nkiBRCA vignette

Philip Schouten

April 13, 2017

1 Introduction

```
> library(nkiBRCA)
```

Shrunken centroid classifiers (1) were trained to predict whether a DNA copy number profile is similar to that of a BRCA1 or BRCA2 mutated DNA copy number profile (2,3). This package includes the information of the BAC array CGH platform that was used to train the classifiers on. In this vignette we will first show how to use the package on BAC array CGH data and then provide suggestions on the use of data from other platforms for classification.

2 Demonstration of classification of BAC array CGH samples

2.1 BAC array CGH platform

The BAC array CGH platform that was used to generate the data on is described in the objects b1.191.ct, b1.371.ct and b2.704.ct. These objects also contain the parameters for classification, which we will discuss later in this vignette. The platform file was mapped on hg18 by aligning the BAC clones. Some locations have been manually corrected.

Here we show the platform and that this platform is included in all the classifier objects:

```
> head(b1.191.ct[,1:8])
```

```
  Order       Clone chrom Genomic.position   Start     End BAC.size  maploc
1  3535  GS-232-B23     1           200000       0       0        0  200000
2  3537    GS-62-L8     1           200050       0       0        0  200050
3  2305  RP4-785P20     1          3284807 3214521 3355092   140571 3284807
4   145   RP1-37J18     1          4542451 4476787 4608114   131327 4542451
5  2309   RP3-438L4     1          7103443 7059893 7146992    87099 7103443
6  2213 RP11-338N10     1          7674468 7641507 7707428    65921 7674468
```

```
> identical(b1.191.ct[,1:8], b1.371.ct[,1:8])
```

```
[1] TRUE
```

```
> identical(b1.191.ct[,1:8], b2.704.ct[,1:8])
```

```
[1] TRUE
```

The platform description contains the following values:

1. Order: print order
2. Clone: BAC clone name
3. chrom: hg18 chromosome
4. Genomic.position: hg18 genomic position
5. Start: hg18 chromosomal start position of the BAC clone
6. End: hg18 chromosomal end position of the BAC clone
7. BAC.size: size of the BAC clone
8. maploc: chromosomal midposition of the BAC clone

2.2 Three classifiers

This package contains the information for 3 classifiers. Two BRCA1 classifiers and 1 BRCA2 classifiers.

The BRCA1 classifier has two versions, one with 371 probes and one with 191 probes because an early development version (371 probes) was tested for the prediction of chemotherapy benefit (4,5). Ever since these two classifiers have been used separately. The early development version contained 371 probes, and uses a cutoff in the discriminative score of 0.63 for classification to predict high dose chemotherapy benefit. The cutoff of 0.63 was trained on predicting the outcome of high dose chemotherapy benefit. The early version was further developed into a classifier with 191 probes and a cutoff of 0.5 to predict BRCA1 association status (2).

Although differences are present between the 191 probe and 371 probe classifier these are practically minimal, resulting in differences in the discriminative score that do not influence the overall predicted class. In our experience both identify BRCA1 mutated and BRCA1 methylated cancer, as well as predict chemotherapy benefit.

The BRCA2 classifier contains 704 probes and has been developed to identify BRCA2 mutated cancers (3). In addition it also predicts for the benefit of high dose chemotherapy (5). Its cutoff is 0.5 (3,5). The data inputted to the BRCA2 classifier need to be segmented using the cghseg package (8).

The functions corresponding to the BRCA1 191, BRCA1 371 and BRCA2 704 classifiers are respectively b1191, b1371, b2704. The parameters are stored in the objects b1.191.ct, b1.371.ct, b2.704.ct.

classifier	number of probes	function	cut-off	originally developed to	input data
BRCA1	191	b1191	0.5	BRCA1-like class	unsegmented log ratios
BRCA1	371	b1371	0.63	BRCA1-like class and high dose chemo benefit	unsegmented log ratios
BRCA2	704	b2704	0.5	BRCA2-like class (validated for high dose chemo benefit)	segmented log ratios

Here we show the classifier values for the three different classifiers:

```
> head(b1.191.ct[,-1:-8])
```

```
  shrunken.centroid.sporadic shrunken.centroid.BRCA1.like      si     s0 priors
1                  0.1908900                    0.1908900 0.14728 0.1542    0.5
2                  0.0242010                    0.0242010 0.18183     NA     NA
3                 -0.0969060                   -0.0969060 0.15005     NA     NA
4                 -0.1725400                   -0.1725400 0.18265     NA     NA
5                  0.0032085                    0.0032085 0.19591     NA     NA
6                 -0.0021208                   -0.0021208 0.14162     NA     NA
```

```
> head (b1.371.ct[,-1:-8])
```

```
  shrunken.centroid.sporadic shrunken.centroid.BRCA1.like      si      s0
1                   0.198570                     0.198570 0.15561 0.16461
2                   0.020298                     0.020298 0.19186      NA
3                  -0.072816                    -0.072816 0.15286      NA
4                  -0.156430                    -0.156430 0.19404      NA
5                   0.021700                     0.021700 0.19974      NA
6                  -0.017411                    -0.017411 0.14860      NA
  priors
1    0.5
2     NA
3     NA
4     NA
5     NA
6     NA
```

```
> head (b2.704.ct[,-1:-8])
```

```
  shrunken.centroid.sporadic shrunken.centroid.BRCA2.like         si    s0
1                    0.17273                      0.17273 0.31420718 0.128
2                   -0.01796                     -0.01796 0.15807365    NA
3                   -0.05052                     -0.05052 0.11449110    NA
4                   -0.06326                     -0.06326 0.09726277    NA
5                   -0.05754                     -0.05754 0.08980518    NA
6                   -0.05262                     -0.05262 0.08087379    NA
  priors
1    0.5
2     NA
3     NA
4     NA
5     NA
6     NA
```

The objects contain the following values:

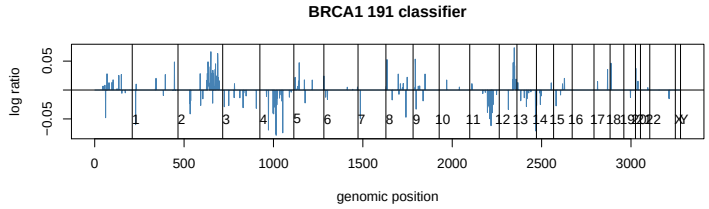
1. shrunken.centroid.sporadic: centroid of the non-BRCA1-like class
2. shrunken.centroid.BRCA1.like: centroid of the BRCA1-like class
3. si: si in the shrunken centroids formula
4. s0: s0 in the shrunken centroids formula.
5. priors: prior probability in the shrunken centroids classifier.

Here we show the number of selected probes in the classifiers:

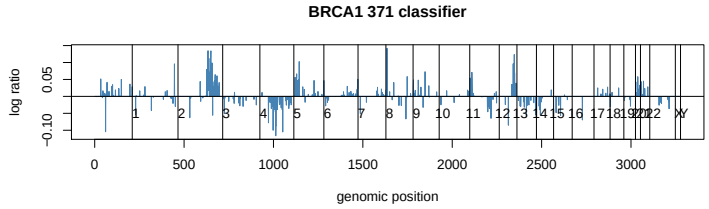
```
> sum((b1.191.ct$shrunken.centroid.BRCA1.like - b1.191.ct$shrunken.centroid.sporadic)!=0)
[1] 191
> sum((b1.371.ct$shrunken.centroid.BRCA1.like - b1.371.ct$shrunken.centroid.sporadic)!=0)
[1] 371
> sum((b2.704.ct$shrunken.centroid.BRCA2.like - b2.704.ct$shrunken.centroid.sporadic)!=0)
[1] 703
```

A visualization of the classifiers:

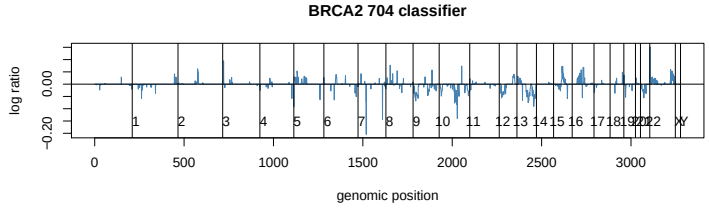
```
> plot((b1.191.ct$shrunken.centroid.BRCA1.like - b1.191.ct$shrunken.centroid.sporadic),
+   type='l', col='steelblue', ylab='log ratio', xlab='genomic position', main='BRCA1 191
> abline(v=cumsum(table(b1.191.ct$chrom)),h=0)
> text(x=cumsum(table(b1.191.ct$chrom))+20, y = rep( -0.05,24), labels=c(1:22,'X','Y'))
```




```
> plot((b1.371.ct$shrunken.centroid.BRCA1.like - b1.371.ct$shrunken.centroid.sporadic),
+   type='l', col='steelblue', ylab='log ratio', xlab='genomic position', main='BRCA1 371
> abline(v=cumsum(table(b1.371.ct$chrom)),h=0)
> text(x=cumsum(table(b1.371.ct$chrom))+20, y = rep( -0.05,24), labels=c(1:22,'X','Y'))
```




```
> plot((b2.704.ct$shrunken.centroid.BRCA2.like - b2.704.ct$shrunken.centroid.sporadic),
+   type='l', col='steelblue', ylab='log ratio', xlab='genomic position', main='BRCA2 704
> abline(v=cumsum(table(b2.704.ct$chrom)),h=0)
> text(x=cumsum(table(b2.704.ct$chrom))+20, y = rep( -0.15,24), labels=c(1:22,'X','Y'))
```



2.3 Example data

Two files of example data are attached to the package. One contains the log ratios of 6 samples, the other the segmented log ratios of the same 6 samples. Two of these samples are BRCA1-like by both the 191 and 371 probe classifier, 2 are BRCA2-like and two are non-BRCA-like (not BRCA1-like, not BRCA2-like). Segmentation was done with the cghseg package.

Here we show the example data in log ratios and segmented log ratios.

```
> head(example.ratios)
```

```
  chrom  maploc    example1     example2    example3    example4    example5
1     1  200000  0.21143548  0.295691392  0.33186595  0.43351383  0.18116566
2     1  200050 -0.44392062  0.006722115 -0.40602861 -0.14235950  0.40217427
3     1 3284807  0.05072993  0.177434108 -0.17224302 -0.21999883 -0.15992900
4     1 4542451  0.22845487 -0.500350375 -0.24091633  0.07421830  0.01486702
5     1 7103443  0.51288076 -0.372330458 -0.01931715  0.06607459 -0.17076901
6     1 7674468 -0.14474415 -0.351813860 -0.18512347 -0.10840547  0.23994466
    example6
1  0.1198949
2  0.3734321
3 -0.1018603
4  0.1187189
5 -0.1007888
6  0.1380388
```

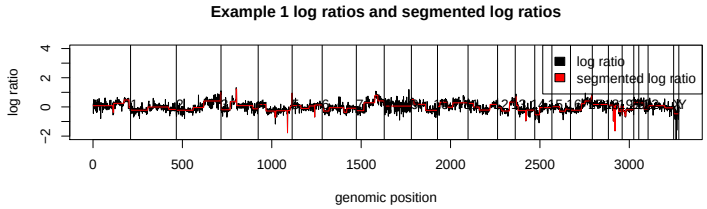
```
> head(example.segments)
```

```
  chrom  maploc   example1   example2   example3    example4    example5
1     1  200000 0.08533588 0.07868787 -0.1143354  0.43351383  0.29166997
2     1  200050 0.08533588 0.07868787 -0.1143354 -0.06583884  0.29166997
3     1 3284807 0.08533588 0.07868787 -0.1143354 -0.06583884 -0.05523652
4     1 4542451 0.08533588 0.07868787 -0.1143354 -0.06583884 -0.05523652
5     1 7103443 0.08533588 0.07868787 -0.1143354 -0.06583884 -0.05523652
6     1 7674468 0.08533588 0.07868787 -0.1143354 -0.06583884 -0.05523652
     example6
1 -0.03539515
2 -0.03539515
3 -0.03539515
4 -0.03539515
5 -0.03539515
6 -0.03539515
```

Chrom and maploc are used as positions, after which the columns with sample (segmented) log ratios are shown:

```
>  plot(example.ratios$example1,
+   type='l', col='black', ylab='log ratio', xlab='genomic position',
+   main="Example 1 log ratios and segmented log ratios", ylim=c(-2,4))
>  lines(example.segments$example1, col='red',cex=3)
>  abline(v=cumsum(table(b1.191.ct$chrom)))
>  text(x=cumsum(table(b1.191.ct$chrom))+20, y = rep(0.2,24), labels=c(1:22,'X','Y'))
```

```
>  legend('topright', fill=c('black', 'red'), legend = c('log ratio',
+   'segmented log ratio'))
```



To classify samples with the BRCA1 classifier we use log ratios we show here how to get the discriminative score, and how to classify with the respective cutoffs. We also show how to classify multiple samples at once:

```
> b1191(example.ratios[,3])

[1] 0.9999901

> b1191(example.ratios[,3]) > 0.5

[1] TRUE

> b1371(example.ratios[,3])

[1] 1

> b1371(example.ratios[,3]) > 0.63

[1] TRUE

> apply(example.ratios[,-1:-2], 2, b1191)

   example1    example2    example3    example4    example5    example6
0.999990137 0.999442987 0.148861801 0.006058065 0.100186932 0.261245269

> apply(example.ratios[,-1:-2], 2, b1191) > 0.5

example1 example2 example3 example4 example5 example6
    TRUE     TRUE    FALSE    FALSE    FALSE    FALSE
```

For the BRCA2 classifier we use segmented ratios (8). We show how to get the discriminative score and how to classify with the cutoff:

```
> b2704(example.segments[,3])

[1] 0.09502983

> b2704(example.segments[,3]) > 0.5

[1] FALSE
```

3 Input data obtained from other platforms than BAC array CGH

Since the BAC array CGH platform is not available anymore, we have demonstrated that it is possible to use data from other platforms to do the classification (6,7).

For these datasets it is important to process them to look similar to the BAC array CGH data. For arrays this means converting the dye signal to raw log ratio and in sequencing data this means converting the read count to raw log ratio. Subsequently, these raw log ratios can be mapped to the hg18 BAC locations by averaging the measurements within a BAC clone. For locations that can't be mapped we recommend interpolating the surrounding locations. Unpublished data suggest that scaling data to the BAC array CGH platform improves classification concordance in most cases. For the BRCA2 classifier the mapped profile needs to be segmented using the cghseg package (8).

These steps do not necessarily mean that the outcomes will be reliable. We always recommend a small set of samples that has been run on any of the validated platforms (in particular Nimblegen 135K, low coverage whole genome sequencing, both were tested with large number of samples. To a lesser extent we validated BAC32K, Nimblegen 720K, SNP6, Molecular Inversion Probe technology and targeted sequencing processed with CopyWriter)(6,7) compared to a new platform. With these samples concordance between platform/processing pipeline can be guaranteed. We did not automate these translation steps (which are fairly straightforward) in this package because we believe it is important to check that the translation works properly.

Array and sequencing platforms seem to be able to produce robust copy number data, however the ranges of inputs for this particular package is potentially large and thus not easily controllable. A package automating such steps would likely be highly restrictive and complex. In our analyses we have found influences on classification with varying prevalence and intensity across datasets. We were unable to derive clear quality control parameters, except by repeated analysis of samples across techniques. Concluding, without validation of some samples processed on both platforms/pipelines and without consistent quality assurance criteria, automating a pipeline would provide an unwarranted feeling of correctness of the analysis.

However, the current functions allow classification of data from other platforms. We are happy to help you translating your dataset, please contact us. If you perform the mapping yourself the following situations may hint at problems with the approach:

- errors: your data is likely out of range
- most classifications BRCA-like: unless biologically explainable, the data likely has a larger amplitude than the training set.
- most classification non-BRCA-like: unless biologically explainable, the data likely has a smaller amplitude than the training set.
- most discriminative scores around 0.5: various situations in which geometric distance is (almost) equally close to both class centroids.

4 References

1. Tibshirani *et. al.* Diagnosis of multiple cancer types by shrunken centroids of gene expression.. *PNAS*, 2002 99(10):6567-72.

2. Joosse *et. al.* Prediction of BRCA1-association in hereditary non-BRCA1/2 breast carcinomas with array-CGH. *Breast Cancer Res Treat.* 2009;116(3):479-89.

3. Joosse *et. al.* Prediction of BRCA2-association in hereditary breast carcinomas using array-CGH.. *Breast Cancer Res Treat.* 2012;132(2):379-89.

4. Vollebergh *et. al.* An aCGH classifier derived from BRCA1-mutated breast cancer and benefit of high-dose platinum-based chemotherapy in HER2-negative breast cancer patients. *Ann Oncol.* 2011;22(7):1561-70.

5. Vollebergh *et. al.* Genomic patterns resembling BRCA1- and BRCA2-mutated breast cancers predict benefit of intensified carboplatin-based chemotherapy. *Breast Cancer Res.* 2014;16(3):R47.

6. Schouten *et. al.* Platform comparisons for identification of breast cancers with a BRCA-like copy number profile. *Breast Cancer Res Treat.* 2013;139(2):317-27.

7. Schouten *et. al.* Robust BRCA1-like classification of copy number profiles of samples repeated across different datasets and platforms. *Mol Oncol.* 2015;9(7):1274-86.

8. Picard *et. al.* Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics.* 2011;12(3):413-28.