Article:
Gusnanto, AS, Tcherveniakov, P, Shuweihdi, F et al. (3 more authors) (2015) Stratifying tumour subtypes based on copy number alteration profiles using next-generation sequence data. Bioinformatics, 31 (16). 2713 - 2720. ISSN 1367-4803

https://doi.org/10.1093/bioinformatics/btv191

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Stratifying tumour subtypes based on copy number alteration profiles using next-generation sequence data

Arief Gusnanto [1]*, Peter Tcherveniakov [2], Farag Shuweihdi [1], Manar Samman [3,4], Pamela Rabbitts [3], and Henry M. Wood [3]

[1]Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom
[2]Department of Thoracic Surgery, St. James Hospital, Leeds LS9 7TF, United Kingdom
[3]Leeds Institute of Cancer and Pathology, University of Leeds, Leeds LS9 7TF, United Kingdom
[4]King Fahd Medical City, Riyadh, Saudi Arabia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The role of personalised medicine and target treatment in the clinical management of cancer patients has become increasingly important in recent years. This has made the task of precise histological substratification of cancers crucial. Increasingly genomic data are being seen as a valuable classifier. Specifically, copy number alteration (CNA) profiles generated by next-generation sequencing (NGS) can become a determinant for tumours subtyping. The principle purpose of this study is to devise a model with good prediction capability for the tumours histological subtypes as a function of both the patients covariates and their genome-wide CNA profiles from NGS data.

Results: We investigate a logistic regression for modelling tumour histological subtypes as a function of the patients' covariates and their CNA profiles, in a mixed model framework. The covariates, such as age and gender, are considered as *fixed* predictors and the genome-wide CNA profiles are considered as *random* predictors. We illustrate the application of this model in lung and oral cancer datasets, and the results indicate that the tumour histological subtypes can be modelled with a good fit. Our cross-validation indicates that the logistic regression exhibits the best prediction relative to other classification methods we considered in this study. The model also exhibits the best agreement in the prediction between smooth-segmented and circular binary-segmented CNA profiles.

Availability: An R package to run a logistic regression is available in http://www1.maths.leeds.ac.uk/∼arief/R/CNALR/

Contact: a.gusnanto@leeds.ac.uk

1 INTRODUCTION

Next-generation sequencing (NGS) technology has greatly transformed our way of interrogating genomes and advanced our understanding of genomic changes such as copy number alterations (CNA). We have previously shown that CNA data can be generated from routine diagnostic biopsy specimens in a very efficient manner using multiplexed, low-coverage sequencing (Wood *et al.* , 2010) and developed a robust algorithm to analyse these data in individual patients (Gusnanto *et al.* , 2012). Our objective in this current study is to analyse multiple samples from two tumour subtypes simultaneously and develop informatic techniques to separate them using their CNA profiles, so that subsequent patients could be better stratified in a predictive manner.

Currently, methods for studying groups of CNA profiles tend to focus on looking for similarities in a homogeneous group of samples (Mermel *et al.* , 2011) or to look for regions of the genome where two groups of samples are very different (de Ronde *et al.* , 2010). We are focusing less on the genomic regions involved, and more on the problem of separating similar groups of samples and predicting to which group a new sample belongs. More specifically, our focus is in utilising the whole (genome-wide) CNA profiles to stratify tumour subtypes, in addition to other covariates or patients characteristics. This strategy enables us to have a wider picture of the contribution of each genomic region in the stratification, in light of the contribution of the other regions.

To model the tumour subtypes, we investigate logistic regression within a random effects model framework, where the contributions of patients' clinical characteristics are considered as fixed effects and those of the CNA profiles are considered as random effects. Since our interest is in the prediction of tumour subtypes of new samples, we perform a cross-validation to identify the prediction error of the model. We compare this prediction error with those from other classification methods.

As a test data set, we have analysed a cohort of lung tumours, all of which are from the two closely related subtypes, squamous cell carcinoma (SCC) and adenocarcinoma (ADC). Until recently these have been treated as part of the larger subgroup non-small cell lung carcinoma (NSCLC), but there is growing evidence that the two groups should be treated as separate diseases (Gazdar , 2010).

As a validation data set, we analysed another cohort of two related tumour subtypes, oral squamous cell carcinoma (OSCC) and oral verrucous carcinoma (OVC). These subtypes have very different prognoses and clinical management, but there can be histological uncertainty in distinguishing between them (Rekha and Angadi , 2010).

The results of our analysis indicate that the logistic regression enables us to stratify tumour subtypes and investigate genome-wide contribution of each genomic region to the stratification. The

*to whom correspondence should be addressed

results also indicate that the logistic regression has good predictive value for new samples: advantageously, this prediction is achieved regardless of the segmentation methods involved in the estimation of CNA. The latter advantage is important because a number of tools exist to obtain CNA estimates from next-generation sequencing data, and they use a variety of pre-processing steps, such as normalisation, and a number of different segmentation methods.

Before we discuss those results, we discuss first the logistic regression in the next section.

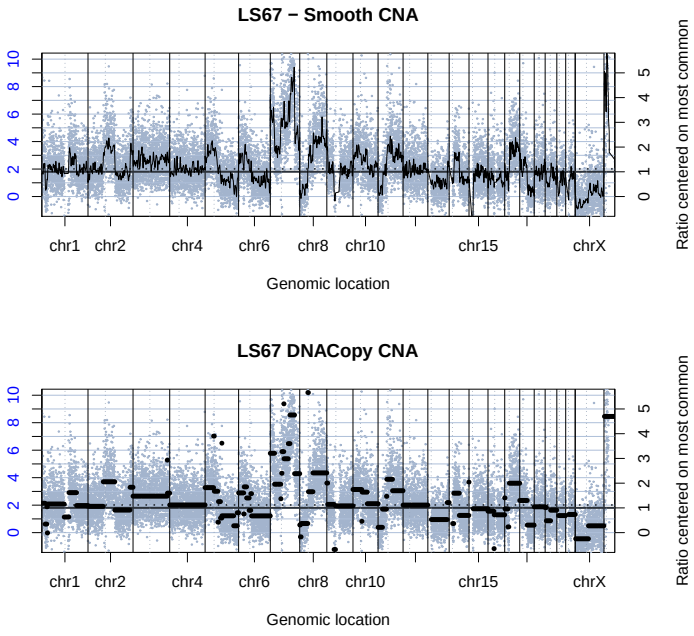

Fig. 1. *Example of CNA estimates from one patient, LS67.*

2 METHODS

2.1 Patients, sequence data and alignments

Seventy six lung cancer patients were included in this study from Leeds Teaching Hospital (UK), comprising of two groups: squamous carcinoma (38 patients) and adenocarcinoma (38 patients). We recorded the patients' clinical characteristics, but in this study we only consider age and gender as covariates. For a validation of our methods, samples from 102 oral cancer patients were collected from Leeds Teaching Hospitals (UK), Queen Victoria Hospital (Sussex, UK), University of Torino (Italy) and the National Guard Hospital (Saudi Arabia), comprising of 45 OSCC samples and 57 OVC samples. Unfortunately, for this cohort of patients, we could not obtain the patients' clinical characteristics, and only work with their CNA profiles. The main body of this manuscript will deal largely with the lung cancer dataset. The experimental validation using the oral cancer dataset will mainly be presented in the supplementary material.

Details on sample preparation, DNA extraction and library preparation are described by Wood *et al*. (2010). Sequences were aligned using the bwa suite version 0.5.9-r16 (Li and Durbin, 2009) against assembly hg19 of the human genome. Only sequences that could be uniquely aligned and with mapping quality ≥ 37 were used.

2.2 Normalisation and CNA estimate

The copy number alteration (CNA) profile from each lung tumour is calculated by 'depth of coverage' from their sequences. For this purpose, the optimal window size for this group of samples is estimated using *NGSoptwin* package to be 150 kbp (Gusnanto *et al.* , 2014). The sequence data from 76 cancer patients are not directly comparable because inevitably the tumour samples are contaminated with normal cells by different degrees. To deal with this problem, we performed a normalisation using the CNAnorm package Gusnanto *et al*. (2012) to obtain the CNA estimates. An example of CNA estimates is presented in Figure 1 for patient LS67, which can be in two forms depending on the segmentation method used:

1. Smooth estimate, where CNA is estimated as smooth segmented lines Huang *et al*. (2007). The main characteristic of the estimate is that the segmented line is smooth and follows the sudden 'jumps' and 'drops' in the CNA profile. This is illustrated in the top row of Figure 1.
2. DNACopy estimate, where CNA is estimated as circular binary segmented lines (Olshen *et al.* , 2004). The main characteristic of the estimate is that the segmented line tends to form relatively long constant segments. This is illustrated in the bottom row of Figure 1.

With 150 kbp window size, we approximately have slightly more than 20,000 windows to cover the whole genome. In this study, we exclude the CNA estimates from the sex chromosomes and the centromere regions, where some missing values can be problematic in the analysis. After we remove them, we have CNA estimates from 17,571 genomic windows in two different forms as described above. For each form of CNA estimate, the CNA profiles from the patients are summarised in a matrix of size 76 by 17,571.

2.3 Logistic regression

At this stage, our data consist of three components: (1) the binary tumour histological subtype (squamous carcinoma=1, adenocarcinoma=0), (2) *age* and *gender* as covariates, and (3) the CNA estimates across 17,571 genomic windows. To model the binary histological subtype, logistic regression is a natural choice because we can extend the standard model to deal with the second type of predictors namely CNA profiles. One of the immediate challenges dealing with this type of data is the dimension of the CNA profiles. We have a data matrix of size $n \times q$ where $76 = n \ll q = 17,571$. A standard logistic regression will fail in the computation because the number of variables far exceeds the number of observations. To proceed, we extend the standard model by assuming that the parameters of CNA to follow a Normal distribution.

Specifically, in vector notation, consider y as a vector of binary tumour histology with $y_i = 1$ if the tumour is squamous carcinoma and $y_i = 0$ if the tumour is adenocarcinoma, for $i = 1, 2, \ldots, n$. We assume that y follow a Binomial(1,μ) distribution with μ is the probability of the tumour being squamous carcinoma. In logistic regression, we model

$$h(\mu) = X\beta + Z\gamma \tag{1}$$

where $h(\cdot)$ is a logit link function that applies elementwise, X is a matrix of fixed covariates of size $n \times p$, β is a p-vector of fixed effects for matrix X, Z is matrix of CNA estimates of size $n \times q$, and γ is a q-vector of random effects for matrix Z. We assume that $\gamma_j, j = 1, \ldots, q$ would follow a Normal(0,τ^2) distribution.

The joint log likelihood of the parameters is given by

$$\log L(\beta, \gamma, \tau^2) = \log p(y|\gamma) + \log p(\gamma) \tag{2}$$

where $p(y|\gamma)$ is the likelihood based on conditional distribution of y given γ, and $p(\gamma)$ is the likelihood based on the Normal distribution of random effects γ. Given the data $y_1, y_2, \ldots, y_n$, the log likelihood is given by

$$\log L(\beta, \gamma, \tau^2) = \sum_i \{y_i \log \mu_i + (1-y_i)\log(1-\mu_i)\} - \frac{1}{2}\lambda \sum_j \gamma_j^2 \tag{3}$$

where μ_i is a logistic function of β and γ, and $\lambda = \frac{1}{\tau^2}$.

2.3.1 Estimation of β and γ at fixed λ The estimation of β and γ in (3) can be performed using iterative weighted least squares (IWLS) at a fixed value of λ. At a fixed value of λ and starting values of β^0 and γ^0, IWLS is performed to find the solution of mixed model equation

$$\begin{pmatrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & Z'\Sigma^{-1}Z + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} X'\Sigma^{-1}Y \\ Z'\Sigma^{-1}Y \end{pmatrix} \quad (4)$$

where "$\prime$" denotes transpose,

$$Y_i = x_i'\beta + z_i'\gamma + \frac{y_i - \mu_i}{\mu_i(1-\mu_i)}$$

is the working vector, Σ is a diagonal matrix with

$$\Sigma_{ii} = \mu_i(1-\mu_i),$$

$\mu = (1 + \exp\{-(X\beta + Z\gamma)\})^{-1}$, and $D \equiv \tau^2 I_q$ where I_q is an identity matrix of size q.

Specifically, in the m-th iteration,

1. update the working vector $Y^{(m)}$ and $\Sigma^{(m)}$ based on $\beta^{(m-1)}$ and $\gamma^{(m-1)}$
2. calculate $\gamma^{(m)} = (Z'\Sigma^{-1}Z + D^{-1})^{-1}Z'\Sigma^{-1}Y$
3. calculate $\beta^{(m)} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$

where in steps (2) and (3) we use the current values of Σ and Y. We iterate the above steps until convergence and, at convergence, we obtain the estimates $\widehat{\beta}$ and $\widehat{\gamma}$. The dimension of matrix $(Z'\Sigma^{-1}Z + D^{-1})$ is in the order of $17,571 \times 17,571$ so that its inversion is a computational bottle neck in the iteration. To overcome this problem, we use a computational modification that are discussed further in Section 2.4.

The approximate standard errors for $\widehat{\beta}$ and $\widehat{\gamma}$ are given by the square root of the diagonal of

$$(X'S^{-1}X)^{-1} \quad (5)$$

and

$$(Z'\Sigma^{-1}Z + D^{-1})^{-1}, \quad (6)$$

respectively, where $S = \Sigma + \tau^2 ZZ'$.

2.3.2 Estimation of τ^2 given β and γ We estimate β and γ above at a given value of τ^2, and τ^2 can be estimated as the one that minimises Akaike's Information Criterion (AIC) or through cross-validation. The calculation of AIC in practice is described further in Section 2.4.

2.4 Computational consideration

The size of the matrix $(Z'\Sigma^{-1}Z + D^{-1})$ involved in the above computation is in the order of 17,571 × 17,571. To reduce the computational burden we consider a modification of the computation to find the solution of the above estimation problem. We can overcome this problem by representing the matrix Z (of size $n \times q$) as a multiplication of special matrices in singular value decomposition (SVD) (Eilers *et al.* , 2001). Let us define

$$Z = UQV'$$

where U and Q are of size $n \times n$, and V is of size $q \times n$, such that $U'U = I_n$, $V'V = I_n$ (but $VV' \neq I_q$), and Q is a diagonal matrix of singular values of Z.

Using this decomposition, the relevant part in 4 can be rewritten as

$$\left((UQ)'\Sigma^{-1}(UQ) + \lambda I_q\right)\gamma^* = (UQ)'\Sigma^{-1}Y \quad (7)$$

where $\gamma = V\gamma^*$. Based on this representation, the updating equation in step (2) in the above iteration gives an update for γ^* (hence also γ), and at convergence we obtain the estimates $\widehat{\beta}$ and $\widehat{\gamma}$ (via $\widehat{\gamma}^*$). The SVD is only done once before IWLS, as the quantities that are updated in the above iteration are only $Y^{(m)}$, $\Sigma^{(m)}$, $\beta^{(m)}$, and $\gamma^{(m)}$ (via $\gamma^{*(m)}$). With this representation, the dimension of the matrix inversion reduces dramatically from $q \times q$ to a managable $n \times n$.

In the estimation of τ^2 (Section 2.3.2), we minimise the AIC

$$\text{AIC} = -2\log L(\widehat{\mu}) + 2\text{df}$$

where

$$L(\mu) = \sum_i \{y_i \log \mu_i + (1-y_i)\log(1-\mu_i)\}$$

and $\widehat{\mu}$ is calculated in a logistic function using $\widehat{\beta}$ and $\widehat{\gamma}$ (via $\widehat{\gamma}^*$). The degrees of freedom is computed as

$$\text{df} = \text{trace}\left\{\left((UQ)'\Sigma^{-1}(UQ) + \lambda I\right)^{-1}\left((UQ)'\Sigma^{-1}(UQ) + \lambda I\right)\right\}$$

(see Pawitan (2001)).

2.5 Cross validation

In building a logistic regression to model tumour histology, we wanted to consider how well the model can be employed to make prediction of new observations. To obtain the prediction of the histology based on the CNA profiles, we performed a cross validation. We randomly split n observations into training set of size n_t and validation set of size n_v ($n_t + n_v = n$) such that

$$y := \begin{bmatrix} y_t \\ \cdots \\ y_v \end{bmatrix}, X := \begin{bmatrix} X_t \\ \cdots \\ X_v \end{bmatrix}, Z := \begin{bmatrix} Z_t \\ \cdots \\ Z_v \end{bmatrix}.$$

The training set serves as the set by which we estimate the model parameters $\widehat{\beta}_t$ and $\widehat{\gamma}_t$. The estimates are then used in the validation set to obtain model prediction

$$\widehat{y}_v = I\left(h^{-1}\left\{X_v\widehat{\beta}_t + Z_v\widehat{\gamma}_t\right\} \geq 0.5\right), \quad (8)$$

where $I(\cdot)$ equals one (squamous group) if the expression inside the brackets is true, and zero (adenocarcinoma group) otherwise.

From this prediction, we obtain how many tumour histologies in the validation set are misclassified (classification error). Denoting $y_v = (y_{v1}\ y_{v2}\ \cdots\ y_{vn_v})'$ and, from (8) $\widehat{y}_v = (\widehat{y}_{v1}\ \widehat{y}_{v2}\ \cdots\ \widehat{y}_{vn_v})'$, we define the classification error as

$$CE = \sum_{k=1}^{n_v} I(y_{vk} \neq \widehat{y}_{vk}). \quad (9)$$

We also calculate this error as percentages out of n_v, which are presented in the supplementary material.

In our application, we perform the cross validation 100 times where, out of 76 observations, 38 observations are randomly selected to be in the training set and the remaining 38 observations are in the validation set. To see whether the classification error obtained by logistic regression is within a reasonable range, we need to consider other classification methods in the cross validation. This enables us to get a better view of the performance of the logistic regression model (Section 2.3) in the prediction of tumour histology.

Some classification models that we consider in the cross validation are: (1) k-nearest neighbour (KNN) with $k = 1, \ldots, 7$, (2) diagonal-quadratic and diagonal-linear discriminant-analysis (DA), (3) partial least squares (PLS), (4) elastic net, (5) lasso, (6) neural network, (7) support vector machines, and (8) smoothed logistic regression (SLR, Huang *et al.* (2009)). For this purpose, we only use the CNA profiles based on the smooth and CBS segmentation, i.e. X_t and X_v are just vector of ones and $\widehat{\beta}_t$ is just an estimate of fixed intercept (from the training set). Note that the elastic-net and lasso models are essentially logistic regression with different penalties to the likelihood that that we describe in Section 2.3. To avoid confusion, we reserve the term "logistic regression" for the one we describe in Section 2.3. We use the terms "elastic-net" and "lasso" to refer to the logistic regression with elastic net and lasso penalties, and use the acronym "SLR" to refer to the smoothed logistic regression.

2.6 Agreement in prediction between segmentation methods

A concern that we have in the analysis is whether the choice of segmentation method that we use to estimate CNA really does determine whether a patient's tumour histology is misclassified or not, given a classification model. In other words, if a patient's tumour is predicted as adenocarcinoma based on smooth-segmented CNA, is it also predicted as adenocarcinoma based on CBS-segmented CNA? To answer this question, we also estimate the agreement in prediction between these two very different segmentation methods.

Let us denote Z^s and Z^d as CNA profiles based on the smooth and CBS segmentation, respectively. Let us also denote $\widehat{\beta}_t^s$ and $\widehat{\gamma}_t^s$ to be the fixed and random parameter estimates from the training set using CNA profiles based on smooth segmentation. Similarly, we denote $\widehat{\beta}_t^d$ and $\widehat{\gamma}_t^d$ as the estimates in the training set using CNA profiles based on CBS (DNACopy) segmentation. We distinguish the predicted tumour histology in the validation set using CNA profiles based on the smooth and CBS segmentation as

$$\widehat{y}_v^s = I\left(h^{-1}\left\{X_v\widehat{\beta}_t^s + Z_v^s\widehat{\gamma}_t^s\right\} \geq 0.5\right) \text{ and} \quad (10)$$

$$\widehat{y}_v^d = I\left(h^{-1}\left\{X_v\widehat{\beta}_t^d + Z_v^d\widehat{\gamma}_t^d\right\} \geq 0.5\right), \quad (11)$$

respectively.

Denoting $\widehat{y}_v^s = (\widehat{y}_{v1}^s\ \widehat{y}_{v2}^s\ \ldots\ \widehat{y}_{vn_v}^s)'$ and $\widehat{y}_v^d = (\widehat{y}_{v1}^d\ \widehat{y}_{v2}^d\ \ldots\ \widehat{y}_{vn_v}^d)'$, we measure the agreement in prediction between the two segmentation methods as

$$\kappa = \frac{1}{n_v}\sum_{j=1}^{n_v} I(\widehat{y}_{vj}^s = \widehat{y}_{vj}^d). \quad (12)$$

A high value of κ indicates that, given a classification model, the prediction in the validation using CNA profiles based on smooth segmentation is in agreement with those based on the CBS segmentation.

2.7 Software

An R package called *CNALR* is available from the corresponding author's webpage (see 'Availability' in the abstract section), including some tools to combine the results of normalisation from *CNAnorm* package (Gusnanto *et al.* , 2012). The *CNAnorm* package normalises each sequence individually, and the *CNALR* package will select relevant quantities from the *CNAnorm* output. The *CNALR* package contains functions to handle the CNA profiles across patients and fit a logistic regression.

3 RESULTS

In this section, we mainly present the results for the lung cancer dataset. For the oral cancer dataset, we present its results mainly in the supplementary material.

3.1 CNA profiles

In this section, we briefly describe the summary of the profiles within each histology. Figure 2 shows (point-wise) means of the CNA profiles across the patients within each of the squamous and adenocarcinoma groups. One point in Figure 2 corresponds to a genomic window, and is the mean of CNA at the same genomic window (as illustrated in Figure 1) across different patients in each histology group. The figure indicates that the means of CNA profiles between the two groups share many similarities, although some differences can be observed. For example, we observe that the overall gains in $3q$ and $2p$ regions are higher in the squamous group than those in the adenocarcinoma group.

Although the pattern in e.g. $3q$ region indicates a separation between the squamous and adenocarcinoma groups, this does not mean that a modelling is not needed. Gains in $2p$ and $3q$ are not exclusive to the squamous group, just more common. If we rely the classification of tumour histology only on the pattern that we observe in the $3q$ region, then effectively we ignore the contribution of other genomic regions in the genome, as well as risk misclassifying any sample which happens by chance to have a small number of CNAs in the "wrong" places. Our results below suggest that the histology prediction of new samples does not necessarily improve if the prediction relies only on CNA profiles from some genomic windows.

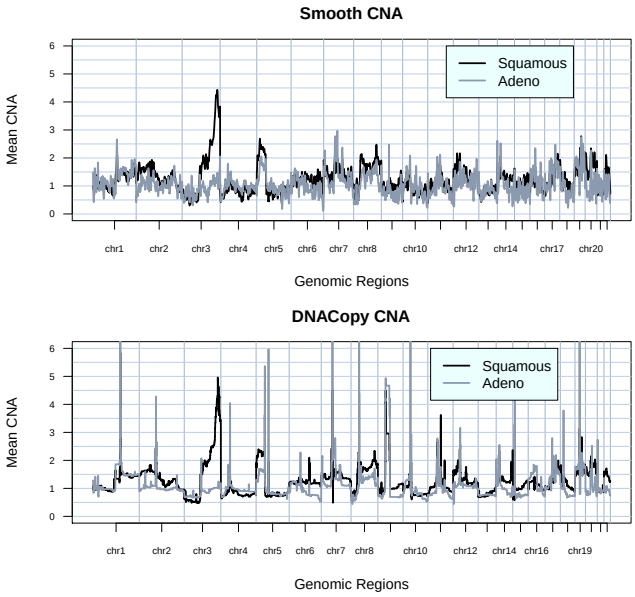

Fig. 2. *Mean of CNA profiles within the squamous group and adenocarcinoma group (lung cancer dataset), based on smooth segmentation and CBS (DNACopy) segmentation.*

3.2 Model fit

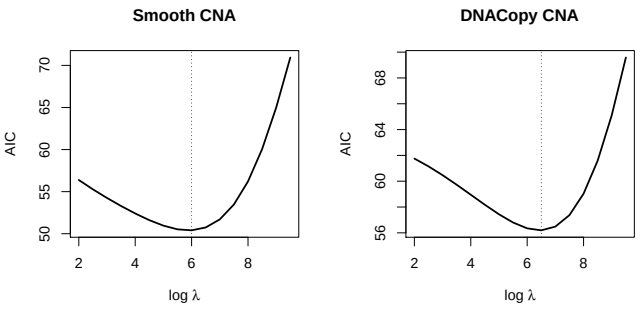

Fig. 3. *Estimation of λ in the model by minimising AIC for CNA profiles in lung cancer dataset.*

We fit the logistic regression on the covariates and CNA profiles of the patients. An important parameter to be estimated from the model is λ, and this is described in Figure 3. The figure indicates that

Summary of fixed predictors

Predictor	Estimate	Std. Error	z value	p-value
(without CNA profiles)				
Age	0.02279	0.02801	0.814	0.416
Gender	-0.65864	0.46839	-1.406	0.160
(with smooth CNA profiles)				
Age	0.02702	0.08789	0.3075	0.7585
Gender	-0.83923	1.32923	-0.6313	0.5278
(with DNACopy CNA profiles)				
Age	0.02840	0.08444	0.3363	0.7366
Gender	-0.73667	1.32644	-0.5554	0.5783

Table 1. *Summary of the fixed predictors (age and gender) in the logistic regression without the inclusion of the CNA profiles (top table), after the inclusion of smooth CNA profiles (midlle table), and DNACopy CNA profiles (bottom table).*

the optimal λ in the model is estimated as $\exp(6)$ and $\exp(6.5)$ for CNA profiles using smooth segmentation and CBS segmentation, respectively.

Using the optimal $\widehat{\lambda}$, the results of estimation in the (fixed) covariates are presented in Table 1. The table indicates that none of the (fixed) covariates is statistically significant (i.e. p-value> 0.05). These results suggest that there is no significance difference in patients' age and gender distribution between the squamous and adenocarcinoma groups.

The random effects estimates of the logistic regression are presented in Figure 4. An immediate pattern that we could see is the magnitude of the estimates for the genomic windows in the *3q* region. This pattern suggests that CNA gains in the region contribute to increase the probability of the tumour to be classified into the squamous group, and CNA losses to the adenocarcinoma group. On the other hand, negative estimates as we see in most of chromosome 4 indicate that CNA gains in the region contribute to increase the probability of the tumour being classified as adenocarcinoma group and CNA losses to the squamous group.

The model fit for our analysis is presented in Figure 5. The figure indicates that the logistic regression can fit the data very well, where the tumour histology is correctly classified.

In Figure 4, we need to be careful in interpreting the pattern of the random effects estimates. None of the individual random effects estimates is statistically significant, in the sense that all of the 95% individual confidence interval for the random effects include zero. This does not mean that there is no information contained in the data. This is just a result of the estimation of more than 17,000 parameters from just 76 observations. This is also not a contradiction when we consider that the model has a good fit as shown in Figure 5. From the construction of the model, although the individual random effects estimates are not significant, there is a linear combination of window-wise CNA profiles that jointly classify the tumour histology.

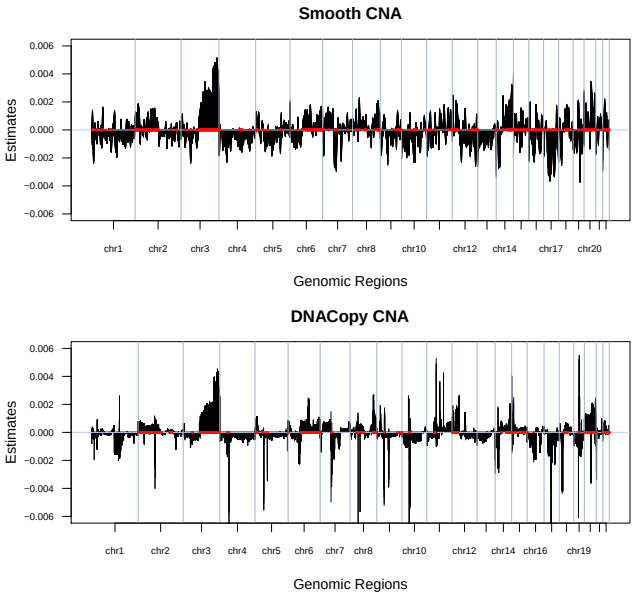

Fig. 4. *Estimates of the random effects in the full model, using CNA profiles from smooth and CBS (DNACopy) segmentation in the lung cancer dataset. Genomic windows with missing values (such as in the centremere regions) are removed from the figure. We currently do not include sex and mitochondrial chromosomes in the analysis. The red dots on the horizontal axis indicate 4549 (smooth) and 3633 (DNACopy) genomic windows (25.9% and 19.8%) that have significant differences of CNA profiles between the squamous and adenocarcinoma groups, using a permutation method. A more detailed view of the random effects estimates in each chromosome is presented in the supplementary material.*

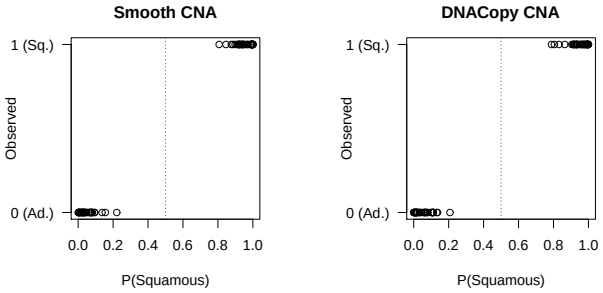

Fig. 5. *The observed tumour histology against its model fit based on the full logistic regression model using smooth and DNACopy CNA profiles in the lung cancer dataset. The vertical lines mark the 50% probability to be in the Squamous histological group. Probability more than 50% are normally classified to the Squamous group.*

3.3 Cross validation

The results of cross validation for the lung cancer dataset are presented in Figure 6. The figure indicates that the logistic regression has the lowest median classification error in comparison to the other classification models that we consider in our study. The use of KNN and discriminant analysis give a relatively high classification error. This high error is consistent regardless of the number of neighbour in KNN or the type of discriminant analysis.

The partial least squares (PLS) gives relatively low median of classification error, which are comparable to that of the logistic regression. In building the classification rule, PLS expands the space on the CNA mean differences (see for example Barker and Rayens

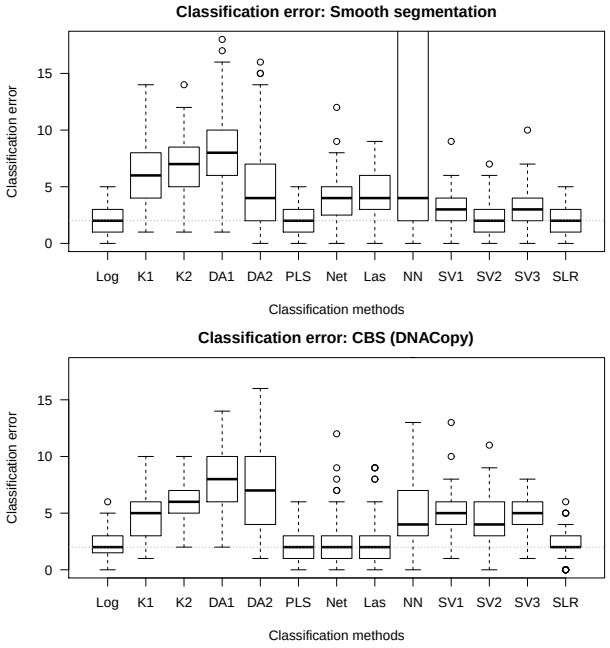

Fig. 6. *Classification error from 100 cross validations (38 observations in each of the training set and validation set) using only the smooth- and CBS-segmented CNA profiles in the lung cancer dataset across different classification models: logistic regression ("Log"), k-nearest neighbour with $k = 1, 2$ ("K1"-"K2"), diagonal quadratic ("DA1") and linear ("DA2") discriminant analysis, partial least squares ("PLS"), elastic net ("Net"), lasso ("Las"), neural network ("NN"), support vector machines with C-classification ("SV1"), nu classification ("SV2"), and bound-constraint classification ("SV3"), and smoothed logistic regression ("SLR"). The horizontal dotted line is the median of classification error of the logistic regression. Figures with "K3"-"K7" are presented in the supplementary material. The significance of the classification error between pair of methods are also presented in the supplementary material.*

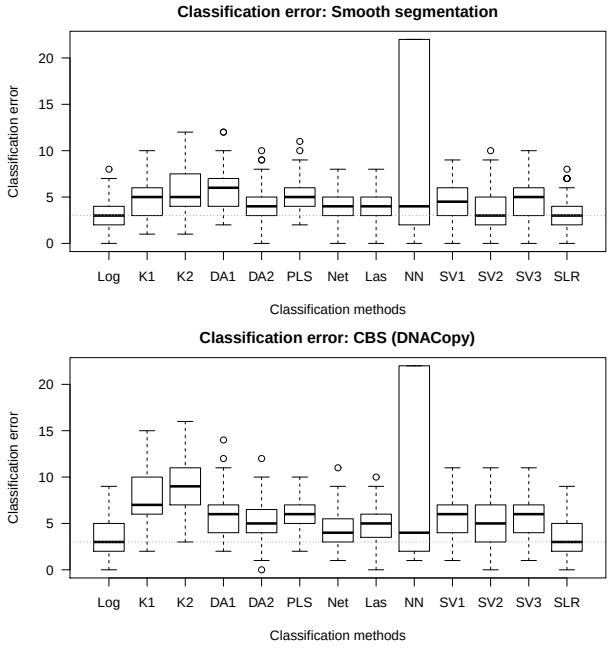

Fig. 7. *Classification error from 100 cross validations (52 and 50 observations in the training set and validation set, respectively) in the oral cancer dataset using only the smooth- and CBS-segmented CNA profiles across different classification models: logistic regression ("Log"), k-nearest neighbour with $k = 1, 2$ ("K1"-"K2"), diagonal quadratic ("DA1") and linear ("DA2") discriminant analysis, partial least squares ("PLS"), elastic net ("Net"), lasso ("Las"), neural network ("NN"), support vector machines with C-classification ("SV1"), nu classification ("SV2"), and bound-constraint classification ("SV3"), and smoothed logistic regression ("SLR"). The horizontal dotted line is the median of classification error of the logistic regression. Figures with "K3"-"K7" are presented in the supplementary material. The significance of the classification error between pair of methods are also presented in the supplementary material.*

(2003)), without being weighted by the between-class covariance matrix. Given that CNA profiles can exhibit dramatic changes of copy number, PLS ignores the between-class covariance-matrix (i.e. focus on the differences on copy number) to give a good prediction.

The elastic-net and lasso models give a slightly higher median classification error than that of the logistic regression when we use smooth-segmented CNA profiles, but not when we use CBS-segmented CNA profiles. This is an interesting result because the elastic-net and lasso models produce sparse solution on the parameters where some of the estimates are zero estimated. Effectively, a variable selection is embedded inside the classification models. Our results indicate that when we use smooth-segmented CNA profiles as predictors, the variable selection does not necessarily produces higher prediction (or lower error). On the CBS-segmented CNA profiles, the variable selection still gives the same low median of classification error as that of logistic regression (Section 2.3).

For the oral cancer data, we have a more challenging situation where it is sometimes more difficult to distinguish oral squamous cell carcinoma (OSCC) from oral verrucous carcinoma (OVC) histologically. Figure 7 describes the prediction error in the oral cancer data. The figure indicates that the logistic regression and SLR have the lowest prediction error in both smooth and CBS (DNAcopy) segmented CNA profiles. Support vector machines with nu-classification has a low prediction error comparable to the logistic regression, only when using the smooth-segmented CNA profiles.

Still within the cross validation, Figure 8 presents how the two segmentation methods (smooth and CBS segmentation) agree in prediction within the validation set. If the agreement is low, then the choice of segmentation method does matter – in terms of classification error – when we use a particular classification method. The contrary can be said if the agreement is high. Figure 8 indicates that we have a high agreement when we use logistic regression and support vector machines. PLS, elastic net, and lasso models have lower agreement in the lung cancer data compared to the logistic regression, and this agreement increases in the oral cancer data to be comparable to the logistic regression.

The overall results of our cross validation suggest that the logistic regression is able to achieve a low classification error while maintaining high agreement in prediction between the smooth- and CBS-segmented CNA profiles.

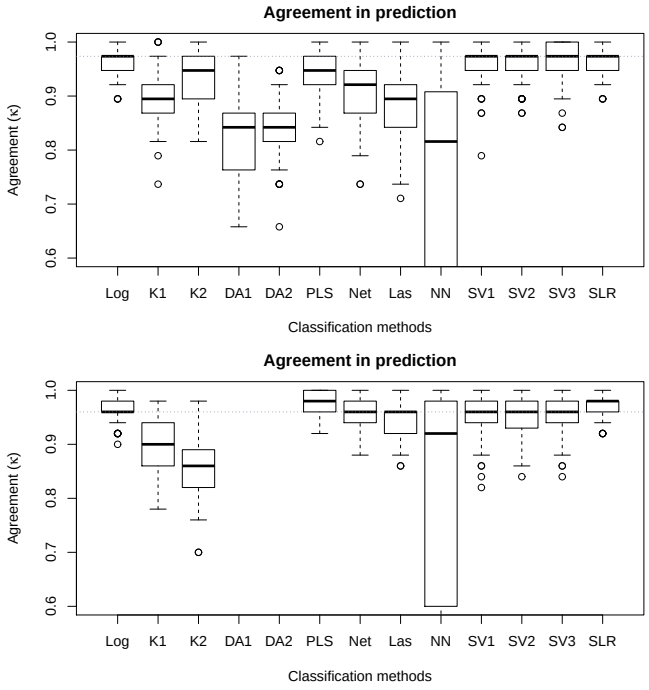

Fig. 8. *Agreement between smooth segmentation and CBS segmentation in prediction from 100 cross validations in the lung cancer dataset (top panel) and oral cancer dataset (bottom panel) across different classification models: logistic regression ("Log"), k-nearest neighbour with $k = 1, 2$ ("K1"-"K2"), diagonal quadratic ("DA1") and linear ("DA2") discriminant analysis, partial least squares ("PLS"), elastic net ("Net"), lasso ("Las"), neural network ("NN"), support vector machines with C-classification ("SV1"), nu classification ("SV2"), and bound-constraint classification ("SV3"), and smoothed logistic regression ("SLR"). For the oral cancer dataset (bottom panel) the agreement for DA1 and DA2 are below 0.6. The horizontal dotted line is the median of agreement in the logistic regression. Figures with "K3"-"K7" are presented in the supplementary material.*

4 DISCUSSION

In predicting tumour histology, the main challenge is to consider a model with low classification error. Further than that, we are also interested in the agreement between smooth- and CBS-segmented CNA profiles in the prediction of new tumour sample. This is critical when we consider that the CNA profiles are derived from low-coverage next-generation sequence data that have undergone several preparation steps. This includes, but is not limited to, mapping to the reference genome, filtering, optimal window estimation, normalisation (including normalisation due to contamination), and segmentation. To have a classification method with a good prediction while having a minimum dependency on a previous preprocessing step is a great advantage.

The logistic regression described in Section 2.3 (also known as *Tikhonov regularisation*) is well known to tend to group variables (genomic window) together (Zou and Hastie , 2005). This property of the logistic regression is likely to predominate in prediction, in terms of classification error and agreement between the segmentation methods, because there are some dependencies of CNA between neighbouring genomic windows. The lasso and elastic net models achieve comparable classification error as the logistic regression only when the CNA profiles are CBS-segmented. The CBS (DNACopy) segmentation generally outputs long segments, and the variable selection effect that the two methods produce does not impair the prediction because the genomic regions that survive the penalisation are able to represent the information of the whole segment.

We have demonstrated the ability of the proposed method to distinguish between two subtypes of lung cancer. We then validated this method in a separate cohort of oral cancer patients, whose diagnosis is not always straightforward, but whose prognosis is very different. However, in theory, it could be used to distinguish between any subtypes of tumour, or to make predictions about disease progression, drug resistance, or outcome.

Currently, several molecular classifiers are used to distinguish cancer subtypes, among them mRNA or protein expression. These are both selective methods. The study of mRNA usually involved the removal of non-coding RNA and micro RNA, the roles of which are increasingly becoming apparent. The study of protein expression usually involves only a few known proteins. Both of these methods will also involve further depletion of valuable sample material. Where there is no known mRNA or protein signature, our method may prove useful in finding previously unsuspected differences in the genomes of two sample groups. Low coverage sequencing is also possible with extremely low quantities of badly degraded DNA (Wood *et al.* , 2010). If mRNA or protein tests are available, then their results can be easily added to the logistic regression method as additional predictors and will further improve performance.

5 CONCLUSION

We have investigated the use of logistic regression to model tumour histology and include the genome-wide CNA profiles as predictors. The model enables us to include clinical characteristics as fixed covariates and CNA profiles as random predictors in a single modelling framework. The model exhibits a good fit and, in a cross-validation, shows minimal classification error. The model also demonstrates the best agreement in prediction between CNA profiles produced by two very different segmentation methods.

ACKNOWLEDGEMENT

Funding: This work was supported by the Yorkshire Cancer Research [L341PG] and Betty Woolsey Endowment. MS is supported by the Saudi Arabian Ministry of Higher Education, King Fahd Medical City, Saudi Arabia.

REFERENCES

Barker, M. and Rayens, W. (2003), Partial least squares for discrimination. *Journal of Chemometrics*, **17**: 166–173

Eilers P.H.C. *et al.* (2001) Classification of microarray data with penalized logistic regression, *International Biomedical Optics Symposium*, San Jose 20-26 January 2001.

Gazdar, A.F. (2010) Should we continue to use the term non-small-cell lung cancer?, *Annals of Oncology* **21** (suppl 7): vii225-vii229, doi: 10.1093/annonc/mdq372

Gusnanto *et al.* (2012) Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, **28**:40-47

Gusnanto *et al.* (2014), Estimating optimal window size for analysis of low-coverage next-generation sequence data, *Bioinformatics* (Advance access) doi: 10.1093/bioinformatics/btu123

Huang J., *et al.* (2007) Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics*, **23**:2463-2469

Huang J, Salim A, Lei K, O'Sullivan K, Pawitan Y. (2009). Classification of array CGH data using smoothed logistic regression, *Statistics in Medicine*, **28**: 3798-3810

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760

Mermel, C.H. *et al.* (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers, *Genome Biology* **12**:R41

Olshen A.B., *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**:557-572

Pawitan, Y. (2001) *In All Likelihood: Statistical Modelling and Inference using Likelihood*, New York: Oxford University Press

Rekha, K.P. and P.V. Angadi (2010) Verrucous carcinoma of the oral cavity: a clinico-pathologic appraisal of 133 cases in Indians, *Oral Maxillofac Surg*, **14**(4): 211-218

de Ronde J.J. *et al.* (2010) KC-SMARTR: An R package for detection of statistically significant aberrations in multi-experiment aCGH data, *BMC Research Notes*, **3**:298

Wood H., *et al.* (2010) Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res.*, **38**:e151

Zou, H., and Hastie, T. (2005) Regularization and variable selection via the elastic net, *J Roy Stat Soc B*, **67**:301-320