Analysis of genomic markers: Make it easy with the R package MPAgemomics
Quentin Grimonprez, Alain Celisse, Guillemette Marot

To cite this version:
Quentin Grimonprez, Alain Celisse, Guillemette Marot. Analysis of genomic markers: Make it easy with the R package MPAgemomics. SMPGD 2014, Jan 2014, Paris, France. 2014. hal-01091543

HAL Id: hal-01091543
https://hal.archives-ouvertes.fr/hal-01091543
Submitted on 5 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Analysis of genomic markers: Make it easy with the R package MPAgenomics

Quentin Grimonprez1, Alain Celisse1,2 and Guilleamette Marot1,2
1Équipe MODAL (Inria Lille Nord Europe). 2Laboratoire Paul Painlevé (Université Lille 1 - CNRS).

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), heterogeneous 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 ρ > 0:
  \[ \sum_{i=1}^{n} (y_i - (X\beta)_i)^2 + \rho \sum_{p=1}^{P} |\beta_p| \]

Implementation in MPAgenomics
• Linear regression: HDPenReg for hinge amount of variables (HDPenReg R package, C+ implementation of LAIR (Elion 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 < λ1 < λ2 < · · · < λmax.
2. Run PELT for each λi (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).

Bibliography