JCBi JOURNAL OF CLINICAL BIOINFORMATICS

RESEARCH Open Access

A novel tree-based procedure for deciphering the genomic spectrum of clinical disease entities

Cyprien Mbogning[1,2*], Hervé Perdry[2,3], Wilson Toussile[2,3] and Philippe Broët[1,2,3,4]

Abstract

Background: Dissecting the genomic spectrum of clinical disease entities is a challenging task. Recursive partitioning (or classification trees) methods provide powerful tools for exploring complex interplay among genomic factors, with respect to a main factor, that can reveal hidden genomic patterns. To take confounding variables into account, the partially linear tree-based regression (PLTR) model has been recently published. It combines regression models and tree-based methodology. It is however computationally burdensome and not well suited for situations for which a large number of exploratory variables is expected.

Methods: We developed a novel procedure that represents an alternative to the original PLTR procedure, and considered different selection criteria. A simulation study with different scenarios has been performed to compare the performances of the proposed procedure to the original PLTR strategy.

Results: The proposed procedure with a Bayesian Information Criterion (BIC) achieved good performances to detect the hidden structure as compared to the original procedure. The novel procedure was used for analyzing patterns of copy-number alterations in lung adenocarcinomas, with respect to Kirsten Rat Sarcoma Viral Oncogene Homolog gene (KRAS) mutation status, while controlling for a cohort effect. Results highlight two subgroups of pure or nearly pure wild-type KRAS tumors with particular copy-number alteration patterns.

Conclusions: The proposed procedure with a BIC criterion represents a powerful and practical alternative to the original procedure. Our procedure performs well in a general framework and is simple to implement.

Keywords: Recursive partitioning, Tree-based regression, Lung cancer, Disease taxonomy, Genomic

Background

Recent advances in large-scale genomic technologies provide an unprecedented amount of data that offer new insights into the molecular portraits of diseases. This information enables to dissect a heterogeneous disease entity into more homogeneous subentities that can be relevant for clinical purposes.

This problem is particularly appealing in oncology where molecular classification of tumors, that are based on the status of specific targeted therapy, rely mainly upon a single molecular event but overlook tumoral genomic complexity. For most solid epithelial tumors, these genomic events are primarily DNA mutations that give selective growth advantages to tumor cells.

A classical example in non-small-cell lung cancer (NSCLC) is the activating EGFR (epidermal growth factor receptor) mutation that predicts the sensitivity to EGFR tyrosine kinase inhibitors. EGFR-mutant lung adenocarcinoma is nowadays almost considered as a distinct disease entity. Such is not the case for KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog gene) mutation that represents one of the most common mutations in NSCLC. With the exception of its well-known mutually exclusive relationship with EGFR mutation, the clinical utility of KRAS mutation status has not been clearly demonstrated [1]. Moreover, it is still unclear whether subgroups exist within KRAS wild-type or KRAS mutated

*Correspondence: cyprien.mbogning@gmail.com
[1] Abirisk consortium WP4, 14-16 Avenue Paul-Vaillant-Couturier, 94807 Villejuif, France
[2] Inserm U669, 14-16 Avenue Paul-Vaillant-Couturier, 94807 Villejuif, France
Full list of author information is available at the end of the article

tumors. The identification of more homogeneous molecular subgroups with respect to KRAS mutational status may provide new genomic taxonomy of NSCLC tumors, that may help for the advancement of personalized medicine.

The aim of the clinical study that prompted this methodological work was to decipher heterogeneity of lung adenocarcinomas with respect to KRAS mutation status based upon whole-genome copy-number alterations. Copy-number alteration (CNA) is one of the main type of genomic alterations that is linked to genome instability and represents a key feature of human carcinomas [2]. In previous cancer studies, association between specific CNAs and point mutations have been reported such as, for example, the relationship between EGFR mutations and copy-gains of 7p12 (which harbors EGFR gene) in lung adenocarcinomas. However, few investigations have been performed for studying the relationships between KRAS mutation and CNAs.

Identifying homogeneous subgroups, with respect to a main factor, based on the complex interplay among genomic alterations is a difficult task that cannot be easily done with standard regression models.

In contrast, recursive partitioning (or tree-based) methods such as CART (Classification And Regression Tree) [3] is a flexible and powerful alternative for exploring high-order interaction between explanatory variables. From a data mining perspective, the purpose of such approach is to decompose a data space recursively into smaller areas that are defined by the set of explanatory variables and tree-structured. The hypothesis space is the set of all possible hyper-rectangular areas. These areas are more homogeneous with respect to the main factor as compared to the whole population. The analysis of the patterns of these areas, that are defined by the explanatory variables, can provide meaningful biological insights. In the context of non-parametric statistical methods, random forests [4] is the classical extension to tree-based methods with many available R packages (for a few: VarSelRF [5], SRF [6], RF [7]). As compared to tree-based methods, a forest that consists of thousands of unpruned trees is more stable and well suited for prediction. However, random forests loose the easy interpretability of CART, which represents the key objective when dissecting the clinico-biological spectrum of clinical disease entities.

For clinical epidemiology studies, an important drawback of classical tree-based methodology is that it does not provide a straightforward way of adjusting for confounding variables. In practice, confounding and explanatory variables are considered in the same way. Thus, the final tree is a mixture of confounder and explanatory variables lacking of clear interpretation and whose joint effects are distorted. This problem was of particular concern in our clinical studies since our series was composed of two different sub-populations (Asian and Caucasian patients). In lung adenocarcinomas, KRAS mutation is found in about one third of the tumours in Caucasian populations, as opposed to less than one tenth is Asian populations. Thus, in our study, it was mandatory to adjust for this confounding factor.

In a pioneering a work, Chen et al. [8] have introduced a new class of regression models, called partially linear tree-based regression models (PLTR). This new framework has been proposed for genetic epidemiology studies in order to assess complex joint gene-gene and gene-environment effects taking into account confounding variables. In practice, PLTR models represent a new class of semi-parametric regression models that integrates the advantages of generalized linear regression and tree-structure models. The linear part is used to model the main effects of confounder variables and the nonparametric tree part is used to capture the distributional shape of the data relying on the complex joint effects of multiple explanatory variables. In their article, Chen et al. have proposed a four-step selection and testing procedure for identifying the optimal tree while adjusting for linear (on the generalized linear scale) confounding variables. This methodology has been recently extended for considering multivariate outcomes [9]. However, Chen's et al. procedure heavily relies on resampling, which is computationally burdensome and not well-suited to situations for which a large number of explanatory variables is expected. In the present work, we propose and evaluate an alternative procedure with different selection criteria, which considers separately the objectives of selection and testing.

We first describe the novel procedure with three different selection criteria. It corresponds to a modified PLTR procedure with four steps, of which the two first are common to the one proposed by Chen et al. A simulation study with different scenarios is presented that compares the power of the proposed procedure to the original PLTR strategy. The proposed procedure is used to decipher heterogeneity of lung adenocarcinomas, with respect to KRAS mutation, based on copy-number alterations.

Methods

In the following we present our novel procedure with different selection criteria. The first two steps are similar to those of the original PLTR procedure (Chen et al.) whereas the last two steps are new. The four steps are summarized in Figure 1 and presented in details below.

Denote $\mathbf{Y}$ the outcome of interest (or the main factor for the application considered in this work), $\mathbf{X}$ the confounding variables (to be modeled linearly), and $\mathbf{Z}$ the explanatory variables. The PLTR model is specified by:

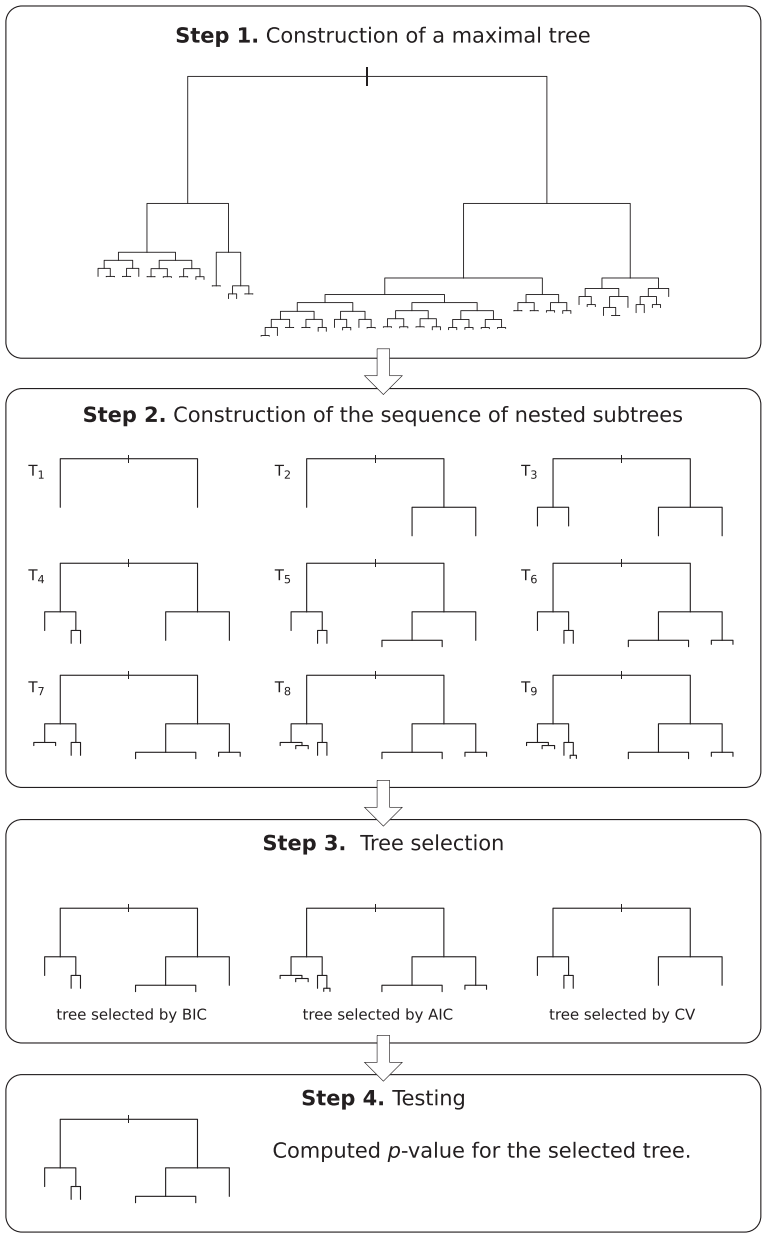

Figure 1 Flow-chart of the four steps of the method.

$$g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right) = \mathbf{X}'\theta + \beta_T F\left(T\left(\mathbf{Z}\right)\right), \quad (1)$$

where $g(\cdot)$ is a known link function (generalized linear model), $F\left(T\left(\mathbf{Z}\right)\right)$ is a vector of indicator variables representing the leaves of the tree $T\left(\mathbf{Z}\right)$.

Step 1: Construction of a maximal tree

The maximal tree is constructed as follows:

- **Initialization:** fit the generalized linear model (GLM) $g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right) = \mathbf{X}'\theta^{(0)} + \beta^{(0)}$
- **Iterations:** iterate the following steps starting with $k = 1$.
  - **fit the tree part:** construct a maximal tree model $T^{(k)}$ based on $\mathbf{Z}$, using $\mathbf{X}'\theta^{(k-1)}$ as offset
  - **fit the leaves of the tree:** fit the GLM $g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right) = \beta_T^{(k)} F\left(T^{(k)}\left(\mathbf{Z}\right)\right)$ using $\mathbf{X}'\theta^{(k-1)}$ as offset
  - **fit the parametric part:** fit the GLM $g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right) = \mathbf{X}'\theta^{(k)}$, with $\beta_T^{(k)} F\left(T^{(k)}\left(\mathbf{Z}\right)\right)$ using as offset

- **Ending conditions:** the algorithm stops when the estimates of θ stabilize within a pre-specified range or after a pre-specified number of iterations.

In the above procedure, an offset is a predictor variable included in the model with coefficient fixed equal to one.

In the construction of the tree, the goodness of a candidate split is assessed for each node by the deviance of a generalized linear model fitted in the node by considering $\mathbf{X}'\theta$ as the offset. More precisely, the goodness of a candidate split is the deviance of the parent node minus the sum of the deviance of the two child nodes. The recursive partitioning stops when the number of cases in each terminal node is smaller than a pre-assigned threshold.

Step 2: Construction of the sequence of nested subtrees

At the end of the previous step, the estimated tree T_R is a maximal tree which generally overfits the data. The second step constructs a sequence of nested subtrees of T_R.

The PLTR model (1) obtained from the previous step is

$$g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right) = \mathbf{X}'\hat{\theta}_{T_R} + \hat{\beta}_{T_R} F\left(T_R\left(\mathbf{Z}\right)\right), \tag{2}$$

where $T_R\left(\mathbf{Z}\right)$ represents the maximal tree at convergence, R being its size (number of terminal nodes or leaves).

Let $D\left(\mathbf{X}\,;\,\hat{\theta}_0\right)$ be the deviance computed under the null hypothesis

$$\mathcal{H}_0 : g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right) = \mathbf{X}'\theta + \beta_0 \tag{3}$$

and $D\left(\mathbf{X}, T(\mathbf{Z})\,;\,\hat{\theta}_T\right)$ the deviance computed under the alternative hypothesis

$$\mathcal{H}_1 : g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right) = \mathbf{X}'\theta + \beta_T F\left(T\left(\mathbf{Z}\right)\right), \tag{4}$$

with a tree $T(\mathbf{Z}) \subseteq T_R(\mathbf{Z})$.

Let $r \leq R$ be the pre-specified maximal size of subtrees to be considered. A sequence of nested candidates subtrees $T_2(\mathbf{Z}) \subset \cdots \subset T_r(\mathbf{Z})$ of $T_R(\mathbf{Z})$ is constructed as follows:

- The procedure is forward with $T_1(\mathbf{Z})$ representing the root of the tree $T_R(\mathbf{Z})$. Let $\left\{T_j^m(\mathbf{Z}),\, m = 1,\ldots,n_j\right\}$ be the set of subtrees of $T_R(\mathbf{Z})$ with j leaves, such that for all $m = 1,\ldots,n_j$, T_{j-1} is a subtree of $T_j^m : T_{j-1}\left(\mathbf{Z}\right) \subset T_j^m\left(\mathbf{Z}\right)$.
- $T_j\left(\mathbf{Z}\right)$ is the subtree of $T_R(\mathbf{Z})$ with j leaves such that $T_{j-1}\left(\mathbf{Z}\right) \subset T_j\left(\mathbf{Z}\right)$, chosen as $T_j = T_j^{m^*}$ with

$$m^* = \arg\max_{m=1,\ldots,n_j}\left[D\left(X\,;\,\hat{\theta}_0\right) - D\left(X, T_j^m(\mathbf{Z})\,;\,\hat{\theta}_{T_j^m}\right)\right].$$

Step 3: tree selection

We select one of the trees of the sequence $T_1 \subset T_2 \subset \cdots \subset T_r$. For this selection step, we use either

- penalized maximum likelihood methods: the Akaïke information criterion(AIC) [10] and the Bayesian information criterion (BIC) [11],
- or a cross-validation method.

The competing models to be considered are:

$$\widehat{\mathcal{M}_j} : g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right) = \mathbf{X}'\hat{\theta}_{T_j} + \hat{\beta}_{T_j} F\left(T_j\left(\mathbf{Z}\right)\right), j = 1,\ldots,r \tag{5}$$

with $F\left(T_1\left(\mathbf{Z}\right)\right) \equiv 1$ representing the situation where the tree is reduced to the root node, that is the null model (3).

BIC and AIC criteria

The BIC criterion for the model $\widehat{\mathcal{M}_j}$ is

$$\mathrm{BIC}\left(\widehat{\mathcal{M}_j}\right) = 2\mathcal{L}\left(\widehat{\mathcal{M}_j}|\hat{\theta}_{T_j}, \hat{\beta}_{T_j}\right) - \delta_j \log\left(N\right),$$

N being the sample size, δ_j the number of free parameters involved in the model $\widehat{\mathcal{M}_j}$ ($\delta_j = \dim\left(\theta\right) + j$) and $\mathcal{L}\left(\widehat{\mathcal{M}_j}|\hat{\theta}_{T_j}, \hat{\beta}_{T_j}\right)$ the log-likelihood for the model $\widehat{\mathcal{M}_j}$.

The model selected by the BIC criterion is $\widehat{\mathcal{M}^{\mathrm{bic}}} = \widehat{\mathcal{M}_{j^{\mathrm{bic}}}}$, where j^{bic} is defined by

$$j^{\mathrm{bic}} = \arg\max_{j=1,\ldots,r} \mathrm{BIC}\left(\widehat{\mathcal{M}_j}\right).$$

We denote $T^{\mathrm{bic}} = T_{j^{\mathrm{bic}}}$ the tree used in the model $\widehat{\mathcal{M}^{\mathrm{bic}}}$.

The AIC criterion for the model $\widehat{\mathcal{M}_j}$ is

$$\mathrm{AIC}\left(\widehat{\mathcal{M}_j}\right) = 2\mathcal{L}\left(\widehat{\mathcal{M}_j}|\hat{\theta}_{T_j}, \hat{\beta}_{T_j}\right) - 2\delta_j,$$

with $\delta_j = \dim\left(\theta\right) + j$. The model selected by the AIC criterion is $\widehat{\mathcal{M}^{\mathrm{aic}}} = \widehat{\mathcal{M}_{j^{\mathrm{aic}}}}$ where j^{aic} is defined by

$$j^{\mathrm{aic}} = \arg\max_{j=1,\ldots,r} \mathrm{AIC}\left(\widehat{\mathcal{M}_j}\right).$$

We denote $T^{\mathrm{aic}} = T_{j^{\mathrm{aic}}}$ the tree used in the model $\widehat{\mathcal{M}^{\mathrm{aic}}}$.

Cross-validation criterion

As an alternative to the penalized maximum likelihood criteria presented above, we propose a cross-validation procedure on the global PLTR model for selecting the optimal tree. The competing models $\widehat{\mathcal{M}_j}$ are those defined in (5).

The original sample $\mathcal{A}$ is randomly partitioned into K equal size subsamples:

$$\mathcal{A} = \bigcup_{\ell=1}^{K} \mathcal{A}_\ell, \text{ with } \mathcal{A}_\ell \cap \mathcal{A}_m = \emptyset \text{ for all } \ell \neq m$$

For $\ell = 1,\ldots,K$, denotes by $\mathcal{A}_{-\ell} = \bigcup_{m\neq\ell} \mathcal{A}_m$ the ℓ^{th} training set, while $\mathcal{A}_\ell$ is the corresponding validation set.

For each $\ell = 1,\ldots,K$, the following steps are performed:

- fit the PLTR model (1) with the sample $\mathcal{A}_{-\ell}$. At the end of step 1, the fitted PLTR model is

$$g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right)=\mathbf{X}'\hat{\theta}_{T_R^\ell}+\hat{\beta}_{T_R^\ell}F\left(T_R^\ell\left(\mathbf{Z}\right)\right),$$

  where $T_R^\ell(\mathbf{Z})$ represents the maximal tree at convergence.
- Construct a sequence of $r-1$ nested subtrees $T_2^\ell(\mathbf{Z}),\ldots,T_r^\ell(\mathbf{Z})$ as in step 2, and determine the underlying PLTR models sequence:

$$\widehat{\mathcal{M}_j^\ell}: g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X},\mathbf{Z}\right)\right)=\mathbf{X}'\hat{\theta}_{T_j^\ell}+\hat{\beta}_{T_j^\ell}F\left(T_j^\ell(\mathbf{Z})\right),\ j=1,\ldots,r$$

- For each $j=1,\ldots,r$, use the validation sample $\mathcal{A}_\ell$ to compute the cross-validation error CV_j^ℓ of the model $\widehat{\mathcal{M}_j^\ell}$.

The mean cross-validation error is

$$\mathrm{CV}_j=\frac{1}{K}\sum_{\ell=1}^{K}\mathrm{CV}_j^\ell.$$

The selected model is $\widehat{\mathcal{M}^{\mathrm{cv}}}=\widehat{\mathcal{M}_{j^{\mathrm{cv}}}}$ where j^{cv} is defined by

$$j^{\mathrm{cv}}=\arg\min_{j=1,\ldots,r}\mathrm{CV}_j.$$

We denote $T^{\mathrm{cv}}=T_{j^{\mathrm{cv}}}$ the tree used in the model $\widehat{\mathcal{M}^{\mathrm{cv}}}$.

Step 4: Testing

To test the null hypothesis (3) versus the alternative (4), we use the statistic

$$\Lambda=2\mathcal{L}\left(\widehat{\mathcal{M}}_{\mathcal{H}_1}\right)-2\mathcal{L}\left(\widehat{\mathcal{M}}_{\mathcal{H}_0}\right).$$

As the model $\widehat{\mathcal{M}}_{\mathcal{H}_1}$ is not obtained as a maximum likelihood estimate, this statistic does not follow the "naïve" $\chi^2(j-1)$ distribution where j is the number of leaves of the tree used in $\widehat{\mathcal{M}}_{\mathcal{H}_1}$. Fan et al. [12] demonstrated that for a variety of models involving non parametric estimators, such generalized likelihood ratio statistics follow a scaled chi-squared distribution. In our case, this implies that for a defined number of leaves j the distribution of Λ is a scaled chi-squared distribution:

$$m\Lambda\sim\chi^2(b). \tag{6}$$

As the theoretical determination of m and b is cumbersome, Fan et al. propose to simulate the null distribution for estimating the constants m and b. In the following, we use the conditional parametric bootstrap procedure described below:

- Generate a new outcome $\mathbf{Y}^b$ from the fitted model $g\left(\mathbb{E}\left(\mathbf{Y}|\mathbf{X}\right)\right)=\mathbf{X}'\hat{\theta}_0+\hat{\beta}_0$
- Fit the complete model (2) with Y^b as the outcome (as in step 1)

$$g\left(\mathbb{E}\left(\mathbf{Y}^b|\mathbf{X},\mathbf{Z}\right)\right)=\mathbf{X}'\hat{\theta}_b+\hat{\beta}_{T_R^b}F\left(T_R^b\left(\mathbf{Z}\right)\right)$$

- Repeat the previous step until the size R is greater than j
- Construct a sequence of candidate optimal subtrees $\left\{T_k^b;\ k=2,\ldots,j\right\}$ as in step 2 (where we take $r=j$) and compute

$$\Lambda^b=2\mathcal{L}\left(\widehat{\mathcal{M}_j}\right)-2\mathcal{L}\left(\widehat{\mathcal{M}_1}\right)$$

- Repeat this process B times
- Estimate b and m from the empirical moments of sample $\Lambda^1,\ldots,\Lambda^B$.

Once b and m have been estimated, a p-value is calculated as $p=\mathbb{P}(X>m\Lambda)$ with $X\sim\chi^2(b)$.

Results

Simulation study

Simulation protocol

A simulation study with a binary outcome (logit link) was conducted to evaluate and compare the performances of the proposed procedure (with the three selection criteria) to the original one proposed by Chen et al.

We have considered three different scenarios for which we used PLTR logistic model similar to the one considered in Chen et al.

- In scenario 1, we simulated four Bernoulli variables G_1, G_2, G_3, G_4 with probabilities 0.3, 0.25, 0.18 and 0.22 respectively, and an outcome Bernoulli variable, denoted Y, according to the following model (null hypothesis):

$$\mathrm{logit}\,\mathbb{P}\left(Y=1|G_1,G_2,G_3,G_4\right)=\beta_1+\theta G_1$$

  with $\beta_1=\log(0.61)$, $\theta=\log(2)$. Here G_1 is the confounding variable and G_2, G_3, G_4 are the explanatory variables.
- In scenario 2, we introduced ten additional Bernoulli variables $G_5,\ldots,G_{14}$ with probabilities $p=0.5$. The Bernoulli variable Y is simulated according to the following model:

$$\begin{aligned}\mathrm{logit}\,\mathbb{P}\left(Y=1|G_1,\ldots,G_{14}\right)=\beta_1+\theta G_1+\beta_2\mathbf{1}_{G_2=1,G_3=0}\\+\beta_3\mathbf{1}_{G_3=1,G_4=0}\\+\beta_4\mathbf{1}_{G_3=1,G_4=1}.\end{aligned}$$

  with $\beta_1=\log(0.45)$, $\theta=\log(2)$, $\beta_2=\log(3.5)$, $\beta_3=\log(2)$ and $\beta_4=\log(4.5)$. This scenario mimics joint effects of G_2, G_3, and G_4. The corresponding tree is displayed in Figure (2). The variables $G_5,\ldots,G_{14}$ are noise variables unrelated to Y.

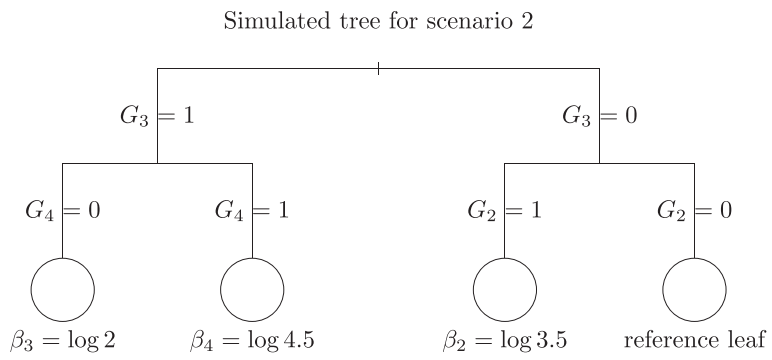

Figure 2 Tree used for scenario 2 simulations. The leaves are represented by circles and the number beneath each node represents the real value of the coefficient consider in each leaf of the tree.

- In scenario 3, we considered a deeper tree with non-independent explanatory variables $G_2, \ldots, G_5$. The model is:

$$\begin{aligned}\text{logit}\,\mathbb{P}\,(Y=1|G_1,\ldots,G_{15}) = \beta_1 + \theta G_1 + \beta_2 \mathbf{1}_{G_2=0,G_3=0,G_4=1} \\ + \beta_3 \mathbf{1}_{G_2=0,G_4=1} \\ + \beta_4 \mathbf{1}_{G_2=1,G_5=0} \\ + \beta_5 \mathbf{1}_{G_2=1,G_3=0,G_5=1} \\ + \beta_6 \mathbf{1}_{G_2=1,G_3=1,G_5=1}.\end{aligned}$$

with $\beta_1 = \log(1.8)$, $\theta = \log(1.35)$, $\beta_2 = \log(1.50)$, $\beta_3 = \log(2)$, $\beta_4 = \log(0.36)$, $\beta_5 = \log(2.5)$ and $\beta_6 = \log(0.36)$. The corresponding tree is displayed in Figure (3).
The Bernoulli variables $G_1, \ldots, G_5$ were generated from the following hierarchical model:

$$G_0 \sim \mathcal{B}(0.2), \quad \text{logit}\,\mathbb{P}(G_i = 1) = \text{logit}(0.2) + G_0 \quad \text{for } i = 2, \ldots, 5$$

and

$$\text{logit}\,\mathbb{P}(G_1 = 1) = \text{logit}\,(0.2) + \log(2) G_0.$$

Hence the variables $G_1, \ldots, G_5$ are marginally dependent. The variables $G_6, \ldots, G_{15}$ are considered as noise variables and are generated independently from a Bernoulli distribution with $p = 0.5$.

For all scenarios the sample was set to $n = 2000$, and 300 datasets were simulated.

Simulation results

Figures 4 and 5 display the quantile-quantile plots of the observed statistics for the "naïve" theoretical χ^2 distribution with degrees of freedom equal to the number of leaves minus 1, and for the scaled χ^2 distribution (equation 6). These figures show that the naïve distribution is inadequate; in contrast, the scaled distribution with estimated m and b fits well the empirical distribution.

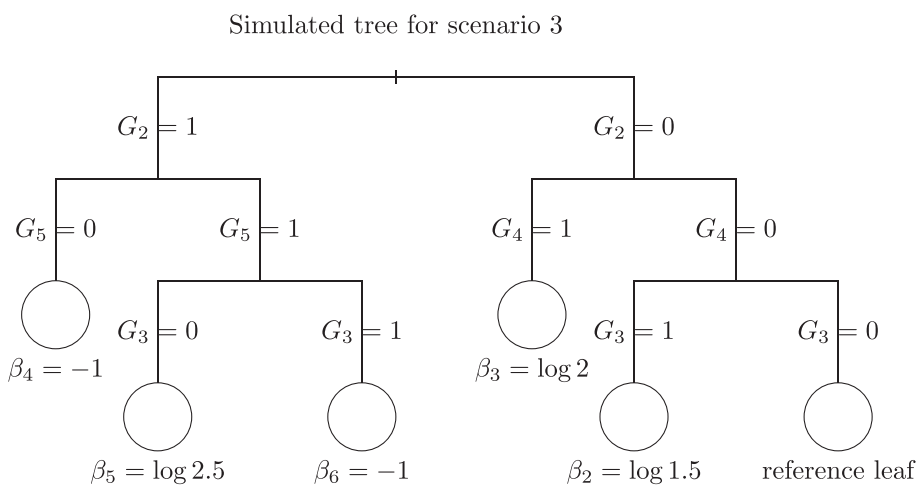

Figure 3 Tree used for scenario 3 simulations. The leaves are represented by circles and the number beneath each node represents the real value of the coefficient consider in each leaf of the tree.

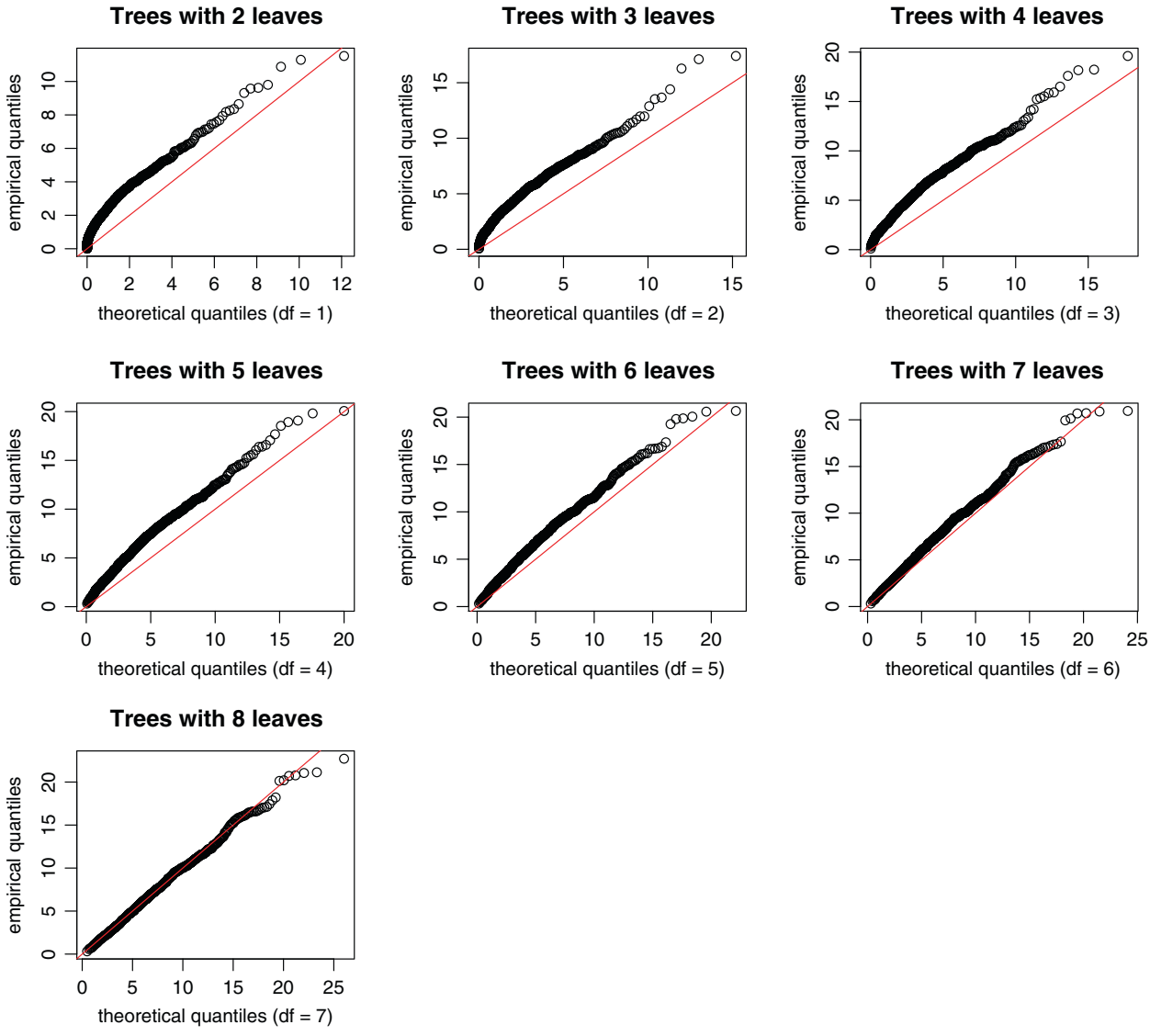

Figure 4 Quantile-quantile plots of the observed statistics versus the "naïve" χ^2 quantiles.

We assess whether or not the trees selected by step 3 in the sequence of nested trees have the correct number of leaves. Under scenario 1 (null hypothesis), a root tree (one leaf) is expected. As seen in Table 1, the procedure with the BIC criterion (BIC) selects the root tree for 98.3% of the simulations, whereas the Chen et al. procedure (named BOOT hereafter) succeeds for only 91% of the simulations. For the 10-fold cross validation procedure (CV), this proportion goes down to 84%, and for the AIC criterion (AIC) it is only 47.3%.

Under scenario 2, the correct number is of 4 leaves. As seen from Table 2, BIC has the best performance with 44.3% of selected trees with four leaves. Moreover, it exhibits the smallest dispersion around the target value. In contrast BOOT selects a tree with only 2 leaves for all the simulations. The performances of CV are inferior to those from BIC, and the dispersion is higher. Finally, AIC selects always trees with too many leaves. Similar results are obtained with scenario 3 (where the correct number of leaves is 6), with increased quality of CV (Table 3).

For the more complex scenario (scenario 3), we computed the ten-fold cross-validation generalization error for each of the 300 simulated data sets for BIC, AIC and CV criteria. The distribution of the generalization errors are displayed in Figure 6. CV and BIC have very similar errors, while AIC have a slightly increased error.

In summary, the procedure using the BIC criterion consistently outperforms the other procedures.

We also investigated which variables are present in the splits of the trees selected by BIC under scenario 2 and 3

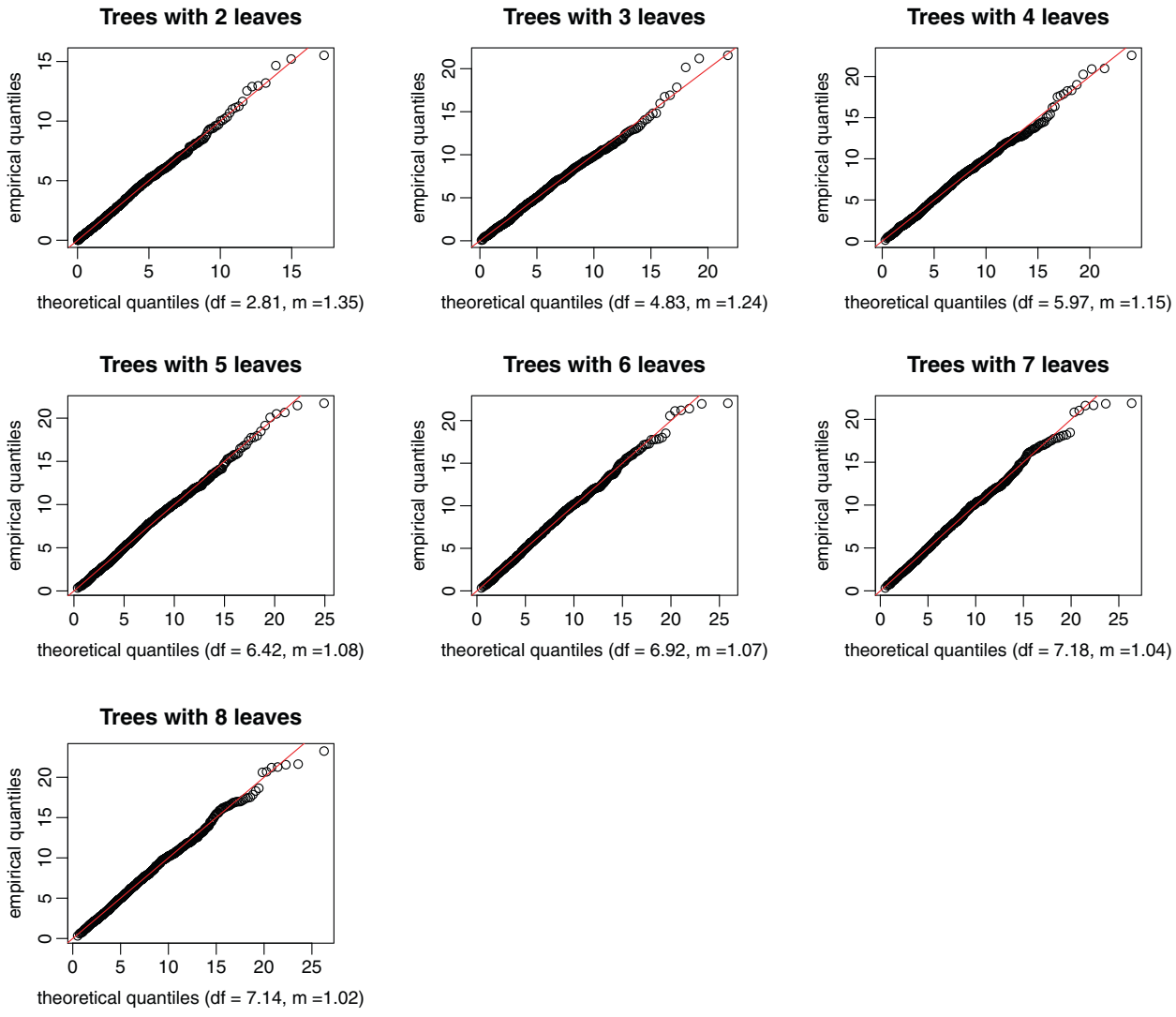

Figure 5 Quantile-quantile plots of the observed statistics versus the scaled χ^2 quantiles.

(Tables 4 and 5). For scenario 2, the so-called correct variables are G_2 to G_4, and in scenario 3, G_2 to G_5. In both scenarios, we refer as incorrect variables the ten noise variables. Under scenario 2, in 18% of the selected trees, at least one noise variable appears; however, all three correct variables are present in 44.3% of the selected trees. In all cases, at least two correct variables were selected. The BIC procedure behaves better under scenario 3, with more than 99% of trees involving all four correct variables, while noise variables appear in 20% of the trees.

Table 1 Number of trees by number of leaves, for the 300 trees selected by the different methods under scenario 1

Leaves	1	2	3	4	5	6
BOOT	273	8	8	7	1	3
CV	252	0	18	10	10	10
BIC	295	4	1	0	0	0
AIC	142	64	54	30	5	5

Table 2 Number of trees by number of leaves, for the 300 trees selected by the different methods under scenario 2

Leaves	1	2	3	4	5	6	7	8	9	10
BOOT	0	300	0	0	0	0	0	0	0	0
CV	0	18	83	61	36	32	21	19	13	17
BIC	0	0	112	133	46	7	2	0	0	0
AIC	0	0	0	0	0	0	3	8	24	265

Analysis of lung adenocarcinomas

Description of the data

The dataset considered in this study is based on a French-Singaporean study (Merlion study) of 230 patients with lung adenocarcinomas [13]. The Western-Europe series

Table 3 Number of trees by number of leaves, for the 300 trees selected by the different methods under scenario 3

Leaves	1	2	3	4	5	6	7	8	9	10
BOOT	0	300	0	0	0	0	0	0	0	0
CV	0	0	41	16	22	154	22	12	17	16
BIC	0	0	0	1	89	162	36	9	3	0
AIC	0	0	0	0	0	0	0	1	2	297

(WE) included 139 tumors and the East-Asian series (EA) included 91 tumors. Clinical characteristics were detailed in a previous published article [13]. DNA was extracted using standard protocols and stored at −80°C until use. Copy number information was issued from Affymetrix Genome-Wide Human SNP 6.0 arrays. Inferences about the copy number status of each genomic segment (copy loss, copy modal, copy gain) were obtained using the modified CGHmix classification procedure [14]. In order to summarize genomic information while keeping a sufficient level of resolution, copy number status was averaged (median estimate) over the 284 main cytogenetic bands. Information about KRAS mutation was extracted from the targeted mutation profiling performed using the Sequenom Massarray 4 platform (Sequenom, San Diego, CA). Here, the KRAS mutation status was defined as the presence or absence of any mutation within KRAS gene. In this dataset, we detected 54 KRAS mutations with 44 cases (31.6%) from the WE series and 10 cases (10.9%) from the EA series.

We compared the results obtained from the Chen et al. procedure [8] to those obtained by the novel procedure with the BIC criterion. The "dependent" variable was the KRAS mutation status (mutation/wild-type). The cohort status (WE/EA) was the confounding binary variable. The 284 copy-number alterations (trinomial variable: copy-loss, modal, copy gain) were considered as candidate explanatory variables. Recursive partitioning stopped as soon as the number of cases in each terminal node was below fifteen.

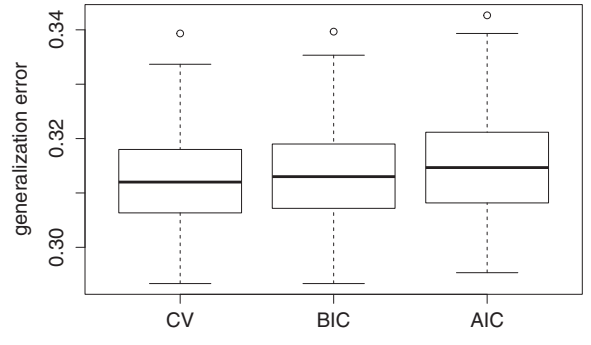

Figure 6 Distribution of the generalization 10-fold cross-validation error for AIC, BIC, CV criteria across the 300 simulated data sets.

Table 4 Variables selected by the procedure using BIC criterion under scenario 2, with global percentages between brackets

		Incorrect variables					
		0	**1**	**2**	**3**	**4**	**5**
Correct	0	0	0	0	0	0	0
Variables	1	0	0	0	0	0	0
	2	134 (44.66%)	22 (7.33%)	10 (3.33%)	0	0	1 (0.33%)
	3	112 (37.33%)	19 (6.33%)	2 (0.66%)	0	0	0

Results

The iterative procedure converged after 15 iterations. The trees selected by Chen's et al. method and by our procedure with the BIC criterion are displayed in Figure 7. Chen's et al. procedure led to two leaves that separated tumors with and without copy-loss of 3q23. The global adjusted p-value associated with the selected tree is 0.0055. This model is a simple cohort-adjusted logistic regression model with 3q23 copy-loss as the unique explanatory variable.

Our procedure with the BIC criterion led to seven leaves. We identified:

(i) two pure or nearly pure wild-type KRAS leaves (with 53 tumors and only one KRAS mutation) characterized by no 3q23 copy-loss and a copy-loss for either 9q12 or 3p11 cytoband,
(ii) a leave with a low rate of KRAS-mutated tumors (8%) characterized by no copy-loss of 3q23, 3p11, 9q12, 14q23 but a copy gain of 8q23 cytoband,
(iii) a leave with a medium rate of KRAS-mutated tumor (15.5%) with no copy-loss of 3q23, 3p11, 14q23 and no copy-gain of 8q23 and 1q32 cytoband,
(iv) the three other leaves were heterogeneous with a mixture of wild-type and KRAS-mutated tumors (43.5%, 52.6%, 51.2%).

Table 5 Variables selected by the procedure using BIC criterion under scenario 3, with global percentages between brackets

		Incorrect variables					
		0	**1**	**2**	**3**	**4**	**5**
Correct	0	0	0	0	0	0	0
variables	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
	3	1 (0.33%)	1 (0.33%)	0	0	0	0
	4	239 (79.66%)	50 (16.66%)	8 (2.66%)	1 (0.33%)	0	0

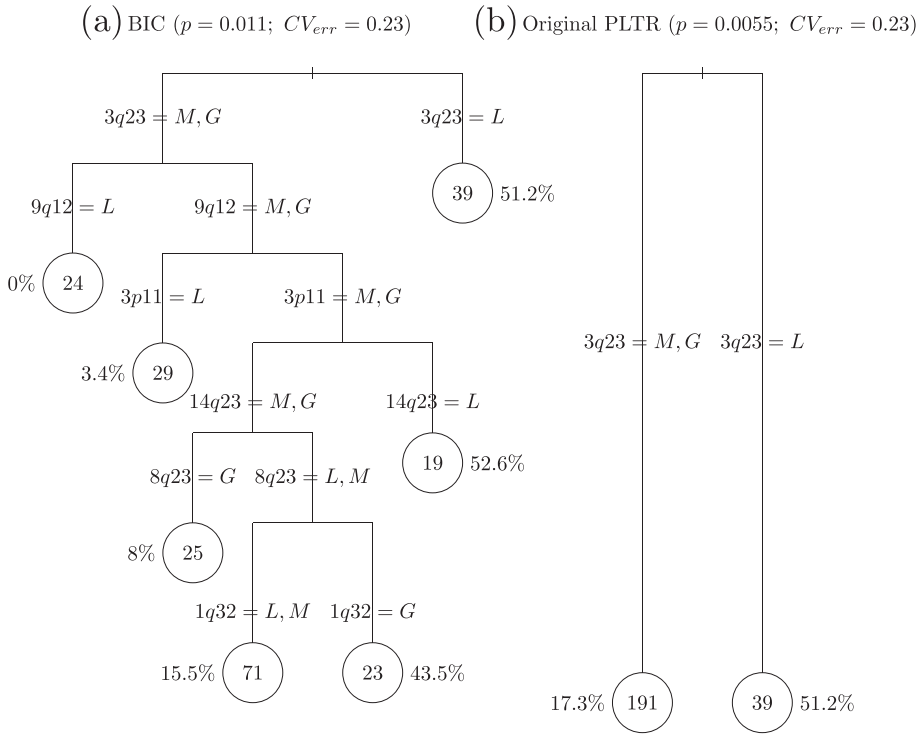

Figure 7 Optimal tree obtained with the two competing methods on the real data set: (a) BIC selected tree, (b) Original PLTR selected tree. The leaves are represented by circles and the number in each leave node represents the number of observations falling inside the node; the percentage represented proportion of cases inside the node.

For the selected tree, the split variables are copy-number aberrations of 1q32, 3p11, 3q23, 8q23, 9p12, and 14q23. The global p-value associated with this tree is 0.011 with a ten-fold cross-validation generalization error of 0.23. These results were obtained after adjustment for a significant cohort effect ($OR = 0.266$, 95% Confidence interval: $[0.12 - 0.56]$) with a higher rate of KRAS for the WE series as compared to the EA series.

We also compared the characteristics of the 53 tumors arising from the two pure or nearly pure wild-type KRAS leaves as compared to the other tumors. There was no significant difference between the two groups regarding the EGFR mutation status ($p = 0.94$). There was a significantly higher proportion of tumors with a large fraction of genome altered (more than 50%) in the pure or nearly pure wild-type KRAS group as compared to the other groups ($p = 1.7 \times 10^{-8}$).

Discussion

Nowadays, there is a growing interest in deciphering the genomic spectrum of clinical disease entities. In this context, recursive partitioning methodology provides a powerful data mining tool for exploring complex interplay between genomic factors, with respect to a main factor, that can reveal hidden genomic patterns. The requirement of adjusting for confounding factors led Chen et al. to develop a semiparametric regression model called PLTR together with an iterative algorithm procedure to select and test the "optimal" tree. A main drawback of the procedure is that it relies on a two levels permutation strategy which can become cumbersome and computationally expensive. In this work, we propose a novel procedure with different selection criteria. As shown from the simulation study, the proposed procedure with the BIC criterion achieves good power to detect the hidden structure as compared to Chen's et al procedure.

When investigating patterns of copy-number alterations in lung adenocarcinomas, with respect to KRAS mutation status and after adjustment for a cohort effect, our proposed strategy highlights two subgroups of pure or nearly pure wild-type KRAS tumors. These subgroups correspond to 53 lung adenocarcinomas having no 3q23 copy-loss but copy-loss for either 9p12 or 3p11 cytoband. It is worth noting that the 3q23 area harbors the PI3KCB gene that participates in the PI3K (Phosphatidylinositol 3-kinase) signaling pathway, well-known to be deregulated in many human cancers. Moreover, PI3K is one of the main effector pathways of RAS, regulating cell growth, cell cycle and cell survival. These wild-type KRAS subgroups are not enriched for EGFR mutation (mutually exclusive

with KRAS mutation) and are composed of tumors having a proportion of copy-number changes significantly higher than expected by chance. The genomic patterns of these two wild-type KRAS subgroup are worth further investigation.

Conclusion

We have proposed a novel recursive partitioning procedure for deciphering the genomic spectrum of clinical disease entities. The proposed procedure represents a powerful and practical alternative to the partially linear tree-based regression model proposed by Chen et al. [8]. Our procedure performs well, is simple to implement, less computationally demanding and can be recommended for investigating new disease taxonomy. The procedure is implemented within an R package known under the acronym 'GPLTR' and will be available very soon on the CRAN site.

We plan to use this novel procedure to identify new sub-groups of multiple sclerosis treated with interferon-beta, with regards to the occurrence of antidrug-antibody response, while adjusting for cohort effect.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CM and WT implemented the proposed procedure. PB coordinated the study. All authors participated in the design of the procedure and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This research is supported in part by the IMI-funded ABIRISK project (http://www.abirisk.eu/).

Author details

[1] Abirisk consortium WP4, 14-16 Avenue Paul-Vaillant-Couturier, 94807 Villejuif, France. [2] Inserm U669, 14-16 Avenue Paul-Vaillant-Couturier, 94807 Villejuif, France. [3] Faculty of Medicine Paris-Sud, 63 rue Gabriel Peri, 94276 Le Kremlin-Bicêtre, France. [4] Assistance Publique – Hôpitaux de Paris, Hôpital Paul Brousse, Villejuif, France.

Received: 23 December 2013 Accepted: 8 April 2014
Published: 16 April 2014

References

1. Roberts P, Stinchcombe T: **Kras mutation: should we test for it, and does it matter?** *J Clin Oncol* 2013, **31**(8):1112–21.
2. Rajagopalan H, Lengauer C: **Aneuploidy and cancer.** *Nature* 2004, **432:**338–341.
3. Breiman L, Olshen JH, Stone CJ: *Classification and Regression Trees*. Belmont, California: Wadsworth International Group; 1984.
4. Breiman L: *Random forest*. Technical Report, Department of Statistics, University of California at Berkeley. 2002.
5. Diaz-Uriarte R, Alvarez de Andrés S: **Gene selection and classification of microarray data using random forest.** *BMC Bioinformatics* 2006, **7**(1):1–13.
6. Guan X, Chance MR, Barnholtz-Sloan JS: **Splitting random forest (srf) for determining compact sets of genes that distinguish between cancer subtypes.** *J Clin Bioinform* 2012, **2**(1):1–12.
7. Liaw A, Wiener M: **Classification and regression by randomforest.** *R News* 2002, **2**(3):18–22.
8. Chen J, Yu K, Hsing A, Therneau TM: **A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects.** *Genet Epidemiol* 2007, **31:**238–251.
9. Yu K, Wheeler W, Li Q, Bergen AW, Caporaso N, Chatterjee N, Chen J: **A partially linear tree-based regression model for multivariate outcomes.** *Biometrics* 2010, **66**(1):89–96.
10. Akaike H: **A new look at the statistical model identification.** *IEEE Trans Automat Control* 1974, **AC-19:**716–723.
11. Schwarz G: **Estimating the dimension of a model.** *Ann Stat* 1978, **6:**461–464.
12. Fan J, Zhang C, Zhang J: **Generalized likelihood ratio statistics and wilks phenomenon.** *Ann Stat* 2001, **29**(1):153–193.
13. Broët P, Dalmasso C, Tan E, Alifano M, Zhang S, Wu J, Lee M, Régnard J, Lim D, Koong H, Agasthian T, Miller L, Lim E, Camilleri-Broët S, Tan P: **Genomic profiles specific to patient ethnicity in lung adenocarcinoma.** *Clin Cancer Res* 2011, **17**(11):3542–50.
14. Dalmasso C, Broët P: **Detection of chromosomal abnormalities using high resolution arrays in clinical cancer research.** *J Biomed Inform* 2011, **44**(6):936–942.

doi:10.1186/2043-9113-4-6
Cite this article as: Mbogning *et al.*: A novel tree-based procedure for deciphering the genomic spectrum of clinical disease entities. *Journal of Clinical Bioinformatics* 2014 **4**:6.