CGHcall: Calling aberrations for array CGH tumor profiles.

Sjoerd Vosse and Mark van de Wiel

August 3, 2009

Department of Pathology
VU University Medical Center

mark.vdwiel@vumc.nl

Contents

1 Overview

CGHcall allows users to make an objective and effective classification of their aCGH data into copy number states (loss, normal, gain or amplification). This document provides an overview on the usage of the CGHcall package. For more detailed information on the algorithm and assumptions we refer to the article (van de Wiel et al., 2007) and its supplementary material. As example data we attached the first five samples of the Wilting dataset (Wilting et al., 2006). After filtering and selecting only the autosomes 4709 datapoints remained.

2 Example

In this section we will use CGHcall to call and visualize the aberrations in the dataset described above. First, we load the package and the data:

```
> library(CGHcall)
> data(WiltingData)
> Wilting <- cghRaw(WiltingData)
```

Next, we apply the `preprocess` function which:

- removes data with unknown or invalid position information.
- shrinks the data to `nchrom` chromosomes.
- removes data with more than `maxmiss` % missing values.
- imputes missing values using `impute.knn` from the package `impute` (Troyanskaya et al., 2001).

```
> cghdata <- preprocess(Wilting, maxmiss = 30, nchrom = 22)
```

```
Changing impute.knn parameter k from 10 to 4 due to small sample size.
Cluster size 3552 broken into 984 2568
Done cluster 984
Cluster size 2568 broken into 1509 1059
Cluster size 1509 broken into 653 856
Done cluster 653
Done cluster 856
Done cluster 1509
Done cluster 1059
Done cluster 2568
```

To be able to compare profiles they need to be normalized. In this package we provide very basic global median or mode normalization. Of course, other methods can be used outside this package. This function also contains smoothing of outliers as implemented in the DNAcopy package (Venkatraman and Olshen, 2007). Furthermore, when the proportion of tumor cells is not 100% the ratios can be corrected. See the article and the supplementary material for more information on cellularity correction (van de Wiel et al., 2007).

```
> tumor.prop <- c(0.75, 0.9, 0.8, 1, 1)
> norm.cghdata <- normalize(cghdata, method = "median", cellularity = tumor.prop,
+     smoothOutliers = TRUE)
```

```
Applying median normalization ...
Smoothing outliers ...
Adjusting for cellularity ...
Cellularity sample 1 :  0.75
Cellularity sample 2 :  0.9
Cellularity sample 3 :  0.8
Cellularity sample 4 :  1
Cellularity sample 5 :  1
```

The next step is segmentation of the data. This package only provides a simple wrapper function that applies the `DNAcopy` algorithm (Venkatraman and Olshen, 2007). Again, other segmentation algorithms may be used. To save time we will limit our analysis to the first two samples from here on.

```
> norm.cghdata <- norm.cghdata[, 1:2]
> seg.cghdata <- segmentData(norm.cghdata, method = "DNAcopy")

Start data segmentation ..
Analyzing: Sample.1
Analyzing: Sample.2
```

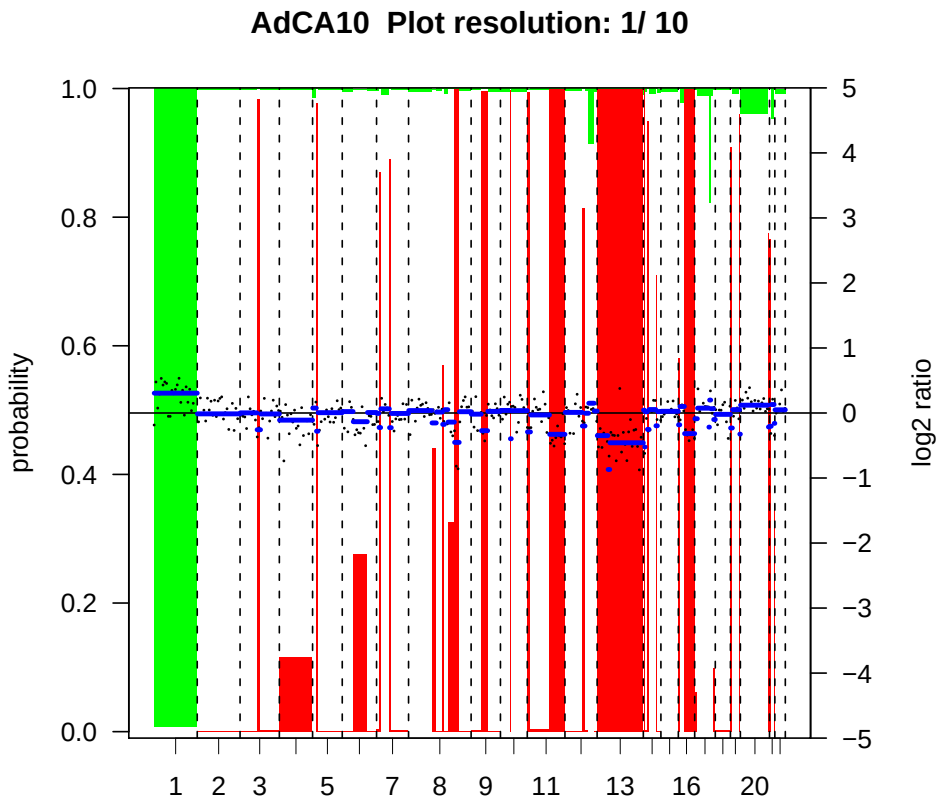
Now that the data have been normalized and segments have been defined, we need to determine which segments should be classified as losses, normal, gains or amplifications.

```
> result <- CGHcall(seg.cghdata)

[1] "changed"
EM algorithm started ...
[1] "Total number of segments present in the data: 118"
[1] "Number of segments used for fitting the model: 118"
Calling iteration 1 :
     j        rl       mudl       musl         mun       mug      mudg      mua
[1,] 2 -3775.002 -0.8566917 -0.2956446 0.005034546 0.3387235 0.5790516 1.079008
          sddl       sdsl        sdn       sdg      sddg       sda
[1,] 0.0935906 0.09305483 0.08015539 0.1525714 0.1528988 0.1528988
Calling iteration 2 :
     j        rl       mudl       musl         mun       mug      mudg      mua
[1,] 2 -3773.482 -0.8673733 -0.2937358 0.009585043 0.3376547 0.5772246 1.077181
           sddl       sdsl        sdn       sdg      sddg      sda
[1,] 0.08711862 0.08654279 0.07152661 0.1503949 0.1507270 0.150727
```

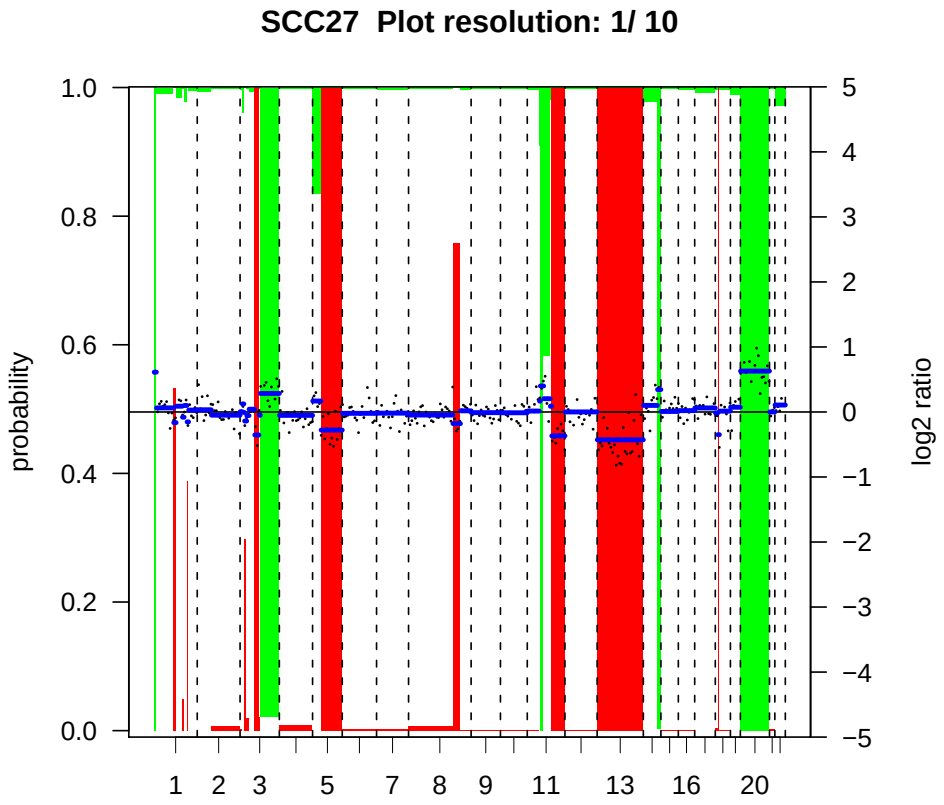
```
EM algorithm done ...
Computing posterior probabilities for all segments ...
FINISHED!
Total time: 1 minutes
```

To visualize the results per profile we use the `plotProfile` function:

```
> plot(result[, 1])

Plotting sample AdCA10
```

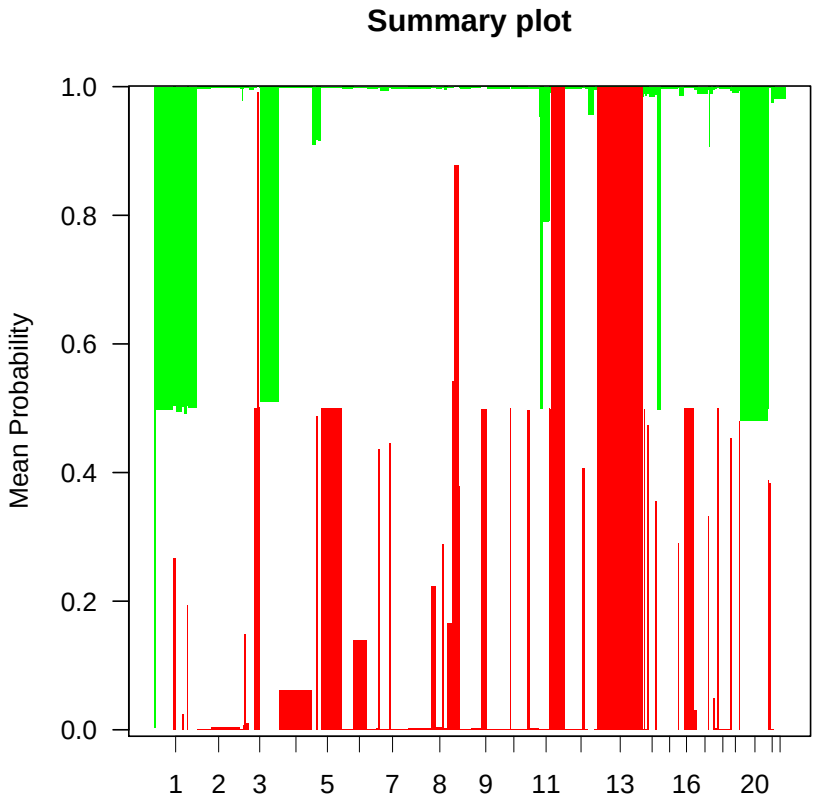


```
> plot(result[, 2])

Plotting sample SCC27
```



Alternatively, we can create a summary plot of all the samples:

```
> summaryPlot(result)

Adding sample AdCA10 to summary plot.
Adding sample SCC27 to summary plot.
```



References

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525.

van de Wiel, M. A., Kim, K. I., Vosse, S. J., van Wieringen, W. N., Wilting, S. M., and Ylstra, B. (2007). CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, 23:892–894.

Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23:657–663.

Wilting, S. M., Snijders, P. J. F., Meijer, G. A., Ylstra, B., van den Ijssel, P. R. L. A., Snijders, A. M., Albertson, D. G., Coffa, J., Schouten, J. P., van de Wiel, M. A., Meijer, C. J. L. M., and Steenbergen, R. D. M. (2006). Increased gene copy numbers at chromosome 20q are frequent in both squamous cell carcinomas and adenocarcinomas of the cervix. *J Pathol*, 209:220–230.