

DNACopy: A Package for Analyzing DNA Copy Data

E. S. Venkatraman and Adam B. Olshen

May 18, 2005

Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center

`olshena@mskcc.org`

Contents

1	Overview	1
2	Data	2
3	An Example	2
4	Tips	6

1 Overview

This document presents an overview of the **DNACopy** package. This package is for analyzing array DNA copy number data, which is usually (but not always) called array Comparative Genomic Hybridization (array CGH) data (Pinkel et al., 1998; Snijders et al., 2001; Lucito et al., 2003). It implements our methodology for finding change-points in these data (Olshen et al., 2004), which are points after which the (log) test over reference ratios have changed location. Our model is that the change-points correspond to positions where the underlying DNA copy number has changed. Therefore, change-points can be used to identify regions of gained and lost copy number. We also provide a function for making relevant plots of these data.

2 Data

We selected a subset of the data set presented in Snijders et al. (2001). We are calling this data set `coriell`. The data correspond to two array CGH studies of fibroblast cell strains. In particular, we chose the studies **GM05296** and **GM13330**. After selecting only the mapped data from chromosomes 1-22 and X, there are 2271 data points. There is accompanying spectral karyotype data (not included), which can serve as a gold standard. The data can be found at http://www.nature.com/ng/journal/v29/n3/supinfo/ng754_S1.html

3 An Example

Here we perform an analysis on the **GM05296** array CGH study described above.

```
> library(DNAcopy)
> data(coriell)
```

Before segmentation the data needs to be made into a CNA object.

```
> CNA.object <- CNA(cbind(coriell$Coriell.05296), coriell$Chromosome,
+   coriell$Position, data.type = "logratio", sampleid = "c05296")
```

We generally recommend smoothing single point outliers before analysis. It is a good idea to check that the smoothing is proper for a particular data set.

```
> smoothed.CNA.object <- smooth.CNA(CNA.object)
```

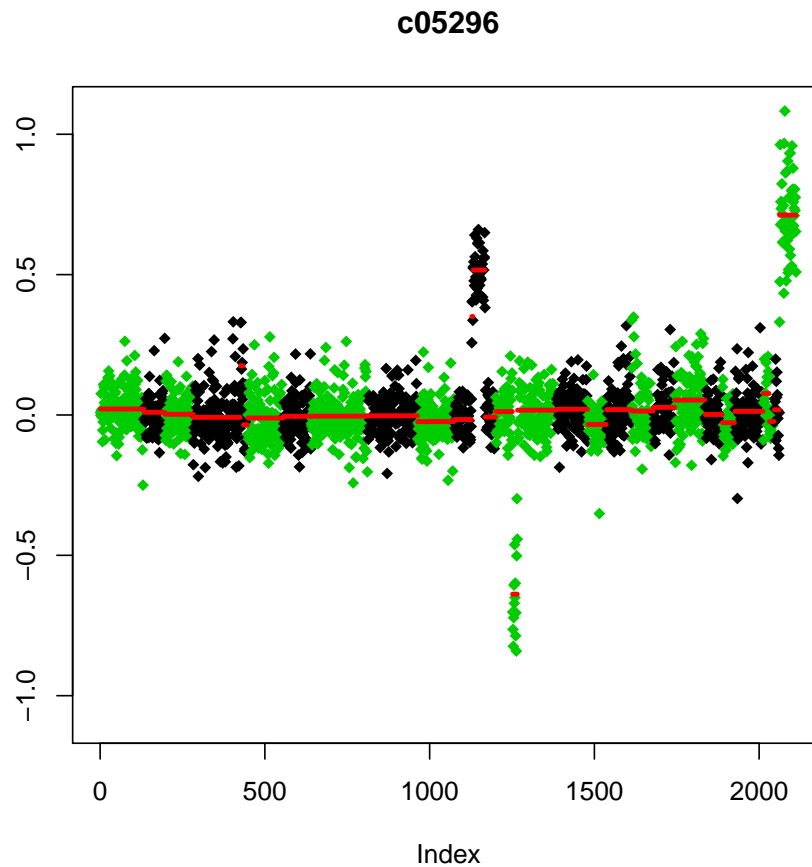
After smoothing, if necessary, the segmentation is run. Here the default parameters are used. A brief discussion of parameters that can be adjusted is in the Tips section.

```
> segment.smoothed.CNA.object <- segment(smoothed.CNA.object, verbose = 1)
```

Analyzing: c05296

There are a number of plots that can be made. The first is ordering the data by chromosome and map positions. The red lines correspond to mean values in segments. Note that the points are in alternate colors to indicate different chromosomes.

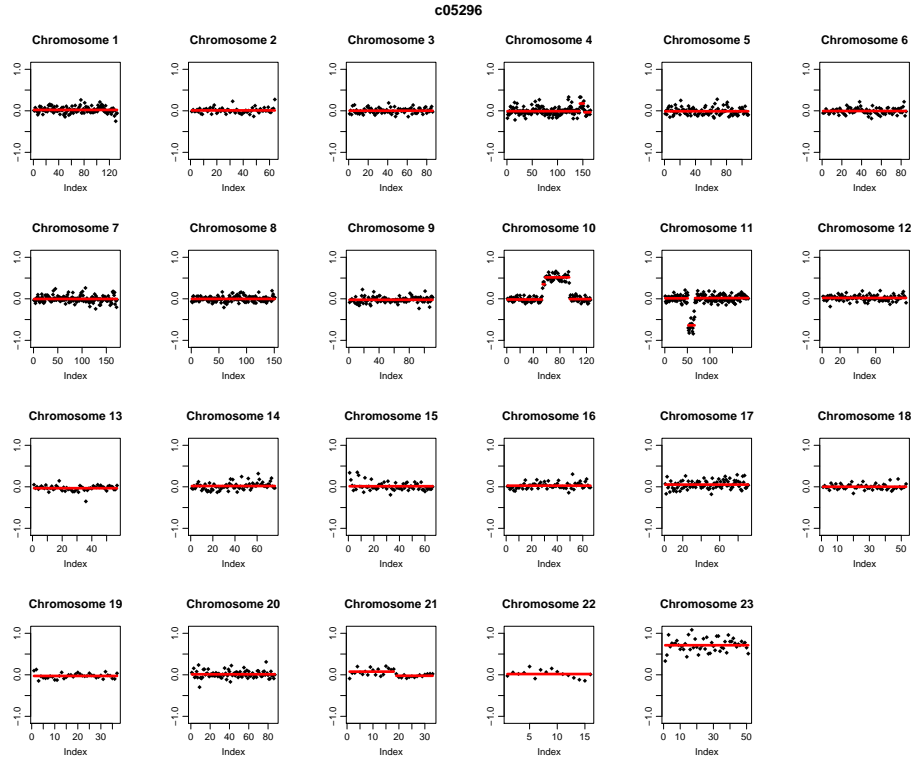
```
> plot(segment.smoothed.CNA.object, plot.type = "w")
```



Another possibility is to plot by chromosome within a study.

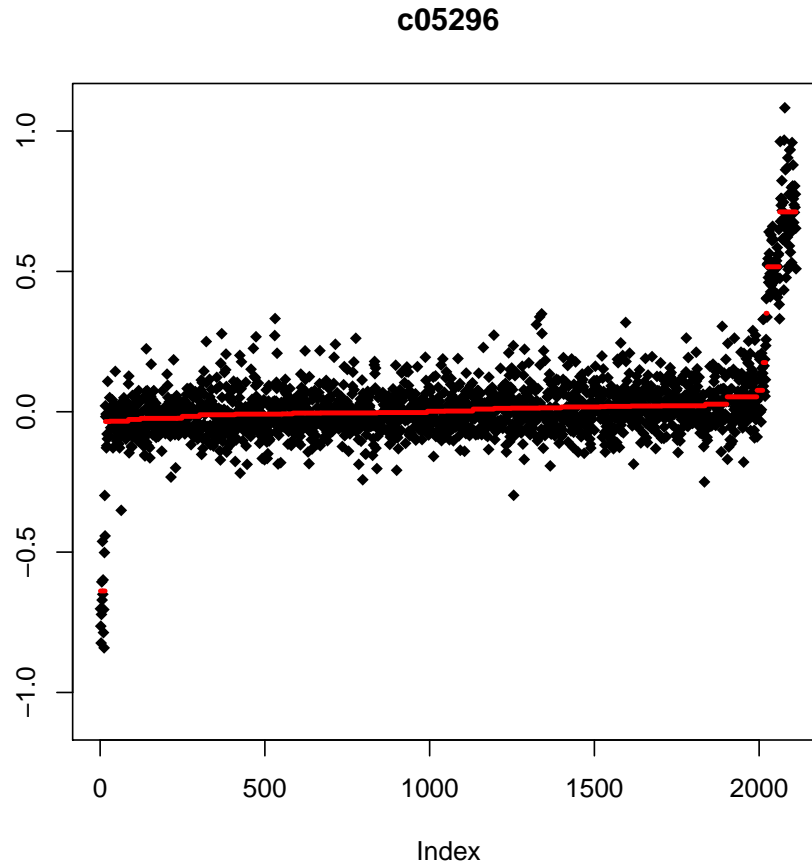
```
> plot(segment.smoothed.CNA.object, plot.type = "s")
```

Setting multi-figure configuration



If there are multiple studies, one could plot by chromosome across studies using the option `plot.type="c"`. A final plot orders the segment by their chromosome means. One can take the plateaus in this plot to determine what the mean values should be for calling segments gains or losses. In this case, maybe 0.4 for gains and -0.6 for losses. For most data, these plateaus are much closer to zero. The next generation of this software will have automatic methods for calling gains and losses.

```
> plot(segment.smoothed.CNA.object, plot.type = "p")
```



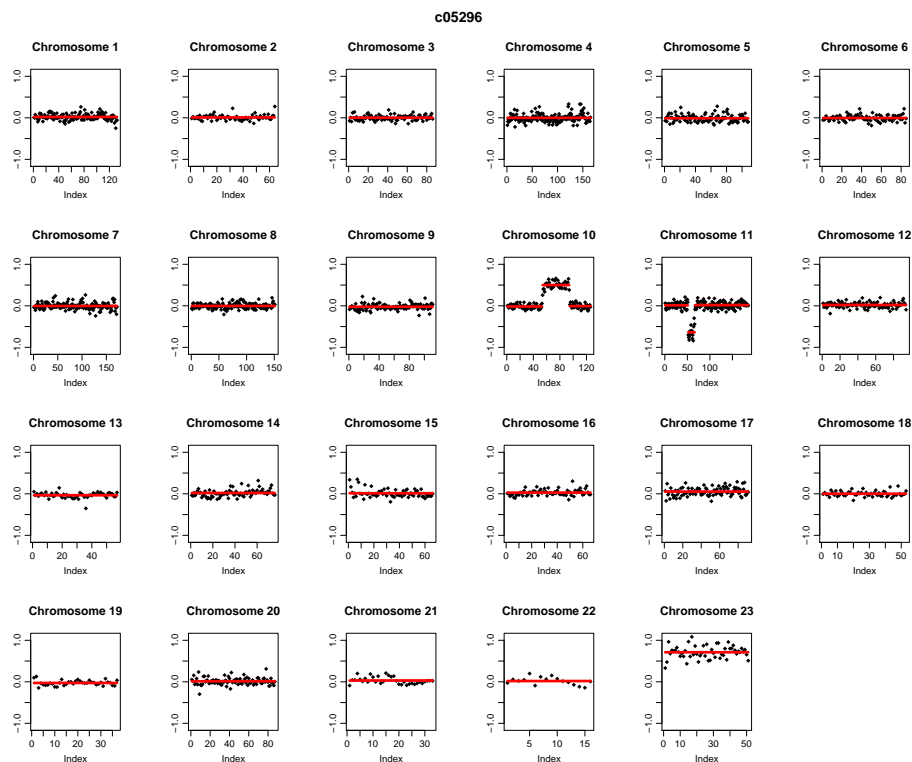
Change-points are often found due to local trends in the data. An undo method is needed to get rid of unnecessary change-points. Below all splits that are not at least three SDs apart are removed. The following plot shows that all splits not corresponding to the gold standard results have been removed.

```
> sdundo.CNA.object <- segment(smoothed.CNA.object, undo.splits = "sdundo",
+   undo.SD = 3, verbose = 1)
```

Analyzing: c05296

```
> plot(sdundo.CNA.object, plot.type = "s")
```

Setting multi-figure configuration



4 Tips

A functions that may be of interest that has not been mentioned is `subset.CNA`. It allows for subsetting of a CNA object by chromosome and sample so that segmentation does not have to be run on a whole data set. Similarly, `subset.DNACopy` allows subsetting of DNACopy objects, which contain the output of segmentation.

The parameter `alpha` controls the sensitivity of the segmentation. The default value is 0.01. Increasing it will lead to more change-points, and, possibly, more false positive change-points. Decreasing it will lead to fewer change-points. The parameter `nperm` controls the number of permutations. Decreasing it will cause the algorithm to go faster (default is 10000). However, the lower `alpha` is, the more permutations that are needed. Another way of speeding up computations is using the `window.size` parameter. In this case, maximization is performed on overlapping windows of a particular size rather than over the whole chromosome.

References

- Lucito, R., Healey, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J., Rostan, S., Nguyen, K., Powers, S., Ye, K., Olshen, A., Venkatraman, E., Norton, L., and Wigler, M. (2003). Representational oligonucleotide microarray analysis: a high resolution method to detect genome copy number variation. *Nat. Genet.*, 13:2291–305.
- Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5:557–72.
- Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W., and Albertson, D. G. (1998). High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20:207–211.
- Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., S, S. L., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D., and Albertson, D. G. (2001). Assembly of microarrays for genome-wide measurement of dna copy number. *Nat. Genet.*, 29:263–4.