MPAgenomics: An R package for multi-patients analysis of genomic markers

Quentin Grimonprez, Alain Celisse, Meyling Cheok, Martin Figeac, Guillemette Marot

MPAgenomics : An R package for multi-patients analysis of genomic markers

Quentin Grimonprez [1,*], Alain Celisse [1,2], Meyling Cheok[3], Martin Figeac[4] and Guillemette Marot [1,5]

January 20, 2014

[1] Modal team, Inria Lille-Nord Europe, France
[2] Laboratoire Paul Painlevé, Université Lille 1, France
[3] Inserm, U837, Team 3, Cancer Research Institute of Lille, France
[4] Plate-forme de génomique fonctionnelle et structurale, IFR-114, Université Lille 2, France
[5] EA 2694, Université Lille 2, France

Abstract

Summary: `MPAgenomics`, standing for multi-patients analysis (MPA) of genomic markers, is an R-package devoted to: (*i*) efficient segmentation, and (*ii*) genomic marker selection from multi-patient copy number and SNP data profiles. It provides wrappers from commonly used packages to facilitate their repeated (sometimes difficult) use, offering an easy-to-use pipeline for beginners in R.

The segmentation of successive multiple profiles (finding losses and gains) is based on a new automatic choice of influential parameters since default ones were misleading in the original packages. Considering multiple profiles in the same time, `MPAgenomics` wraps efficient penalized regression methods to select relevant markers associated with a given response.

Availability: The R-package `MPAgenomics` is available on R-forge at http://r-forge.r-project.org/R/?group_id=1658.

Contact: quentin.grimonprez@inria.fr

1 Introduction

Analyzing data from genome-wide SNP arrays within R requires several packages, e.g. `aroma` for normalization of Affymetrix SNP6.0 arrays [2], `changepoint` for segmentation of copy number profiles [7], `cghcall` for labelling segments [10], and `glmnet` for penalized regressions [5]. Each package performs a specific task in the whole analysis but none of them is related to the others. Output formats of given packages are often not compatible with input formats required by the other, making their use awkward for beginners in R. One main contribution of the `MPAgenomics` R package is to aggregate these commonly used packages, providing wrappers to inter-relate them automatically.

At each step of the analysis a large amount of packages are available to perform normalization, segmentation or marker selection. A careful choice of only a few methods is required to provide an easy-to-use and efficient tool. For instance the `MPAgenomics` segmentation step is based on

*to whom correspondence should be addressed

PELT [7], which has been proved to be the most reliable method to segment copy number profiles among 17 competitors [6]. Furthermore the `MPAgenomics` package improves on the native PELT method by providing an automatic data-driven choice of its penalty parameter.

In this application note, we describe two different pipelines implemented in the R package `MPAgenomics`. Both of them perform the whole analysis from raw data to normalization, and then to either successive segmented profiles, or a list of genomic markers selected from all available profiles.

2 MPAgenomics package

One interest of `MPAgenomics` is to provide a simple automatic way to combine the elementary steps described in what follows. Each of them can however be used separately by more advanced users.

2.1 Data normalization

The normalization process in `MPAgenomics` contains *technical biaises correction*, and *copy number and allele B fraction estimation*. Following [3], *allele B fraction* refers to the proportion of the total signal coming from allele B. Normalization methods are available for Affymetrix arrays (GenomeWideSNP 5 & 6, ...). The estimation of the total copy number and allele B fraction is made by *CRMAv2* [4] originally implemented in the `aroma` packages. For studies with matched normal-tumor samples, a better estimation is suggested for the allele B fraction of the tumoral sample with the TumorBoost method [3].

The use of `aroma` packages is difficult for neophytes due to the complex folder architecture it requires and the lack of internal documentation of the R package. `MPAgenomics` implements a wrapper to process all these normalization steps and build the folder architecture automatically. Furthermore different graphs such as the copy number signal can be saved in the working folder architecture for further visualization.

2.2 Segmentation method

Following [6] the PELT segmentation method [7] is implemented in `MPAgenomics`. It relies on a penalty $\lambda \log(n)$ penalizing too many segments, with a profile of length n and λ a parameter to choose. We observed that using the default parameter $\lambda = 1$ on a real dataset of 70 profiles [8] leads to over-segmented regions (too many segments). Since λ is crucial, `MPAgenomics` suggests an automatic sample-specific choice of λ (see Section 3).

The implemented segmentation method is available both for copy number and allele B fraction profiles. The allele B fraction segmentation is only made from heterozygous SNPs. The resulting signal is centered and symmetrized around 0, which makes it similar to the usual copy number.

2.3 Calling method

From each segmented profile, the *CGHcall* method [10] is run to label every segment in terms of *loss*, *normal*, and *gain*. Segmentation and labeled segments are available in two settings. One is aroma-based and exploits the folder architecture and the files generated along the process. The second does not depend on aroma. It is particularly relevant for more advanced users with their own normalized data.

In the aroma-dependent function, segmentation and calling are performed with the same wrapper. The calling is run for each profile separately. Results are saved in *.BED* format in the working folder architecture.

2.4 Genomic marker selection

The goal is to select genomic markers (e.g. SNPs or CNV) associated with a given response from all patient profiles simultaneously.

For each individual i ($1 \leq i \leq I$), y_i denotes the response and $x_{i,p}$ the corresponding normalized value of copy number or allele B fraction signal at genomic position p ($1 \leq p \leq P$). By default normalization is done as in Section 2.1 without between-array normalization. There is no need to perform segmentation and calling before the multi-patients analysis.

Due to the huge number of markers ($P \gg I$) `MPAgenomics` uses the *lasso* [9] regularization method to select very few ones. It consists in minimizing $\beta \in \mathbb{R}^P \mapsto g(\beta)$, where

$$g_\rho(\beta) = \sum_{i=1}^{I} (y_i - (X\beta)_i)^2 + \rho \sum_{p=1}^{P} |\beta_p| \ ,$$

with $(X\beta)_i = \sum_p x_{i,p}\beta_p$ and $\rho > 0$ controlling the number of non-zero coordinates of β. After minimization, non-zero coefficients β_p correspond to influential positions to predict the response.

`MPAgenomics` drastically improves on ongoing packages in terms of computation time. With the linear regression model, it efficiently provides the exact solution by use of the new R package `HDPenReg`, which is an optimized implementation of the *lars* algorithm specially dedicated to huge number of markers. Logistic regression is also available with binary responses. `MPAgenomics` wraps the `glmnet` package [5] in the whole process. Unlike `HDPenReg` it does not provide the exact solution but is computationally very efficient.

With `glmnet` and `HDPenReg`, the regularization parameter ρ is chosen by k-folds cross-validation [1]. The selected variables are the most relevant ones regarding to the response.

2.5 MPAgenomics vignette

The tutorial of the `MPAgenomics` R package is obtained by running the following commands in the R console:

```
library(MPAgenomics)
vignette("MPAgenomics")
```

An example explains how to quickly analyze data. More details on each step or wrapper are given to help advanced users to run each function separately.

3 Sample-specific parameter in MPA

First we detail the *sample-specific* choice of the λ parameter we propose in the PELT method (segmentation). Then we illustrate its potential improvement upon a common parameter choice on real data from [8].

3.1 Proposed method

For each profile the PELT method is run on a grid of λ values. Fig. 1 displays the number of segments with respects to λ on an example. The widest range of λ for which the number

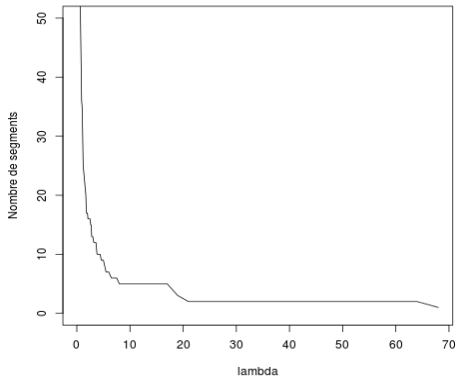

Figure 1: Number of segments for each λ in the penalty of PELT.

of segments remains unchanged (and larger than 1) indicates a high confidence in resulting segmentation.

The optimal data-driven λ is the left-most value of the widest range such that the number of segments is larger than 1. Otherwise we consider there is only one segment in the profile.

3.2 Sample-specific versus common parameter

The *sample-specific choice of* λ in Section 3.1 is compared with a *common choice of* λ depending on the signal-to-noise ratio within each group of profiles.

Profiles from the real dataset [8] are clustered into groups with homogeneous signal-to-noise ratio (SNR) by use of Gaussian mixture model. Three groups are provided by the BIC criterion. Figure 2 displays results (chromosome 1) for each profile (patient) from 1 to 70. Following Section 3.1 the widest range of λ values is plotted for each profile. Colors (black, red, and green) indicate the SNR level in each group (respectively low, middle, and high).

Whereas the lowest SNR group only contains ranges of λ with small values, other groups correspond to ranges of both small and large values of λ. A common choice of λ within each of these two groups lead to erroneous segmentations. The same conclusion applies to other chromosomes and criteria such as variance, which justifies the implementation of the sample-specific choice of λ in `MPAgenomics`.

4 Conclusion

`MPAgenomics` provides user-friendly wrappers for normalization and multi-patients analysis of genomic data. It also provides automatic choices of crucial parameters for segmentation and marker selection. Even though normalization is provided for Affymetrix arrays, others steps (segmentation, calling, and marker selection) can be applied to high-throughput sequencing data.

Acknowledgement

We thank Serge Iovleff for his help implementing `HDPenReg`, and Samuel Blanck for relevant remarks when testing first versions of `MPAgenomics`. We also thank Claude Preudhomme and

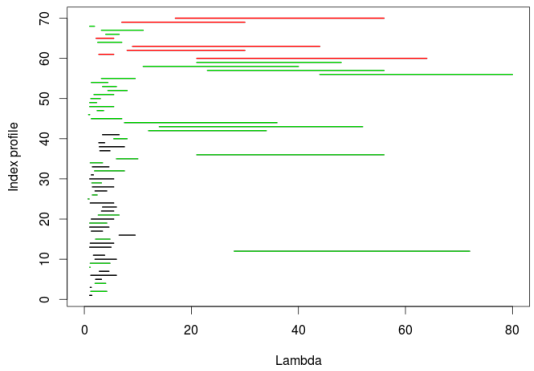

Figure 2: Widest ranges of λ (x-axis) for 70 copy number profiles (chromosome 1) (y-axis). Colors indicate clusters of signal-to-noise ration (black $<$ red $<$ green).

Olivier Nibourel for providing the data, and for their helpful clinical competences to interpret the results.

Funding: The developpment of this package was funded by the Inria Action de Développement Technologique named *MPAGenomics*.

References

[1] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[2] Henrik Bengtsson. aroma - An R Object-oriented Microarray Analysis environment. Preprint in Mathematical Sciences 2004:18, Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Sweden, 2004.

[3] Henrik Bengtsson, Pierre Neuvial, and Terence P. Speed. Tumorboost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics*, 11, 2010.

[4] Henrik Bengtsson, Pratyaska Wirapati, and Terence P. Speed. A single-array preprocessing method for estimating full-resolution raw copys from all affymetrix genotyping arrays including genomewidesnp 5 & 6. *Bioinformatics*, 25(17):2149–2156, 2009.

[5] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2 2010.

[6] Toby Hocking, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Valentina Boeva, Julie Cappo, Olivier Delattre, Francis Bach, and Jean-Philippe Vert. Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinformatics*, 14(1):164, 2013.

[7] Rebecca Killick and Idris Eckley. *changepoint: An R package for changepoint analysis*, 2013. R package version 1.1.

[8] Aline Renneville, Raouf Ben Abdelali, Sylvie Chevret, Olivier Nibourel, Meyling Cheok, Cecile Pautas, Remy Dulery, Thomas Boyer, Jean-Michel Cayuela, Sandrine Hayette, Emmanuel Raffoux, Hassan Farhat, Nicolas Boissel, Christine Terre, Herve Dombret, Sylvie Castaigne, and Claude Preudhomme. Clinical impact of gene mutations and lesions detected by snp-array karyotyping in acute myeloid leukemia patients in the context of gemtuzumab ozogamicin treatment: Results of the alfa-0701 trial. *Oncotarget*, 4(9), 2013.

[9] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[10] Mark van de Wiel and Sjoerd Vosse. *CGHcall: Calling aberrations for array CGH tumor profiles.*, 2012. R package version 2.20.0.