Vega: Variational Segmentation for Copy Number Detection

Sandro Morganella Luigi Cerulo Giuseppe Viglietto
Michele Ceccarelli

Contents

1 Overview

This document describes classes and functions of Vega (Variational Estimator of Genomic Aberrations) package. Vega algorithm allows to segment copy number profiles from array comparative genomic hybridization (aCGH) data. This package implements the algorithm described in "VEGA: Variational segmentation for copy number detection" [**?**].

Here we show as users can use Vega to perform the copy number segmentation task. The Log R Ratio (LRR) data used in the presented examples concern the mantle cell lymphoma (MCL) Granta-519 cell line previously published by DeLeeuw *et al.* [**?**].

2 Installation

Install VEGA on your computer by using the following command:

```
source("http://bioconductor.org/biocLite.R")
biocLite("Vega")
```

3 Vega .RData Description

In Vega package you find the LRR data of the mantle cell lymphoma (MCL) Granta-519 cell line previously published by DeLeeuw *et al.*The data are obtained by using SMRT aCGH containing 97 299 elements, representing 32 433 overlapping genomic segments spanning the entire human genome [**?**].

Load Vega package:

```
> library(Vega)
```

Load Granta-519 data:

```
> data(G519)
```

G519 data is organized as a matrix having four columns:

- Chromosome

- Start position for the observed probe
- End position for the observed probe
- The measured LRR for the probe

```
> G519[1:16,]
```

```
      Chromosome Probe Start Position Probe End Position LRR
 [1,] "1"        "0"                 "0"                "0.02"
 [2,] "1"        "0"                 "0"                "0.59"
 [3,] "1"        "0"                 "0"                "0.09"
 [4,] "1"        "0"                 "0"                "-0.1"
 [5,] "1"        "0"                 "0"                "0.62"
 [6,] "1"        "0"                 "0"                "-0.08"
 [7,] "1"        "50617"             "217280"           "0.12"
 [8,] "1"        "722781"            "902347"           "0.17"
 [9,] "1"        "851036"            "1037875"          "-0.13"
[10,] "1"        "899476"            "1049818"          "0.2"
[11,] "1"        "946127"            "1096338"          "0.26"
[12,] "1"        "1154648"           "1320545"          "-0.09"
[13,] "1"        "1286206"           "1419769"          "0.34"
[14,] "1"        "1337829"           "1419769"          "0.36"
[15,] "1"        "1501818"           "1681097"          "0.55"
[16,] "1"        "1511518"           "1659697"          "0.45"
```

4 Run Vega

In order to run Vega algorithm the user needs to use the function `vega`. This function can be used passing only two arguments: the data (a matrix having the structure described in the previous section) and the list of chromosomes that will be analyzed (see Appendix **??** for more details). In order to run Vega on Granta-519 data use the following command:

```
> seg <- vega(CNVdata=G519, chromosomes=c(1:22,"X","Y"))
```

Note that the previous segmentation is computed considering all chromosomes of Granta-519 (`c(1:22,"X","Y")`).

If you want analyze only a subset of chromosomes you can conveniently set the argument `chromosomes`. For example if you want to run Vega algorithm on the chromosomes 1 and X you will use the following command

```
> seg_1X <- vega(CNVdata=G519, chromosomes=c(1,"X"))
```

The computed segmentation can be found in the R object `seg`:

```
> seg[1:5,]
```

```
     Chromosome bp Start   bp End     Num of Markers Mean                 Label
[1,] "1"        "0"        "3098099"  "36"           "0.218055555555556"  "1"
[2,] "1"        "3103888"  "9047388"  "66"           "-0.245"             "-1"
[3,] "1"        "9295827"  "10847085" "18"           "0.110555555555556"  "0"
[4,] "1"        "10809955" "33814729" "253"          "-0.210197628458498" "-1"
[5,] "1"        "33732317" "47694543" "152"          "0.0951973684210526" "0"
```

Analyzing the first line of the output, we can deduce that on the chromosome 1 the region from the bp 0 to the bp 3 098 099 (which contains 36 probes) has a LRR mean value of ≈ 0.218 and it is considered as a chromosomal gain.

In addition you can save the computed segmentation into a tab delimited file by setting the argument `out_file_name` with the name of the file (for example segmentation.txt). For more detail about this file see the next subsection (**??**). Other two parameters can be chosen by the user: `beta` and `min_region_size` for more details on these parameters see Appendix **??**.

4.1 Output File Format

If you want to save the computed segmentation into a tab delimited file you can run VEGA with the additional parameter `out_file_name`. This file has a row for each segmented region and for each of them it has seven features (columns of the file):

Chromosome: the chromosome containing the region

bp Start: the genomic start position (in bp) of the region

bp End: the genomic end position (in bp) of the region

Num of Markers: the number of markers contained in the region

Mean: the mean value of the LRR of all probes contained in the region

Label: indicates the computed label of the region: loss (-1), normal (0) and gain (1).

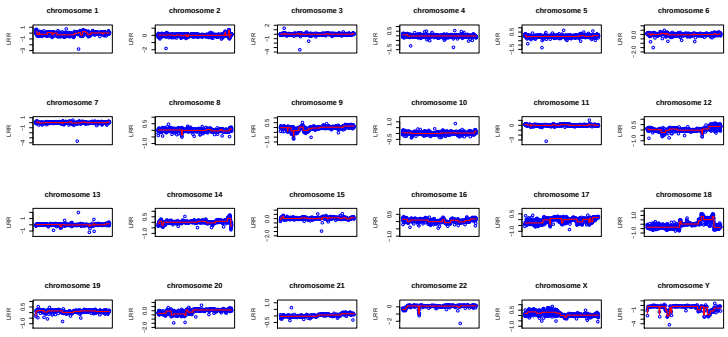
5 Segmentation Plot Utility

In Vega a plot utility called `plotSegmentation` is provided to show an overview combining LRR data and computed segmentation. The user can specify both the chromosomes and the segmentation informations that have to be shown. In particular if the user wants to plot the LRR mean value of each region he needs to set the argument `opt=0` (which is the default value). In contrast if the user wants to plot the computed label of each region (loss, normal and gain) he needs to set `opt=1`. In the following of this section some plot examples are provided.

5.1 Plot All Chromosomes with the LRR Mean Values

The next commands allows to plot an overview of all analyzed chromosomes (`chromosomes=c(1:22,"X","Y")`) in which the LRR mean values are reported (`opt=0`).

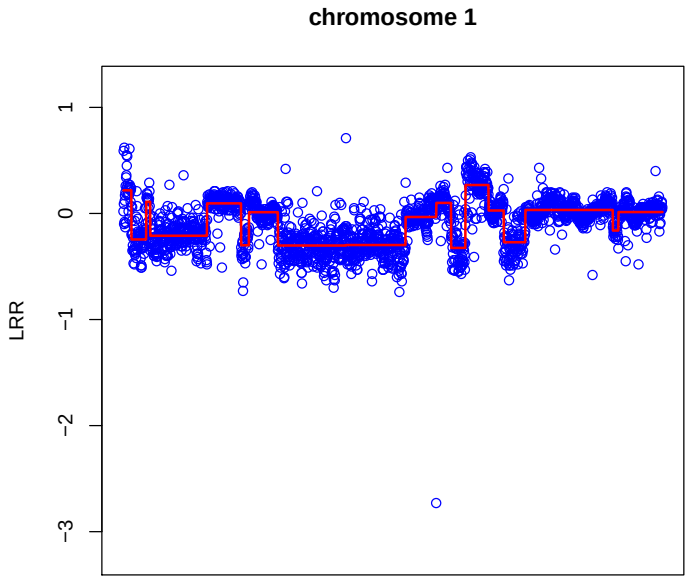
```
> plotSegmentation(G519, seg, chromosomes=c(1:22,"X","Y"), opt=0)
```



5.2 Plot Chromosome 1 with the Respective LRR Mean Values

The next command plots only the chromosome 1 (`chromosomes=c(1)`) with the LRR mean value (`opt=0`).

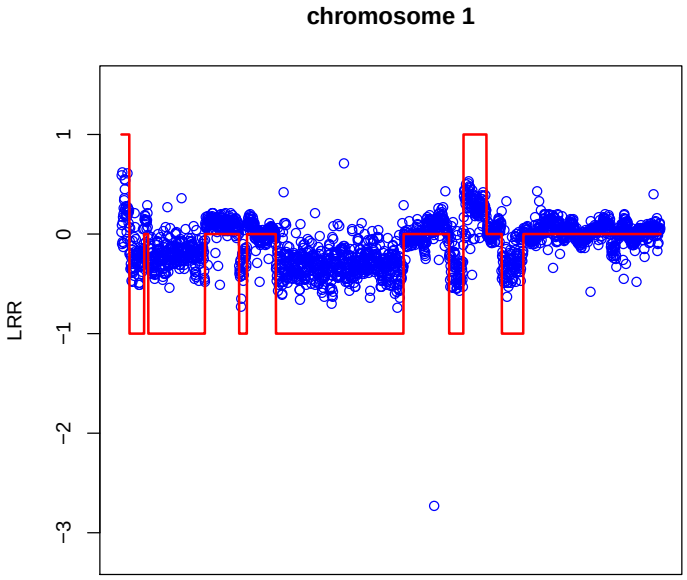
```
> plotSegmentation(G519, seg, chromosomes=c(1), opt=0)
```



5.3 Plot Chromosome 1 with the Respective Labels

The next command plots only the chromosome 1 (`chromosomes=c(1)`) with the label computed for each region (`opt=1`)

```
> plotSegmentation(G519, seg, chromosomes=c(1), opt=1)
```



From this plot we can notice that on the chromosome 1 we find 8 mutations (2 gains and 6 losses).

A Vega: Function Description

A.1 `vega`

The function `vega` computes the segmentation on the aCGH data passed as argument. The header of `vega` follows:

```
vega(CNVdata, chromosomes, out_file_name="segmentation.txt",
     beta=0.5, min_region_size=2)
```

`CNVdata`: This argument is a matrix containing all informations about the observations: Chromosome, Probe Start Position, Probe End Position and LRR. For more details see Section **??**.

`chromosomes`: This argument is used to list the chromosomes that have to be processed. By using `c(1:22, "X", "Y")` all chromosomes will be segmented.

`out_file_name`: (default value:"") This name is used to save the computed segmentation into a tab delimited file. If the default value is used no file will be saved. For more details see Section **??**.

`beta`: (default value 0.5) This argument is used to define the stop condition of Vega algorithm (see [**?**] for more details).

`min_region_size:` (default value 2) This argument specifies the minimum size allowed for the segmented regions.

A.2 `plotSegmentation`

The function `plotSegmentation` plots observations and segmentation results. the header of `plotSegmentation` function follows:

```
plotSegmentation(CNVdata, segmentation, chromosomes, opt = 0)
```

`CNVdata:` This argument specifies the matrix containing all informations about the observations: Chromosome, Probe Start Position, Probe End Position and LRR. For more details see Section **??**.

`segmentation:` This argument is the segmentation computed by `vega` function on the observations contained in `CNVdata`.

`chromosomes:` This argument is used to list the chromosomes that have to be plotted. By using `c(1:22, "X", "Y")` all chromosomes will be plotted.

`opt:` (default value 0) This argument is used to choose the segmentation informations that have to be plotted. If `opt=0` then the LRR mean value of each segmented region is shown. If `opt=1` the label of each region is shown where levels -1, 0 and 1 are associated with loss, normal and gain respectively.

References

[1] DeLeeuw RJ. *et al* (2004). Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes, *Human Molecular Genetics* **13**(17):1827-1837.

[2] Ishkanian AS. *et al.* (2004). A tiling resolution DNA microarray with complete coverage of the human genome, *Nature Genetics* **36**:299-303.

[3] Morganella S. *et al.* (2010).VEGA: Variational segmentation for copy number detection, *Bioinformatics.*