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

To cite this version:

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 Grimponprez1, Alain Celisse1,2, and Guillelmette Marot1,2,3
1Équipe MODAL (Inria Lille Nord Europe), 2Laboratoire Paul Painlevé (Université Lille 1 - CNRS), 3Équipe d’Accueil 2694 (Université Lille 2)

Context
Data
Asymmetric genome-wide SNP
t 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.

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
- Differentials 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 \( \lambda \) parameter in PELT
- PELT depends on a parameter to calibrate.
- MPAGenomics: automatic calibration of \( \lambda \).

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

Calibration of \( \lambda \) (segmentation)

- PELT default parameter is misleading.
- MPAGenomics: automatic data-driven choice of \( \lambda \).

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

Sample-specific parameter versus common \( \lambda \)
- \( \lambda \), Compute the signal-to-noise ratio (SNR) for each profile.
- Cluster profiles according to SNR (Gaussian mixture).
- For each cluster, choose \( \lambda \).

Sample-specific parameter versus common \( \lambda \)
- Common \( \lambda \) within each cluster is misleading (Figure 2 right).

Markers selection

Strategy
- Select genomic markers (e.g. SNPs or CNV) associated with a response \( y \).
- Lasso method for sparse selection (few markers) with \( \rho > 0 \) :
  \[
  \sum_{i=1}^{p} |y_i - (X_\lambda)_{i\cdot}|^2 + \rho \sum_{p=1}^{P} |\beta_p|
  \]

Implementation in MPAGenomics
- Linear regression: HDPenReg for large amount of variables (HDPenReg R package, C++ implementation of LAIR (Efron et al., 2004)).
- Logistic regression: wrapper of glmnet R package (Friedman et al., 2010).
- Choice of \( \rho \) by k-fold cross validation.

Bibliography