

Analysis of genomic markers: Make it easy with the R package MPAGenomics



Quentin Grimonprez¹, Alain Celisse^{1,2} and Guillemette Marot^{1,2,3}

¹Équipe MODAL (Inria Lille Nord Europe), ²Laboratoire Paul Painlevé (Université Lille 1 - CNRS),

³Équipe d'Accueil 2694 (Université Lille 2)



Context

Data

Affymetrix genome-wide SNP6 arrays.

About 200 biological samples with two types of profiles :

- copy-number: ~ 1.8 million probes (SNPs + CN)
- allele B fraction: proportion of total signal from allele B (~ 930.000 SNPs).

Goal

- Create an R package : pipeline for beginners in R to easily perform data analysis from genome-wide SNP arrays.
- Calibration method for the segmentation parameter.

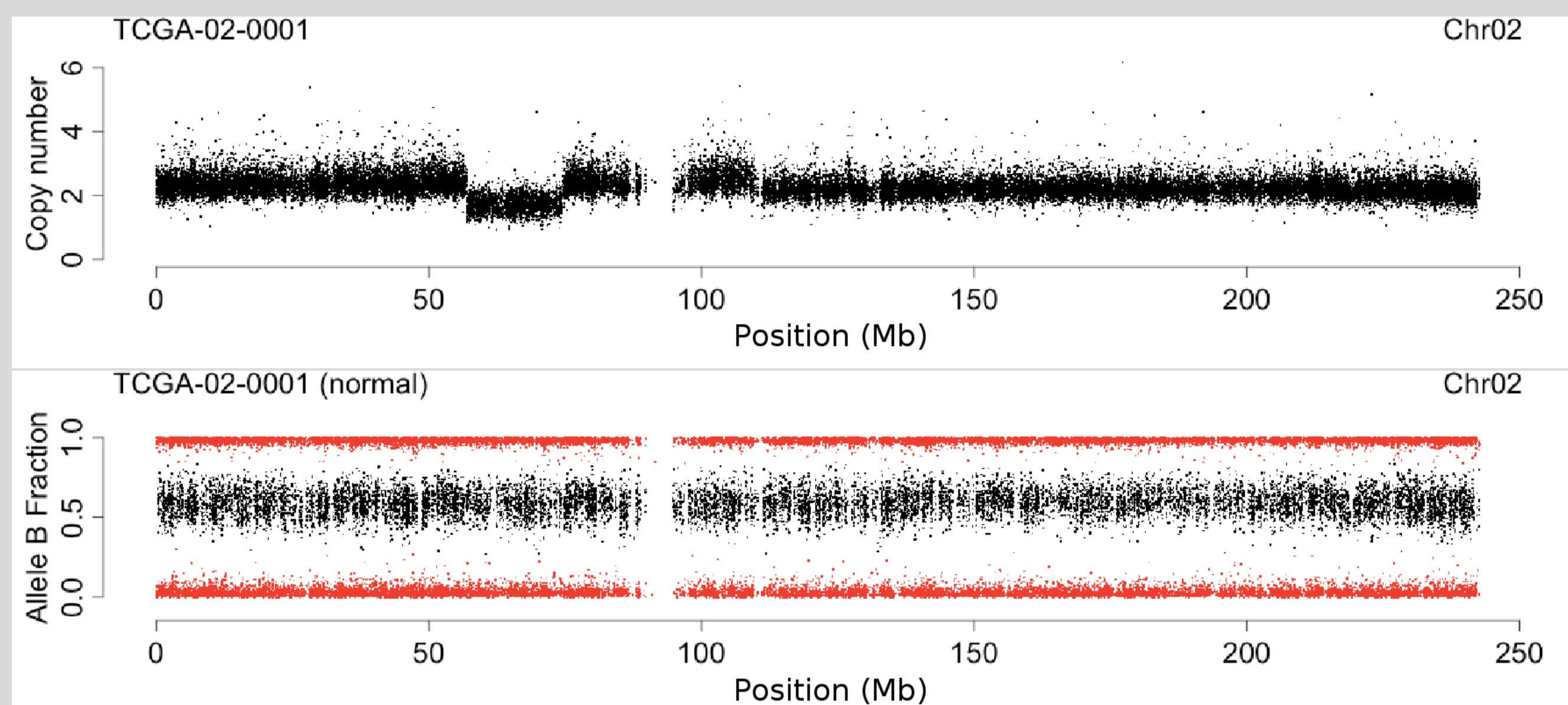


Figure 1: Top : Copy-number profile. Bottom : Allele B fraction profile : homozygous SNPs (red), heterozygous SNPs (black).

Data Normalization

Packages aroma

- Technical biases correction
- Copy-number & allele B fraction calculation
- *TumorBoost* : better allele B fraction correction for studies with matched normal-tumor samples

Difficulties for beginners :

- Complicated internal documentation
- **Heavy architecture** to deal with
- No way to perform the whole analysis straightforwardly

MPAGenomics contribution

1. Normalize data via **MPAGenomics**
 - Easily build architecture
 - Provide automatic wrappers of **aroma** functions
2. Provide normalized data

Segmentation

Copy-number

Copy-number signal is segmented by the PELT segmentation method from **changepoint** package (Killick et al., 2013).

Allele B fraction

Heterozygous SNPs are kept and the signal is symmetrized. Then, the signal is segmented the same way as the copy-number signal.

Calibration of λ parameter in PELT

- PELT depends on a parameter to calibrate.
- **MPAGenomics**: automatic calibration of λ .

Calling method

- Assign labels (loss, normal or gain) to segments (copy-number).
- **CGHcall** package (van de Wiel et al., 2007).

Markers selection

Strategy

- Select genomic markers (e.g. SNPs or CNV) associated with a response y .
- Lasso method for sparse selection (few markers) with $\rho > 0$:

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

Implementation in MPAGenomics

- **Linear regression**: **HDPenReg** for huge amount of variables (**HDPenReg**: R package, C++ implementation of LARS (Efron et al., 2004)).
- **Logistic regression**: wrapper of **glmnet** R package (Friedman et al., 2010).
- Choice of ρ by k -fold cross validation.

Calibration of λ (segmentation)

- PELT default parameter is misleading.

- **MPAGenomics**: automatic data-driven choice of λ

Strategy

1. Grid of λ : $0 < \lambda_1 < \lambda_2 < \dots < \lambda_{\max}$.
2. Run PELT for each λ_i (see Figure 2 left).
3. Choose λ corresponding to the widest range such that the number of segments is constant (> 1).

Sample-specific parameter versus common λ

1. Common λ :

- Compute the signal-to-noise ratio (SNR) for each profile.
- Cluster profiles according to SNR (Gaussian mixture).
- For each cluster, choose λ .

2. Sample-specific λ :

MPAGenomics provides an automatic choice of λ for each profile.

Sample-specific parameter versus common λ

Common λ within each cluster is misleading (Figure 2 right).

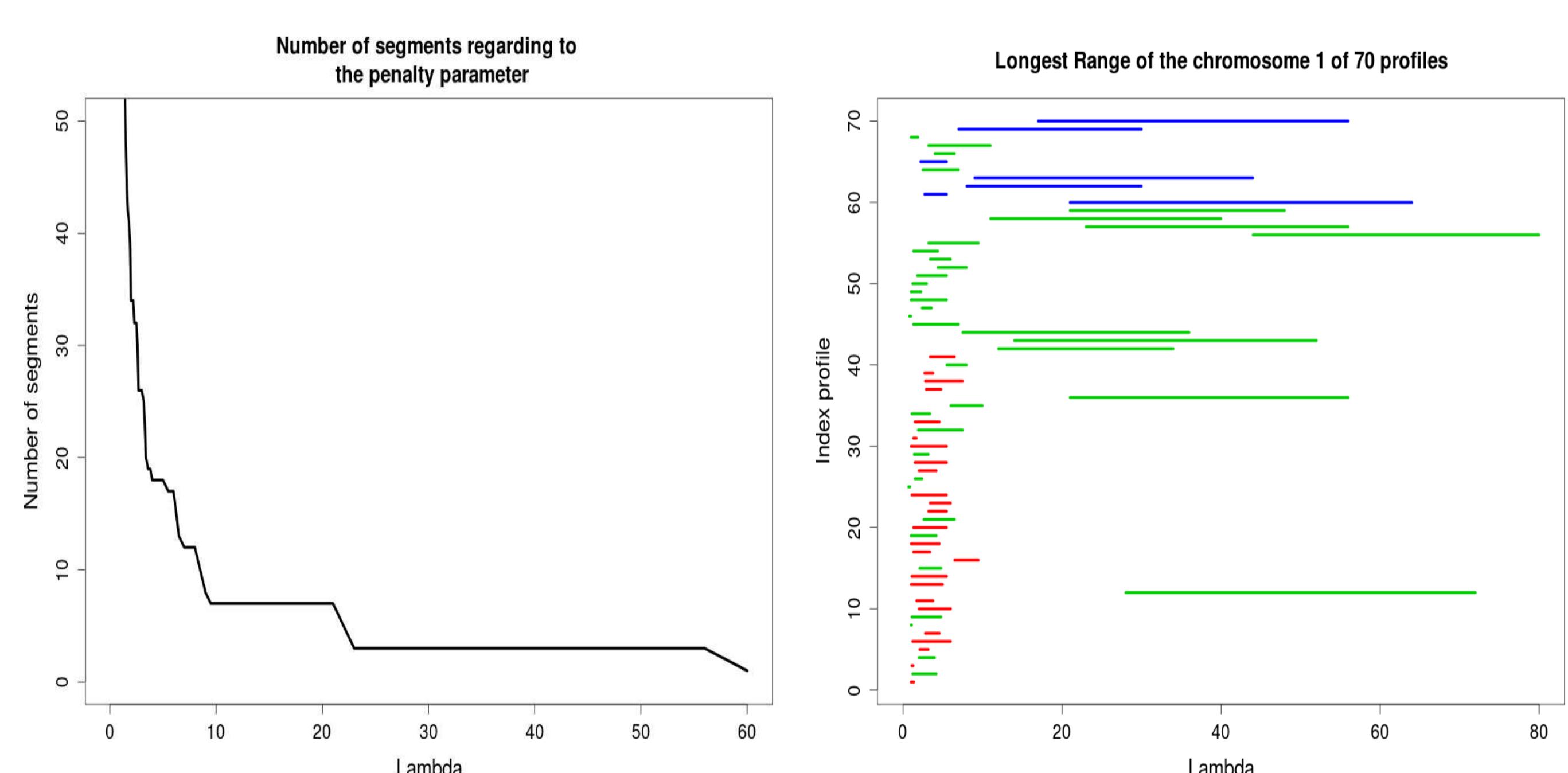


Figure 2: Left: Number of segments versus λ (chromosome 1). Right: Ranges of optimal λ for each profile. Color indicates a given signal-to-noise ratio (red/green/blue).

Bibliography

- [1] Bengtsson H, (2004) aroma - An R Object-oriented Microarray Analysis environment, *Preprint in Mathematical Sciences*.
- [2] Bengtsson H, Wirapati P, Speed,T.P. (2009) A single-array preprocessing method for estimating full-resolution raw copycs from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6, *Bioinformatics*, **25**, 2149-2156.
- [3] Bengtsson H, et al (2010) TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays, *BMC Bioinformatics*, **11**, 1-17.
- [4] Efron B, et al (2010) Least angle regression, *Annals of Statistics*, **32**, 407-499.
- [5] Friedman J, et al (2007) Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, **33**, 1-22.
- [6] Hocking T, et al (2013) Learning smoothing models of copy number profiles using breakpoint annotations, *BMC Bioinformatics*, **14**, 164.
- [7] Killick R, Eckley I (2013) changepoint: An R package for changepoint analysis, *R package version 1.1*, <http://CRAN.R-project.org/package=changepoint>.
- [8] Tibshirani R, (1994) Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.
- [9] van de Wiel M, et al. (2007) CGHcall: Calling aberrations for array CGH tumor profiles, *Bioinformatics*, **23**, 892-894.