```

```
     1   2
1  677 710
2   83 704
3  687 353
4  574 886
5  493  13
6  853  22
7  234 299
8  381  18
9  527 140
10 502 828
```

```
> CA(obj)
```

```
            1         2
1   1.2180633 1.0038317
2  -0.6629601 1.9796429
3   1.2938908 1.3000931
4   2.8049309 2.1270254
5   0.2104208 0.3261277
```

```
6   1.3961819 0.1669671
7   1.3846664 1.7155701
8   0.6653007 1.0901007
9  -0.5603489 2.5049052
10  2.1979889 0.6245305
```

```
> CB(obj)
```

```
            1          2
1   1.4447594  1.4913310
2   1.6835406  1.8989601
3  -0.3557915  2.6851352
4   2.9892648  0.9716800
5   3.6057273  1.3679556
6   0.7918260  0.7987002
7   2.8280835  0.2539961
8  -0.3950830  2.9025483
9   0.9506834 -1.1234216
10  0.3971347  0.9641129
```

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 2.7.5, genomewidesnp6Crlmm 1.0.4, oligoClasses 1.9.58
- Loaded via a namespace (and not attached): affyio 1.15.2, Biostrings 2.15.25, IRanges 1.5.74