Analyse multi-patients de données génomiques

Quentin GRIMONPREZ[1], Alain Celisse[1,2] and Guillemette MAROT[1,3]

[1] MΘDAL team, Inria Lille-Nord Europe, France
[2] Laboratoire Paul Painlevé, Université Lille 1, France
[3] EA 2694, Université Lille 2, France

SUMMARY

1 Introduction

Introduction

Data

- Patients with leukemia (about 300).
- Profiles obtained using SNP arrays with DNA from patients at diagnosis and at remission.
- Informations about the relapse of patients.

Goal

- Select important markers for predicting the relapse of patients.

Partners

- Laboratoire d'hématologie, CHRU Lille.
- Plate-forme de génomique fonctionnelle et structurale, Université Lille 2.

2
Markers selection

Markers selection

- Select genomic markers (e.g. SNPs or CNV) associated with a response $y \in \mathbb{R}^p$.
- $X \in \mathcal{M}_{n,p}(\mathbb{R})$ containing the different signals.
- Lasso [Tibshirani, 1994] method for sparse selection (few markers) with $\lambda > 0$:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

R packages

- Package `glmnet` for logistic regression [Friedman et al., 2010]
- Re-implementation of LARS algorithm [Efron et al., 2004] for linear regression

R Package `HDPenReg`

- Lars algorithm in C++ with R interface
- Choice of λ by k-fold cross validation.
- Faster than `lars` package
- Work with bigger data than `lars` package

Conclusion

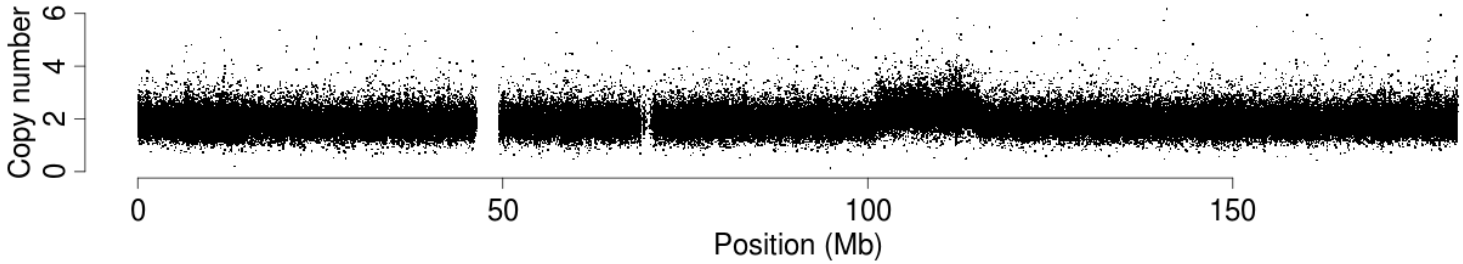
- No interesting results
- According to M.D., an abnormality is present in 5-10% of patients at the same position. The number of abnormalities is more important.

3
Segmentation & Calling

Segmentation

Goal

- Segment copy-number signal



Segmentation

Statistic modeling

- $y = (y_1, \ldots, y_n)$ a signal of length n
- Let $r_1, r_2, \ldots, r_{K-1}$ position of breakpoints
- Let $\mu_1, \mu_2, \ldots, \mu_K$ means of segments
- $\forall t \in s_i = [r_i, r_{i+1}[\,, y_t = \mu_i + N_t$ où $N_t \sim \mathcal{N}(0, \sigma^2)$
  $i = 1, \ldots, K-1$

Maximum of likelihood

- Maximize : $\log p(y; R, \mu) = \sum_{k=1}^{K} \sum_{t \in s_k} \log\, p(y_t; \mu_k) + Pen$
- In Gaussian case :
  $\log p(y; R, \mu) = C_1 - C_2 \sum_{k=1}^{K} \sum_{t \in s_k} (y_t - \mu_k)^2$

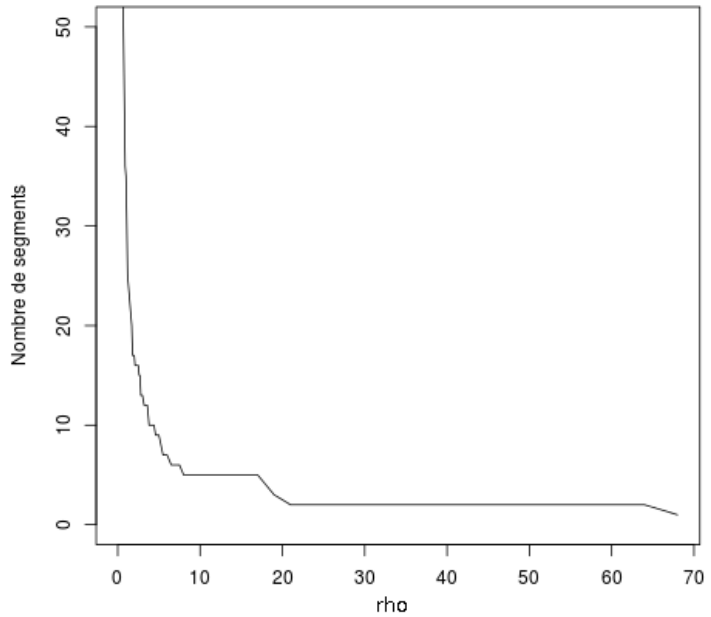
Method recommanded in [Hocking et al., 2013]

- PELT method [Killick et al., 2012] from `changepoint` package
- Default parameter for each profile and each chromosome
  $\Rightarrow$ In practise, over-segmentation of our data.
- Calibrate the penalty parameter ρ of the penalty $\rho \log(n)$ according to the data

Calibration of PELT parameter

Sample specific parameter

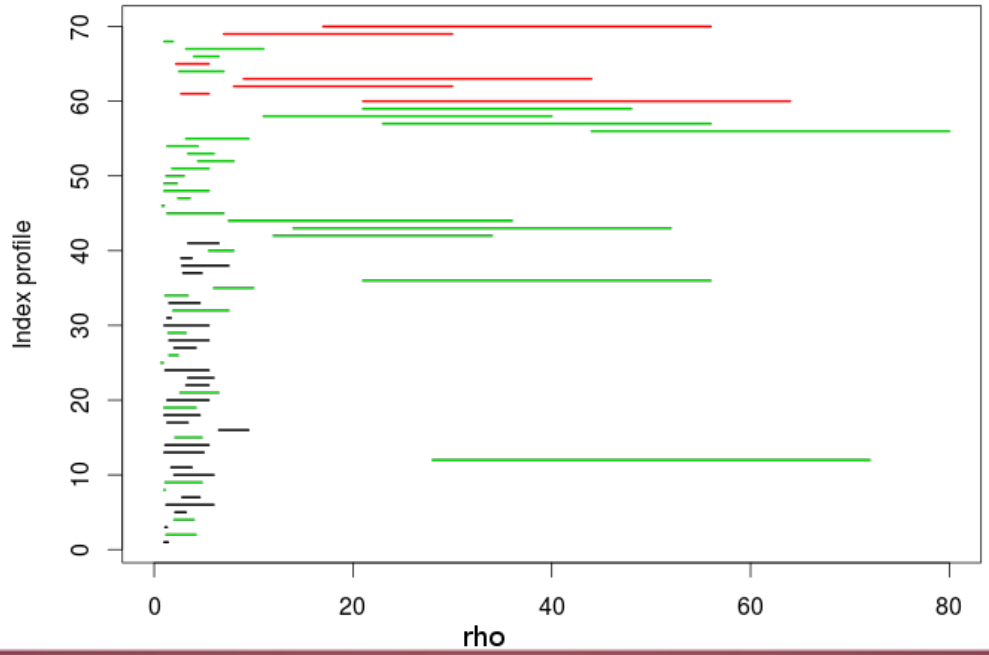
- Grid of ρ : $0 < \rho_1 < \cdots < \rho_{\text{max}}$
- Run PELT for each ρ_i
- Choose ρ corresponding to the widest range such that the number of segments is constant



Calibration of PELT parameter

Common parameter by group of variance

- Compute variance for each sample
- Cluster profiles according to variance with a Gaussian Mixture Model
- For each cluster, choose ρ



Calling

Goals

- Put a label on each segment (gain, loss, ...)
- Isolate non-normal segments

Method

- CGHcall package [van de Wiel et al., 2007]
- based on Gaussian Mixture Model

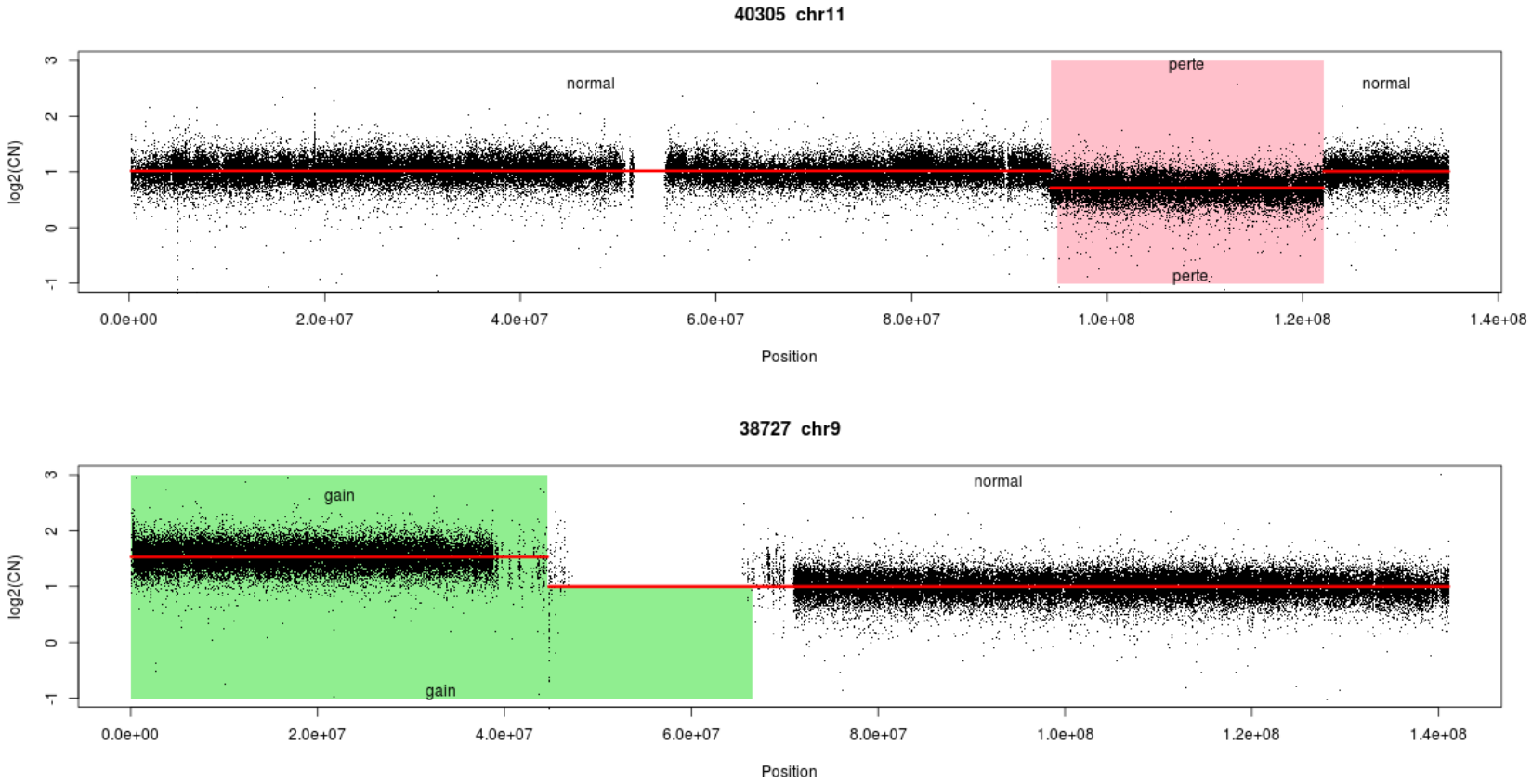
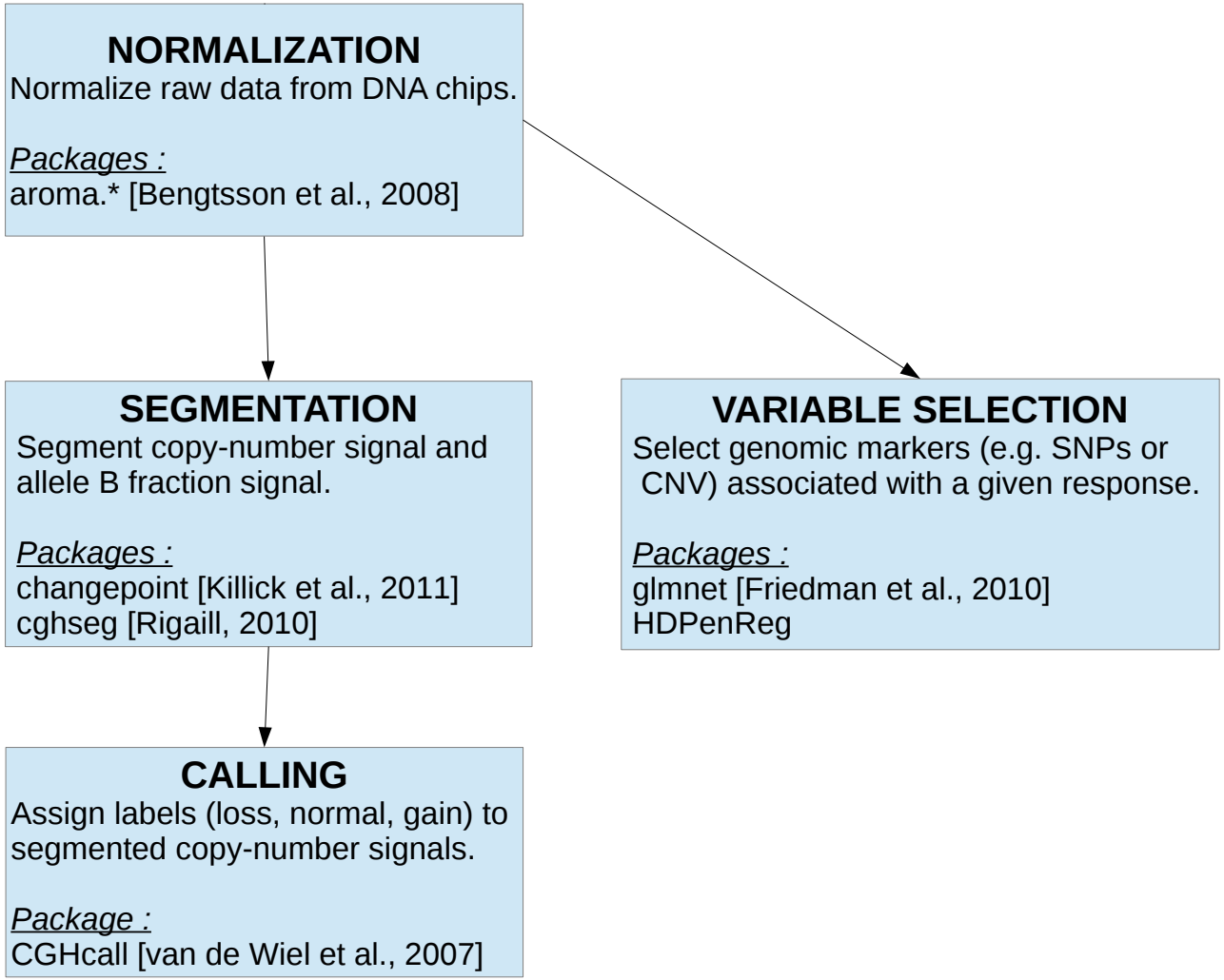


FIGURE: Examples of segmented and annotated signals. On each graph, CGHcall annotation vs human annotation.

4
MPAgenomics package

MPAgenomics package



5 Conclusion

Conclusion

Easy way to perform automatically multi-patients analysis

R Packages

- `HDPenReg` available on R-forge
- `MPAgenomics` available on R-forge $\Rightarrow$ `https://r-forge.r-project.org/R/?group_id=1658`

Futures Development

- Galaxy interface
- Joint segmentation

Acknowledgments

- Guillemette Marot[1,3]
- Alain Celisse[1,2]
- Meyling Cheok[4]
- Martin Figeac[5]
- Samuel Blanck[1]

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

Bibliographie I

Bengtsson, H. (2004).

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

Bengtsson, H., Neuvial, P., and Speed, T. P. (2010).

Tumorboost : Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays.
BMC Bioinformatics, 11.

Bengtsson, H., Simpson, K., Bullard, J., and Hansen, K. (2008).

aroma.affymetrix : A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory.
Technical Report 745, Department of Statistics, University of California, Berkeley.

Birgé, L. and Massart, P. (2007).

Minimal penalties for gaussian model selection.
Probability Theory and Related Fields, 138(1-2) :33–73.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004).

Least angle regression.
Annals of Statistics, 32 :407–499.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010).

Regularization paths for generalized linear models via coordinate descent.
Journal of Statistical Software, 33(1) :1–22.

Bibliographie II

Hocking, T., Schleiermacher, G., Janoueix-Lerosey, I., Boeva, V., Cappo, J., Delattre, O., Bach, F., and Vert, J.-P. (2013).
Learning smoothing models of copy number profiles using breakpoint annotations.
BMC Bioinformatics, 14(1) :164.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012).
Optimal detection of changepoints with a linear computational cost.
Journal of the American Statistical Association, 107(500) :1590–1598.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005).
A statistical approach for array CGH data analysis.
BMC Bioinformatics, 6(1) :27.

Rigaill, G. (2010).
Pruned dynamic programming for optimal multiple change-point detection.

Tibshirani, R. (1994).
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society, Series B, 58 :267–288.

van de Wiel, M. A., Kim, K. I., Vosse, S. J., van Wieringen, W. N., Wilting, S. M., and Ylstra, B. (2007).
Cghcall : calling aberrations for array cgh tumor profiles.
Bioinformatics, 23(7) :892–894.

MERCI

Inria Lille-Nord Europe
Villeneuve d'Ascq
www.inria.fr/centre/lille