Infrastructure classes for high-throughput SNP data

Robert Scharpf and Benilton Carvalho

March 31, 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  2 2
3  2 2
4  1 3
5  2 2
6  2 3
7  3 2
8  2 1
9  3 3
10 1 3
```

```
> confs(obj)
```

```
           1         2
1  0.06760618 0.7261021
2  0.29601625 0.4679405
3  0.53837389 0.7068292
4  0.21021933 0.6140317
5  0.57852719 0.5355958
```

```
6  0.94508672 0.3215877
7  0.87742105 0.9354942
8  0.13843089 0.7900740
9  0.07132831 0.7496762
10 0.08423912 0.3540286
```

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  0.03246144 0.8610999
2  0.75463294 0.8730548
3  0.80308832 0.4893138
4  0.94702843 0.2832298
5  0.14015230 0.6948647
6  0.02761163 0.3376757
7  0.31202309 0.1605430
8  0.06293254 0.9832105
9  0.48829142 0.6346867
10 0.69820280 0.6231847
```

```
> A(obj)
```

```
     1   2
1  484 310
2  503  23
3  910 935
4  290 251
5  362 493
6  103 919
7  746 988
8  729 288
9  212 482
10 316 649
```

```
> B(obj)
```

```
     1   2
1  442   6
2  871  83
3  398 834
4  103  20
5  967 554
6  317 784
7  929 524
8  144 521
9  921 649
10 969 714
```

```
> CA(obj)
```

```
            1          2
1   0.6640014  1.3007666
2   1.6490427 -0.6824386
3  -0.3378797  1.7708979
4   2.5684436  2.2697932
5  -0.2505532 -0.3140521
```

```
6   1.7252625  1.8218891
7   0.4566130  3.8721221
8   1.2683090  0.6563965
9   2.9491808  2.2254479
10  2.5372163  0.7464832
```

```
> CB(obj)
```

```
            1          2
1  0.59674563  1.4253802
2  0.78462741  0.1499211
3  0.08052294 -0.4850030
4  0.93078761  1.0641703
5  0.19604213  1.2765442
6  1.12021241  0.4323002
7  0.51693415 -0.4764148
8  0.76344649  1.2119607
9  0.63714004  1.0307953
10 1.49759177  2.6432544
```

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 1.9.50
- Loaded via a namespace (and not attached): affyio 1.15.2, Biostrings 2.15.25, IRanges 1.5.69