VegaMC: A Package Implementing a Variational Piecewise Smooth Model for Identification of Driver Chromosomal Imbalances in Cancer

Sandro Morganella Michele Ceccarelli

Contents

1 Overview

VegaMC enables the detection of driver chromosomal imbalances (deletion, amplification and loss of heterozygosity (LOH)) from array comparative genomic hybridization (aCGH) data. VegaMC performs a joint segmentation of aCGH data coming from a cohort of disease affected patients. Segmented regions are used into a statistical framework to distinguish between driver and passenger mutations. In this way, significant imbalances can be detected by the associated p-values. A very interesting feature of VegaMC, is that, it has been implemented to be easily integrated with the output produced by PennCNV tool [1]. PennCNV is a widely used tool in Bioinformatics, it analyzes raw files and produces for each probe the respective value of Log R Ratio (LRR) and B Allele Frequency (BAF). In addition, VegaMC produces in output two web pages allowing a rapid navigation between both detected regions and altered genes. In the web page that summarizes the altered genes, the user finds the link to the respective Ensembl gene web page. An accurate implementation of the algorithm allows the accurate and rapid detection of significant chromosomal imbalances.

Below we lists the main interesting features of VegaMC:

- ◇ Designed to be integrated within PennCNV protocol
- ◇ Detection of LOH alterations
- ◇ Two web pages enable an easily and rapid navigation of both detected regions and altered genes
- ◇ Accurate analysis of large datasets in a short time

For More details on the usage of VegaMC visit the home page of the package:

`http://bioinformatics.biogem.it/download/vegamc/vegamc`

2 Installation and Dependencies

In order to install VegaMC from Bioconductor repository start R and enter:

```
> if (!requireNamespace("BiocManager", quietly=TRUE))
    > install.packages("BiocManager")
> BiocManager::install("VegaMC")
```

You can load all functions of VegaMC by entering:

```
> library(VegaMC)
```

In order to download gene information, VegaMC need of biomaRt that provides an interface to BioMart databases (e.g. Ensembl, COSMIC ,Wormbase and Gramene). In addition VegaMC is compatible with the genoset package that offers an extension of the Bioconductor `eSet` object for genome arrays.

3 Input Data Format

The main input of VegaMC is the dataset file that has a row for each probe. The first three columns of the file specify:

- ◇ The probe name
- ◇ The chromosome in which the probe is located
- ◇ The genomic position of the probe

The other columns of the matrix report the LRRs observed for the probe (and optionally the BAFs). **Note that this format reflects the format of PennCNV**. An example of input file can be found in `inst/template` folder (breast_Affy500K.txt).

Note that the probes must be ordered by the respective genomic positions within the chromosome. In some cases (as for PennCNV) data are not sorted. VegaMC provides a function that performs this sorting: `sortData` (see Section 6).

In VegaMC the user can find an example dataset (breast_Affy500K.txt) composed of 10 breast tumor aCGH samples profiled by high-resolution Affymetrix 500K Mapping Array (GEO accession GSE7545) [2]. Raw data were preprocessed in accord to PennCNV protocol and both LRR and BAF were obtained. For space requirements only the observation for the first 4 chromosomes are reported, the complete dataset is available at:

`http://bioinformatics.biogem.it/download/vegamc/vegamc`

In order to copy the example dataset in the current work directory enter:

```
> file.copy(system.file("example/breast_Affy500K.txt",
+     package="VegaMC"),".")
```

```
[1] TRUE
```

Note that LRR and BAF of a sample must be reported in this order: LRR followed by BAF. breast_Affy500K.txt agrees with this format.

4 VegaMC

VegaMC analysis is composed by three different steps: segmentation of the dataset, classification of the regions, assessment of statistical significance for each region. In the next sections all steps are described providing also a description of their main specific input parameters.

4.1 Segmentation

VegaMC performs a joint segmentation of all samples to detect the regions that have a similar LRR profile among the samples. In order to perform this joint segmentation, VegaMC extends an algorithm based on a popular variational model [3]. In particular, VegaMC implements the weighted multi-channel version of this model. Segmentation is separately performed on each chromosome and results are strictly related to the so called scale parameter: as the scale grows the segmentation gets coarser. In VegaMC a data-driven approach is used to compute the optimal scale value: VegaMC computes the jump of the scale required to merge two adjacent regions (parameter `beta` of VegaMC). if the allowed jump (`beta`) is 0 then the computed segmentation will be composed of N regions (each region will contain just a probe). In contrast, if the allowed jump is very large (ideally ∞) then the segmentation will contain just a region for each chromosome.

4.2 Classification

Aim of classification step is the labeling of each detected region. For each detected region and for each sample, classification looks for deletions, amplifications and LOHs (if BAF is observed). This step is performed by a threshold-based approach.

4.2.1 Classification of Deletions and Amplifications

In order to distinguish between deleted and amplified regions, two thresholds are used. Given the region i of the sample j then:

- ◇ the region is considered as a deletion if$||\mu_{ij}|| <$ `loss_threshold`
- ◇ the region is considered as an amplification if $||\mu_{ij}|| >$ `gain_threshold`

where $||\mu_{ij}||$ is the respective mean.

4.2.2 Classification of LOHs

LOH is receiving greater attention as a mechanism of possible tumor initiation. LOH is located in region with no copy number alteration (LRR= 0) and it is characterized by a BAF trend that moves away from the value of 0.5 corresponding to heterozygous genotype AB.

In order to detect LOH regions, two parameters are used:

- ◇ `loh_threshold`: BAF values out of the range:
  ]`loh_threshold`,(1−`loh_threshold`)[
  are considered to belong to homozygous genotypes AA and BB.
- ◇ `loh_frequency`: Minimum fraction of homozygous genotypes needed for marking a region as LOH.

In other words, given the region i of the sample j, the region is considered to be a LOH if:

$$\Big(\sum_{ij}(BAF) \notin \ ]\texttt{loh\_threshold}, (1 - \texttt{loh\_threshold})[\ \Big) > \texttt{loh\_frequency} \tag{1}$$

Note that in order to detect LOHs, the dataset must contain the BAF associated to each observed probe. This is the case of the example file in `/inst/template` folder breast_Affy500K.txt

4.3 Assessment of Statistical Significance

The main problem in identification of chromosomal imbalances is the distinction between driver and passenger mutations. Driver alterations are functionally related to the disease under study, while, passenger alterations are subject-specific random somatic mutations. Aim of this step is the computation of the p-value

associated to each segmented region, which provides the evidence for driver mutations. VegaMC assigns a p-value to each possible aberration: loss (deletion), gain (amplification) and LOH. Here we use an approach similar to the one described in [4]. Starting from a discretized representation of the data obtained by the classification step, VegaMC performs a conservative permutation test to obtain the null distribution associated to the hypothesis: *All observed aberrations are passengers.*

5 VegaMC: Run Analysis

In order to perform the analysis, the user needs to call only the function `vegaMC`. Below, the list of the input argument of this function are listed:

- ◇ `dataset`: The file containing the observations for dataset under inspection. The first three columns describe the name, the chromosome and the position respectively. The other columns of the matrix report the LRR and the BAF (if available) of each sample.
- ◇ `output_file_name`: (Default `output`) The file name used to save the results.
- ◇ `beta`: (Default 0.5) This parameter is used in the segmentation step to compute the stop condition. This parameter specifies the maximum jump allowed for the updating of the scale parameter. If `beta`=0 then the computed segmentation will be composed of a region for each probe (all regions will contain just a probe). In contrast, if `beta`$\rightarrow \infty$) then the segmentation will contain just a region for each chromosome.
- ◇ `min_region_bp_size`: (Default 1000) VegaMC does not report the regions short then this size (in bp). Note that this is only an output parameter, indeed, VegaMC also uses the "short" regions in all algorithmic steps.
- ◇ `classification` : (Default `FALSE`) If this value is TRUE multiple testing corrections is performed.
- ◇ `loss_threshold`: (Default -0.2) Value used to mark a region as a deletion (loss). If the wighted mean of a region $||\mu_{ij}||$ is lower than this threshold, then the region is marked as a deletion (loss).
- ◇ `loss_threshold`: (Default 0.2) Value used to mark a region as an amplification (gain). If the wighted mean of a region $||\mu_{ij}||$ is greater than this threshold, then the region is marked as an amplification (gain).
- ◇ `baf` : (Default `TRUE`) This parameter specifies if the dataset also contains BAF measurements. `baf`=`TRUE` means that BAF measurements are available, in this case VegaMC is able to compute LOH imbalances. If `baf`=`FALSE` the dataset only contains the LRR values and LOH detection is not possible.
- ◇ `loh_threshold` : (Default 0.75) Threshold used to distinguish between homozygous and heterozygous genotypes. If the BAF is greater than `loh_threshold` or lower then (1−`loh_threshold`) then the respective probe is considered to be homozygous.

- ◇ `loh_frequency` : (Default 0.8) Minimum fraction of homozygous probes needed for marking a region as LOHs. Regions with a fraction of homozygous probes greater than this threshold are marked as LOH.
- ◇ `bs` : (Default 1000) Number of permutation bootstraps performed to compute the null distribution.
- ◇ `pval_threshold` : (Default 0.05) Significance level used to reject the null hypothesis. If the p-value of an aberration (loss, gain, LOH) is not greater than this threshold, then the region is considered to be significant and, consequently, it is considered a driver mutation.
- ◇ `html` : (Default `TRUE`) If this value is `TRUE`, then in output will be produced an html file called `output_file_name`.html in which a summary of all detected regions is reported.
- ◇ `getGenes` : (Default `TRUE`) If this value is `TRUE`, then in output will be produced an html file called `output_file_name`Genes.html in which the list of all genes overlapping the significant regions is reported.
- ◇ `mart_database`(Default `ensembl`) BioMart database name you want to connect to. Possible database names can be retrieved with the function listMarts of biomaRt package.
- ◇ `ensembl_dataset` : (Default `hsapiens_gene_ensembl`) BioMart dataset used to get information from Ensembl BioMart database.

The following command performs the analysis on the breast dataset with default parameter settings:

```
> results <-
+ vegaMC("breast_Affy500K.txt",
+         output_file_name="breast.analysis.default");
```

The next command performs the analysis with the following user-defined parameter setting:

- ◇ Increasing of the jump for the update of the scale parameter (`beta`) from default value of 0.5 to 1
- ◇ Removing regions shorter than 2000 bp (`min_region_pb_size`)
- ◇ Increasing the number of permutation bootstraps (`bs`) from the default value of 1000 to 5000
- ◇ Generation of html pages disabled

```
> results <-
+ vegaMC("breast_Affy500K.txt",
+         output_file_name="breast.analysis",
+         beta=1,        min_region_bp_size=2000, bs=5000,
+         html=FALSE, getGenes=FALSE)
```

Output The function `vegaMC` returns a matrix with a row for each detected region. In addition this command also creates in the current folder the tab delimited file breast.analysis.default (reporting the matrix content of the matrix `results`), the html file breast.analysis.default.html (which provides both a summary of the input parameter setting and all detected regions) and the html file file breast.analysis.defaultGenes.html (which lists the genes overlapping the significant regions).

5.1 VegaMC: View Results

After the execution of the previous command, the object `results` contains all information on the detected regions. This object is a matrix having a row for each detected regions. The name of the columns can be displayed by using the following command:

```
> colnames(results)
```

- ◇ `Chromosome`: The chromosome of the region
- ◇ `bp Start`: The position in which the region starts (in bp)
- ◇ `bp End`: The position in which the region ends (in bp)
- ◇ `Region Size`: The size of the region (in bp)
- ◇ `Mean`: The weighted mean $||\mu_{ij}||$ of the region
- ◇ `Loss p-value`: The p-value associated to the probability to have a driver deletion
- ◇ `Gain p-value`: The p-value associated to the probability to have a driver amplification
- ◇ `LOH p-value`: The p-value associated to the probability to have a driver LOH
- ◇ `% Loss`: The percentage of sample showing a deletion for this region
- ◇ `% Gain`: The percentage of sample showing an amplification for this region
- ◇ `% LOH`: The percentage of sample showing a LOH for this region
- ◇ `Probe Size`: The number of probes composing the region
- ◇ `Loss Mean`:Mean of LRR computed only on the samples that show a loss
- ◇ `Gain Mean`: Mean of LRR computed only on the samples that show a gain
- ◇ `LOH Mean`: Mean of LRR computed only on the samples that show a LOH
- ◇ `Focal-score Loss`: Focal Score associated to deletion
- ◇ `Focal-score Gain`:Focal Score associated to amplification
- ◇ `Focal-score LOH`: Focal Score associated to LOH

This matrix is automatically saved in the current work directory as a tab delimited file. For default the name used to save the file is `output`.

In addition, VegaMC creates two html pages that can be visualized by a web browser. These web pages are created to provide a rapid and easy access to the produced results. For default, generation of html page is active and the file `output`.html and are `output`Genes.html are created in the current work directory. These pages are created when options html `getGenes` are `TRUE` (default values).

output.html Web Page This page reports a summary of the analysis and it is composed of several sections that can be easily navigated by using the link on the top of the page. The first section is **Summary of Input Parameters**, it contains the parameter setting used for the analysis. The second section is **Results Summary**, it reports the number of identified regions and the number of each specific mutation. The third section is **List of All Regions** in which the information contained in the matrix `results` are summarized. The next sections report the list of significant regions detected for each imbalances (deletion, amplification and LOH). Therefore, in each section we only find the regions that have a p-value lower than the specified threshold (0.05 for default). Tables can be sorted by column.

outputGenes.html Web Page This page shows all genes located into a significant region. For each gene a set of information is reported. The **Ensembl Gene ID** and the **External Gene ID** report the name of gene. By clicking on the Ensembl Gene ID the user is automatically redirected on the Ensembl web page of the gene. The genomic position of the gene and the respective **Strand** and **Cytoband** are also reported. In addition the table also contains a column with a description of the gene (if available in Ensembl BioMart database). The last column reports the p-value associated to the region overlapping the gene. Tables can sorted by column.

6 Analysis of Gastrointestinal Stromal Tumors (GIST) Dataset

From the home page of VegaMC you can download a tab delimited file (gist.lrr_baf.txt) containing LRR and BAF for the dataset GISTs:

```
http://bioinformatics.biogem.it/download/vegamc/vegamc
```

GISTs data were published in [5] where 25 fresh tissue specimens of GISTs were collected and hybridized by Affymetrix Genome Wide SNP 6.0 (GEO identifier GSE20710). Raw data were preprocessed by PennCNV tool and the available file is the output of PennCNV which contains observation for $\sim$ 1.6 million of probes.

The file produced by PennCNV is not sorted by the genomic position of the probe. VegaMC provides a function that sort the data. The function `sortData`

takes in input a matrix and returns the matrix ordered by the genomic position. The arguments of this function follow:

- ◇ `dataset`: The matrix that have to be ordered
- ◇ `output_file_name`: The name used to save the results into a tab delimited file

Given that `gist` data is directly produced by PennCNV, default parameter setting can be used::

```
> sortData("gist.txt", "gist.sorted.txt")
```

Now we are ready to run VegaMC. The following command performs a segmentation in which each segmented region must contain at least 5 probes and the resulting list of regions contains only the regions having size greater than 1000 bp:

```
> gist.results <-
+     vegaMC("gist.sorted.txt",
+             output_file_name="gist.analysis.default",
+             min_region_bp_size=1000)
```

Execution time of VegaMC on gist dataset is about 10 minutes. This execution time is very interesting, given that gist dataset is not a toy example ($\sim$ 1.6 million of probes on each of the 25 samples). Execution time was measured on a Mac OS X with 4GB of RAM and CPU2×2.8 GHz Quad-Core Intel Xeon.

References

[1] PennCNV: A free software tool for Copy Number Variation (CNV) detection from SNP genotyping arrays. `http://www.openbioinformatics.org/penncnv`.

[2] Haverty PM. *et al.* (2008). High-resolution genomic and expression analyses of copy number alterations in breast tumors. *Genes Chromosomes Cancer* **47**(6):530-542.

[3] Morganella S. *et al.* (2010).VEGA: Variational segmentation for copy number detection. *Bioinformatics* **26**(24):3020-3027.

[4] Morganella S. *et al.* (2011). Finding recurrent copy number alterations preserving within-sample homogeneity. *Bioinformatics* **27**(21):2949-2956.

[5] Astolfi A. *et al.* (2010). Molecular portrait of gastrointestinal stromal tumors: an integrative analysis of gene expression profiling and high-resolution genomic copy number. *Laboratory Investigation* **90**(9):1285-1294.