BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 8 2013, pages 1081–1082
doi:10.1093/bioinformatics/btt082

Gene expression Advance Access publication February 17, 2013

PLRS: a flexible tool for the joint analysis of DNA copy number and mRNA expression data

Gwenaël G.R. Leday[1,*] and Mark A. van de Wiel[2]

[1]Department of Mathematics, VU University, De Boelelaan 1081a, 1081HV Amsterdam and [2]Department of Epidemiology and Biostatistics, VU University Medical Center, 1007MB Amsterdam, The Netherlands

Associate Editor: Martin Bishop

ABSTRACT

Summary: DNA copy number and mRNA expression are commonly used data types in cancer studies. Available software for integrative analysis arbitrarily fixes the parametric form of the association between the two molecular levels and hence offers no opportunities for modelling it. We present a new tool for flexible modelling of this association. PLRS uses a wide class of interpretable models including popular ones and incorporates prior biological knowledge. It is capable to identify the gene-specific type of relationship between gene copy number and mRNA expression. Moreover, it tests the strength of the association and provides confidence intervals. We illustrate PLRS using glioblastoma data from The Cancer Genome Atlas.

Availability and implementation: PLRS is implemented as an R package and available from Bioconductor (as of version 2.12; http://bioconductor.org). Additional code for parallel computations is available as Supplementary Material.

Contact: g.g.r.leday@vu.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 23, 2012; revised on January 11, 2013; accepted on February 12, 2013

1 INTRODUCTION

DNA copy number aberrations are characteristics of the cancer cell. These aberrations are gains and losses of chromosomal DNA, which may alter expression levels of mRNA transcripts. The identification of genes for which an abnormal copy number affects gene expression is important in cancer studies, as these genes are likely to be relevant for tumourigenesis. Here, we present a new tool for exploratory and confirmatory analysis of such effects.

For a given gene, copy number and mRNA expression are generally believed to be *concordant*. The exact form of the association is usually not established. In fact, the shape is likely to differ between genes because of the presence of different (post-) transcriptional regulatory mechanisms. Tools that investigate the interaction between the two molecular levels assist in better understanding of regulatory mechanisms.

Numerous software packages have been proposed for joint analysis of copy number and gene expression data (Chari *et al.*, 2008; Lê Cao *et al.*, 2009; Lee and Kim, 2009; Louhimo and Hautaniemi, 2011; Salari *et al.*, 2010; van Wieringen *et al.*, 2006). However, most of these fix the association between DNA and RNA a priori, typically a linear or piecewise constant line. Hence, these approaches do not permit investigation or identification of the shape of the association. Recently, the need for more subtle models has been highlighted (Leday *et al.*, 2013; Nemes *et al.*, 2012; Solvang *et al.*, 2011) to reflect the biological mechanisms between the two molecular levels. Here, we describe the R package PLRS that implements the framework recently proposed by Leday *et al.* (2013). PLRS uses piecewise linear regression splines, which allow multiple linear lines, and are a wide class of interpretable models including the linear and piecewise constant ones. It enforces concordance by restricting relevant model parameters. In addition, PLRS tests the strength of the overall association, identifies its functional shape and provides confidence intervals for the estimated curve. We illustrate PLRS using a dataset from 160 glioblastoma samples obtained from The Cancer Genome Atlas (TCGA).

2 MODEL

PLRS models *cis*-relationships between copy number and mRNA expression by piecewise linear regression splines (Leday *et al.*, 2013). The relevance of this class of models is multifold. Unlike other methods, PLRS combines copy number data from various steps of the preprocessing, namely, the *segmented* and *called* data (van de Wiel *et al.*, 2011). Segmented data are continuous (log$_2$-values) and provide the (relative) amount of DNA copies (gene dosage), whereas called data represent discrete states associated with the various types of copy number aberration; the biological literature commonly distinguishes four of these: 'loss' (less than two copies of genomic DNA), 'normal' (two copies), 'gain' (three to four copies) and 'amplification' (more than four copies). Second, PLRS allows the effect of DNA on mRNA to differ across types of aberrations. This is biologically plausible: the efficacy of mechanisms that compensate for genomic aberrations may differ between losses, gains and amplifications. Third, good interpretability is ensured by the piecewise linearity of the model and a set of restrictions on the parameters. For example, copy number is concordant with gene expression and 'normal' copy number cannot severely alter gene expression.

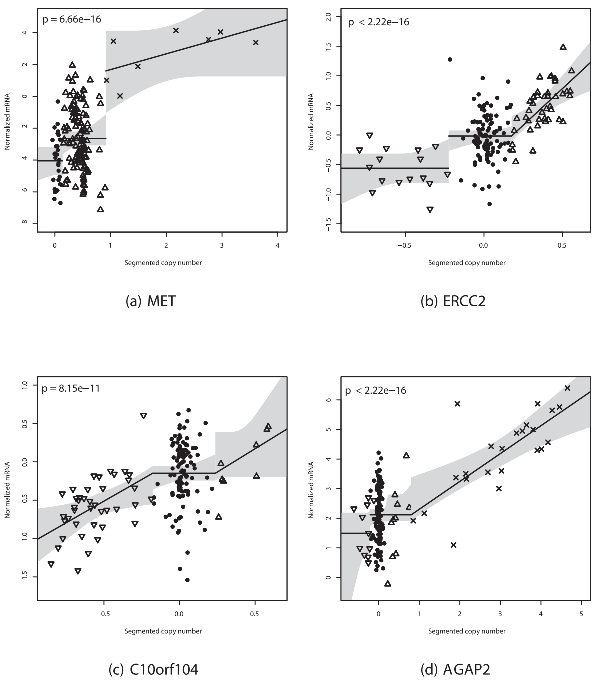
In this context, the R package PLRS implements various statistical procedures to detect which and how gene copy number abnormalities alter the gene expression level. Identification of the functional form of the association is achieved by model selection, which automatically merges copy number states when their

[*]To whom correspondence should be addressed.

association with mRNA expression can be captured with one regression line. Simultaneous confidence intervals on the selected curve are provided for more detailed description. Finally, a statistical test evaluates the significance of the overall association by testing the null hypothesis: copy number does not affect mRNA expression, leading to a single horizontal line.

3 RESULTS

We applied PLRS to a dataset of 160 glioblastoma tumour samples obtained from TCGA (http://cancergenome.nih.gov/; Verhaak *et al.*, 2010) for which copy number (Agilent CGH Microarray 244A) and mRNA expression (Agilent 244K platform) were available. We found that for many known cancer genes, the expression level is strongly associated with DNA aberrations (cf. Supplementary Material). Figure 1 depicts the DNA–mRNA association for four genes, including known cancer genes *MET*, *ERCC2* and *AGAP2*. Clearly, relationships are different and demonstrate that the flexibility of the PLRS model allows new insights in the association. For gene *MET*, we observe that the effect of amplifications extends that of gains more than proportionally. For *ERCC2*, the expression level of samples with loss and normal copy number differs in average and expression increases linearly with dosage. Amplifications of gene *AGAP2* have a strong effect on mRNA expression, whereas gains have none. The effect as defined by PLRS is broad and expressed by both an intercept and a slope for each copy number aberration state. The variety of models resulting from PLRS contrasts with most other methods, which impose a unique parametric form to all genes. Our method lets the data decide what is most appropriate. As a consequence, PLRS has more power than other standard methods for detecting relatively large effects occurring in small subgroups of samples (Leday *et al.*, 2013). Note that other non-linear techniques, e.g. based on mutual information, can be competitive but less interpretable.



(a) MET

(b) ERCC2

(c) C10orf104

(d) AGAP2

Fig. 1. DNA–mRNA associations for four genes in the TCGA dataset. *X*-axis: Gene dosage (segmented values), *y*-axis: mRNA gene expression. Copy number states are indicated by symbols: loss (open inverted triangles), normal (open circles), gain (open triangles) and amplification (multiplication symbols). Grey surfaces correspond to 95% uniform confidence bands. The top left value corresponds to the *P*-value of the PLRS test

4 CONCLUSION

PLRS is a tool for flexible modelling of the association between DNA copy number and mRNA expression. We demonstrated its potential to reveal interesting relationships. It is particularly useful for (i) a detailed understanding of the relationship between DNA copy number and mRNA expression and (ii) powerful detection of copy number-induced sample subgroup-specific effects, thereby acknowledging heterogeneity of many cancers. The software can also be used for studying the effect of DNA copy number on microRNA expression.

Funding: The Center for Medical Systems Biology (CMSB), established by the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NGI/NWO).

Conflict of Interest: none declared.

REFERENCES

Chari,R. *et al.* (2008) SIGMA2: a system for the integrative genomic multidimensional analysis of cancer genomes, epigenomes, and transcriptomes. *BMC Bioinformatics*, **9**, 422.

Lê Cao,K.A. *et al.* (2009) integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, **25**, 2855–2856.

Leday,G.G.R. *et al.* (2013) Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines. *Ann. Appl. Stat*, doi:10.1214/12-AOAS605.

Lee,M. and Kim,Y. (2009) CHESS (CgHExpreSS): a comprehensive analysis tool for the analysis of genomic alterations and their effects on the expression profile of the genome. *BMC Bioinformatics*, **10**, 424.

Louhimo,R. and Hautaniemi,S. (2011) CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics*, **27**, 887–888.

Nemes,S. *et al.* (2012) Segmented regression, a versatile tool to analyze mRNA levels in relation to DNA copy number aberrations. *Gene Chromosome Cancer*, **51**, 77–82.

Salari,K. *et al.* (2010) DR-integrator: a new analytic tool for integrating DNA copy number and gene expression data. *Bioinformatics*, **26**, 414–416.

Solvang,H. *et al.* (2011) Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer. *BMC Bioinformatics*, **12**, 197.

van de Wiel,M.A. *et al.* (2011) Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief Bioinform.*, **12**, 10–21.

van Wieringen,W.N. *et al.* (2006) ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data. *Bioinformatics*, **22**, 1919–1920.

Verhaak,R.G. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*, **17**, 98–110.