SegCorr a statistical procedure for the detection of genomic regions of correlated expression

Eleni Ioanna Delatola, Emilie Lebarbier, Tristan Mary-Huard, François Radvanyi, Stéphane Robin, Jennifer Wong

Delatola *et al. BMC Bioinformatics* (2017) 18:333
DOI 10.1186/s12859-017-1742-5

BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

SegCorr a statistical procedure for the detection of genomic regions of correlated expression

Eleni Ioanna Delatola[1,2,3,4*], Emilie Lebarbier[1,2], Tristan Mary-Huard[1,2,5], François Radvanyi[3,4], Stéphane Robin[1,2] and Jennifer Wong[3,4,6,7]

Abstract

Background: Detecting local correlations in expression between neighboring genes along the genome has proved to be an effective strategy to identify possible causes of transcriptional deregulation in cancer. It has been successfully used to illustrate the role of mechanisms such as copy number variation (CNV) or epigenetic alterations as factors that may significantly alter expression in large chromosomal regions (gene silencing or gene activation).

Results: The identification of correlated regions requires segmenting the gene expression correlation matrix into regions of homogeneously correlated genes and assessing whether the observed local correlation is significantly higher than the background chromosomal correlation. A unified statistical framework is proposed to achieve these two tasks, where optimal segmentation is efficiently performed using dynamic programming algorithm, and detection of highly correlated regions is then achieved using an exact test procedure. We also propose a simple and efficient procedure to correct the expression signal for mechanisms already known to impact expression correlation. The performance and robustness of the proposed procedure, called SegCorr, are evaluated on simulated data. The procedure is illustrated on cancer data, where the signal is corrected for correlations caused by copy number variation. It permitted the detection of regions with high correlations linked to epigenetic marks like DNA methylation.

Conclusions: SegCorr is a novel method that performs correlation matrix segmentation and applies a test procedure in order to detect highly correlated regions in gene expression.

Keywords: Gene expression, Chromosomes, Correlation matrix segmentation, CNV, DNA Methylation, SegCorr

Background

In the last decade, the study of local co-expression of neighboring genes along the chromosome has become a question of major importance in cancer biology [6]. The development of "Omics" technologies have permitted the identification of several mechanisms inducing local gene regulation, that may be due to a common transcription factor [11] or common epigenetic marks [14, 34]. Copy number variation due to polymorphism or to genomic instability in cancer is also a possible cause for observing a correlation between neighboring genes [1], as their expressions are likely to be affected by the same copy number variation (CNV). It has further been observed that local regulations may occur in specific nuclear domains, as the nuclear region is an environment which may favor or not transcription [4].

Investigating the impact of a specific source of regulation (TF, CNV, epigenetic modifications such as DNA methylation and histone modifications) on the expression has now become a common practice for which statistical tools are readily available. However, only a few methods have been proposed to focus on the direct analysis of gene expression correlation along the chromosomes. The direct analysis of correlations may have different purposes:

(i) one can aim at detecting all potential chromosomal domains of co-expression, then investigating to which extend known causal mechanisms are responsible for the observed co-expression patterns,

*Correspondence: eldelatola@yahoo.gr
[1] AgroParisTech UMR518, 75005 Paris, France
[2] INRA UMR518, 75005 Paris, France
Full list of author information is available at the end of the article

(ii) one can aim at detecting chromosomal domains of co-expression where correlations are not caused by already known sources of regulation, in order to identify new potential mechanisms impacting transcription.

Addressing problems (*i*) and (*ii*) is crucial to fully understand transcriptional deregulation and/or to model gene regulation. We first consider problem (*i*) and provide a precise definition of our purpose: one aims at identifying correlated regions, i.e. blocks of neighboring genes, the expression of which displays correlations across patient samples that are significantly higher than expected. Indeed, it has been observed that background correlation between adjacent genes along the genome does exist. This background correlation should not be confounded with the co-expression that can be locally observed due to the aforementioned mechanisms. Consequently, we do not consider here methods that only account for this background correlation in the statistical modeling (for instance to improve the detection of differentially expressed genes), such as [24], [40] or [30]. Also note that we focus on methods that detect correlated regions on the basis of expression data solely. This excludes strategies that look for clusters of adjacent genes based on correlations between gene expression and a given phenotype or response, such as Rendersome [24], DIGMAP [41] or REEF [10].

Several approaches have been proposed to tackle problem (*i*). CluGene [13] uses a clustering method accounting for the chromosomal organization of the genes, while G-NEST [20] and TCM [28] rely on sliding windows procedures. The principle of the latter approach is to compute correlation scores for genes falling within the window, then to detect local peaks of high correlation scores. While these procedures have been successfully applied to cancer data, all tackle the detection of correlated region using heuristics. As such, they suffer from classical limitations associated with these techniques, including local optimum (for clustering algorithms) or detection instability according to the choice of the window size (for sliding windows).

It is now well known that the problem of finding regions in a spatially ordered signal can be cast as a segmentation problem, for which standard statistical models exist, along with efficient algorithms to find the globally optimal solution [3]. According to our definition, the detection of correlated regions boils down to the block-diagonal segmentation of the correlation matrix between gene expressions. Such an approach has been proposed in image processing [22], finance [18] and bioinformatics for CNV analysis [42], but to the best of our knowledge it has never been considered for the detection of correlated expression regions.

While problem (*i*) can be addressed on the basis of only expression data, problem (*ii*) requires the additional measurement of the signal one needs to account for. For example, consider that one seeks for locally expressed co-regulation events that are not due to copy number variations but due to other causes such as epigenetic mechanisms. The strategy we adopt here consists in first correcting the expression data for potential cancer CNV contribution, then in applying the procedure described to solve problem (*i*) on the corrected signal. The corrected signal is obtained by regressing the initial expression signal on the CNV signal. Although quite simple, the strategy turns out to be efficient in practice. An alternative strategy would be to jointly model both the expression and the signals to correct for, and then propose within this framework a correction. Such a strategy would necessitate to adapt the modeling to the specific combination of signals one has at hand. In comparison, the regression procedure proposed here can be applied to any kind and any number of signals one needs to correct for.

The outline of the present article is the following. In Section 'Correlation matrix segmentation' (Methods) we propose a parametric statistical framework for the problem of correlated region identification. Finding regions of co-regulated genes can then be achieved by maximum likelihood inference (to find the boundaries of each region along with their correlation levels). Moreover, we propose a procedure to correct for known sources of correlation. An exact test procedure to assess the significance of the correlation with respect to background correlation is proposed in Section 'Assessing correlation significance' (Methods). We introduce a simple procedure to correct expression data beforehand for some known (and quantified) sources of correlation. Because the background correlation level is a priori unknown, an estimator of this quantity is also proposed. The performance of the resulting procedure, called SegCorr hereafter, is illustrated in Section 'Simulation study' (Results) on simulated data, along with a comparison with the TCM algorithm proposed in [28]. Finally, a case study on cancer data is presented in Section 'Bladder cancer data' (Results), in which we identify some regions with high correlation between gene expression and the local DNA methylation level.

Methods

Correlation matrix segmentation

Statistical model

We consider the following expression matrix:

$$Y = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ Y_{21} & \cdots & Y_{2p} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix}$$

where Y_{ij} stands for the expression of gene j ($j = 1, \ldots, p$) observed in patient i ($i = 1, \ldots, n$). The i-th row of this matrix is denoted Y_i and corresponds to the expression vector of all genes in patient i. In order to detect regions of correlated expression, we consider the following statistical model. Profiles $\{Y_i\}_{1 \leq i \leq n}$ are supposed to be i.i.d, normalized (centered and standardized), following a Gaussian distribution with block-diagonal correlation matrix G:

$$G = \begin{bmatrix} \Sigma_1 & & \\ & \Sigma_k & \\ & & \Sigma_K \end{bmatrix} \quad \text{with} \quad \Sigma_k = \begin{bmatrix} 1 & \cdots & \rho_k \\ \vdots & \ddots & \vdots \\ \rho_k & \cdots & 1 \end{bmatrix}. \tag{1}$$

The model states that genes are spread into K contiguous regions, with respective lengths p_k ($k = 1, \ldots, K$, $\sum_{1 \leq k \leq K} p_k = p$), the length of a region being the number of genes it contains. Genes belonging to different regions are supposed to be independent, whereas genes belonging to a same region are supposed to share the same pairwise correlation coefficient ρ_k. This amounts to assume that some specific effect (e.g. methylation) affects the expression of all genes belonging to the region. More specifically, let U_k denote the vector of the region effect (accross patients). For all genes j from region k, the model can be written as $Y_{ij} = U_{ik} + E_{ij}$. The error terms E_{ij} are all independent and independent from U_{ik} such that $\mathbb{V}(U_{ik})/\mathbb{V}(Y_{ij}) = \rho_k$, where $\mathbb{V}(U)$ stands for the variance of U.

While different technologies (microarrays, RNA-seq) may provide different types of signal (continuous, counts), an appropriate transformation may be applied to make the Gaussian assumption reasonable. For example, in the context of segmentation, [7] showed that Gaussian segmentation applied to $\log(1 + x)$-transformed RNA-seq data performs as well as negative binomial segmentation applied to the raw data.

Accounting for known sources of regulation

As mentioned in the Introduction, a second task (*ii*) can be to detect correlated regions which are not due to an already known mechanism. To this aim, one may first correct the expression signal using the following regression model :

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \tag{2}$$

where x_{ij} stands for the covariate observed in patient i for gene j. For instance, in the illustration of Section 'CNV-dependent regions', x_{ij} is the copy number associated to patient i at location of gene j. The corrected signal is then $\widetilde{Y}_{ij} = Y_{ij} - \widehat{\beta_0} - \widehat{\beta_1} x_{ij}$. Note that $\widehat{\beta_0}$ and $\widehat{\beta_1}$ can be obtained as ordinary least-square estimates. Indeed, it suffices to assume that (ϵ_{ij}) are independent among patients (but not among genes) to get the standard linear regression estimates (see [2], Chapter 8). Once the correction has been made, the model described in Section 'Statistical model' can be applied to the corrected signal $\widetilde{Y}_{ij}$.

Note that the correction procedure could be based on more sophisticated modellings of the relationship between gene expression and mechanisms such as CNV or methylation, e.g. the ones proposed in [19, 23, 38]. The difference between the observation and the prediction obtained from one such model (i.e. the residuals) could then be used as the corrected signal.

Lastly, the proposed correction procedure can be adapted straightforwardly to handle count data such as provided by RNAseq technologies. Indeed, Model (2) can be rephrased in the generalized linear model framework and Pearson residuals can be used as $\widetilde{Y}_{ij}$ (see e.g. [12] for a general introduction or [15] for the specific case of negative binomial regression).

Inference of correlated regions

Parameter inference in Model (1) amounts to estimating the number of regions K, the region boundaries $0 = \tau_0 < \tau_1 < \cdots < \tau_K = p$, and the correlation parameters $\rho_1, \ldots, \rho_K$ within each of these regions. Here, we consider a maximum penalized likelihood approach. First, we show that for a given K the optimal region boundaries and correlation coefficients can be efficiently obtained using dynamic programming. The number of regions can then be selected using a penalized likelihood criterion. For a fixed K, the estimation problem can be formulated as follows:

$$\arg \max_{\tau_1 < \cdots < \tau_{K-1}} \max_{\rho_1, \ldots, \rho_K} \mathcal{L} \tag{3}$$

where the log-likelihood $\mathcal{L}$ is $-\left(n \log |G| + \mathrm{tr}\left[YG^{-1} (Y)^\top\right]\right)/2$. Here, thanks to the block diagonal structure of the correlation matrix in Model (1), the log-likelihood can be rewritten as

$$\begin{aligned} -2\mathcal{L} &= \sum_k \left\{ n \log |\Sigma_k| + \mathrm{tr}\left[Y^{(k)} \Sigma_k^{-1} (Y^{(k)})^\top \right] \right\} \\ &= -2 \sum_k \mathcal{L}(\tau_{k-1} + 1, \tau_k) \quad = -2 \sum_k \mathcal{L}_k \end{aligned} \tag{4}$$

where $Y^{(k)}$ stands for the set of expression from Y corresponding to genes included in the k-th region, and $\mathcal{L}_k = \mathcal{L}(\tau_{k-1} + 1, \tau_k)$ is the log-likelihood corresponding to region k, i.e. corresponding to measurements of genes from $\tau_{k-1} + 1$ to τ_k. While log-likelihood (4) is derived in a Gaussian setting, it can be used for count data, as the Pearson residuals mentioned in Section 'Accounting for known sources of regulation' have an approximate Gaussian distribution.

Thanks to the additivity of the likelihood over the regions, the optimization problem (3) boils down to

$$\arg \max_{\tau_1<\cdots<\tau_{K-1}} \sum_k \max_{\rho_k} \mathcal{L}_k. \qquad (5)$$

Inference when K is fixed We first show that for a given region k with known boundaries, explicit expressions can be obtained for both the ML estimator $\widehat{\rho}_k$ and the likelihood $\mathcal{L}_k$ at the optimum:

Lemma 1 *For a region k with fixed boundaries* $[\tau_{k-1}+1, \tau_k]$, *the maximum of* $\mathcal{L}_k$ *with respect to* ρ_k *is reached for*

$$\widehat{\rho}_k = \frac{\sum_{j=\tau_{k-1}+1}^{\tau_k} \sum_{\ell=\tau_{k-1}+1}^{\tau_k} \widehat{G}_{j\ell} - p_k}{p_k^2 - p_k}$$

where $\widehat{G}_{j\ell} := n^{-1} \sum_{i=1}^{n} Y_{ij} Y_{i\ell}$. *Furthermore, the maximal value of* $\mathcal{L}_k$ *is given by:*

$$-2\widehat{\mathcal{L}}_k = n\left[p_k + (p_k-1)\log\left(1-\widehat{\rho}_k\right) + \log\left(1+(p_k-1)\widehat{\rho}_k\right)\right].$$

The proof is given in Additional file 1. The expression of Problem (5) is now

$$\arg \max_{\tau_1<\cdots<\tau_{K-1}} \sum_k \widehat{\mathcal{L}}_k$$

which is additive with respect to the $\widehat{\mathcal{L}}_k$ terms that can be straightforwardly computed thanks to Lemma 1. Consequently, optimization can be performed via Dynamic Programming (DP, [17], [25]). The optimal boundaries, and correlation estimators can be obtained at computational cost $\mathcal{O}(Kp^2)$.

Lasso-type approaches have been proposed to tackle segmentation problems in a faster way (see e.g. [36]). First, note that such methods rely on a relaxation of the original problem, so that the result may be different from the exact solution of problem (3). Furthermore, in the context of matrix segmentation, such approaches have been proposed ([5, 21]), which do not allow to capture the longitudinal structure (i.e. blocks of neighboring genes).

Model selection To choose the number of regions, we adopt the model selection strategy proposed in [17]. For each $1 \leq K \leq K_{\max}$, we define the maximal log-likelihood for K regions as

$$L_K = \max_{\tau_1<\cdots<\tau_{K-1}} \sum_k \widehat{\mathcal{L}}(\tau_{k-1}+1, \tau_k).$$

Furthermore, the normalized log-likelihood is defined as

$$\widetilde{L}_K = \frac{L_{K_{\max}} - L_K}{L_{K_{\max}} - L_1} (\widetilde{K}_{\max} - \widetilde{K}_1) + 1,$$

where $\widetilde{K}_j = 5 \times j + 2 \times j \log(p/j)$ is the penalty function. [17] suggests to estimate the number of regions $\widehat{K}$ as the value of K such that $\widetilde{L}_K$ displays the largest slope change. Namely, we take

$$\widehat{K} = \arg\min_K \left\{ (\widetilde{L}_K - \widetilde{L}_{K+1}) - (\widetilde{L}_{K+1} - \widetilde{L}_{K+2}) > S \right\}, \qquad (6)$$

where the value of threshold S is predefined. Throughout the paper, we used $S = 0.7$ as suggested in [17]. The robustness of the results with respect to other values for threshold S is investigated in Section 'Simulation study'. This global approach (dynamic programming and model selection) has been applied with success for CNV detection (see [25] and [16] for a comparative study).

Assessing correlation significance

It has been observed [9, 28, 32, 34] that background correlations may exist between adjacent genes along the genome, i.e. one expects the correlation level in any region to be positive. As a consequence, one has to check whether a given region exhibits a correlation level that is significantly higher than the background correlation level ρ_0, that is observed by default.

Test procedure Once the correlation matrix segmentation is performed, it is possible to identify regions with high correlation levels by testing $H_0 : \rho_k = \rho_0$ vs $H_1 : \rho_k > \rho_0$. This can be done using the following test statistic for region k:

$$T_k = \sum_i^n \left(Y_{i\bullet}^{(k)} - Y_{\bullet\bullet}^{(k)} \right)^2$$

where $Y_{i\bullet}^{(k)} = p_k^{-1} \sum_{j=\tau_{k-1}+1}^{\tau_k} Y_{ij}$ and $Y_{\bullet\bullet}^{(k)} = n^{-1} \sum_{i=1}^{n} Y_{i\bullet}^{(k)}$. Assuming Model (1) is true, test statistic T_k has distribution

$$T_k \sim \lambda(p_k, \rho_k)\chi^2_{n-1} \text{ where } \lambda(p_k, \rho_k) = \frac{(1+(p_k-1)\rho_k)}{p_k}.$$

Here χ^2_{n-1} stands for the chi-square distribution with $n-1$ degrees of freedom. The proof is given in Additional file 1. We emphasize that this test is exact and does not rely on any resampling strategy.

Consequently, the p-value associated to region k is given by

$$\mathbb{P}\left(\lambda(p_k, \rho_0) Z > T_k^{obs}\right), \text{where } Z \sim \chi^2_{n-1}.$$

Statistical power We now study the ability of the proposed test to detect a region with width p_0 where the correlation ρ is higher than in the background. The probability to detect such a region depends on both p_0 and ρ and is given by

$$\begin{aligned} Po(n,p_0,\rho) &= \Pr\{T > \lambda(p_0,\rho_0)q_{n-1,1-\alpha}\} \\ &= \Pr\left\{Z > \frac{\lambda(p_0,\rho_0)}{\lambda(p_0,\rho)}q_{n-1,1-\alpha}\right\} \end{aligned}$$

where $Z \sim \chi^2_{n-1}$ and $q_{n-1,1-\alpha}$ is the $1-\alpha$ quantile for the χ^2_{n-1} distribution. Figure 1 (Top) displays the evolution of power for different values of p_0 and ρ. Here ρ_0 and n are fixed at 0.15 and 100, respectively. The nominal levels of α are 5, 0.5 and 0.05%. These levels correspond to realistic thresholds, once multiple testing corrections such as Bonferroni or FDR are performed. One can observe that even for small values of ρ, the power is high whatever the nominal level as long as the number of genes in the considered region is equal to or higher than 5. Figure 1 also shows that the procedure will probably fail to find regions of size 3, if the correlation is not 0.7 or higher (to obtain a power of 0.8). On the same graph (Bottom), one observes that a sample of size 50 is sufficient to efficiently detect regions of size 5, as long as the correlation is higher than 0.6. Larger samples will be required if one wants to efficiently detect regions with smaller correlation levels.

Background correlation estimation The test procedure requires the knowledge of parameter ρ_0 that is unknown in practice. However, it can be estimated using

$$\widehat{\rho_0} = |\underset{i>1}{\text{median}}(\text{corr}(Y^{j-1}, Y^j))| \qquad (7)$$

where Y^j stands for the vector of expression of gene j for the n patients. Under the assumption that most pairs of adjacent genes display a ρ_0 correlation, i.e. only a few number of regions with moderate sizes exhibit a high level of correlation, $\widehat{\rho_0}$ is a robust estimator of the background correlation. The behavior of estimator (7) is investigated in Section 'Simulation study'.

Results

Simulation study

In this section, we first study the quality of the proposed estimator of ρ_0. Then we study the ability of SegCorr to detect correlated regions and compare its performance with this of TCM algorithm. The robustness of the method with respect to the choice of the model selection threshold S will be investigated in Section 'Study of the model selection threshold S' on real data, since very little difference were observed on the simulated data (results

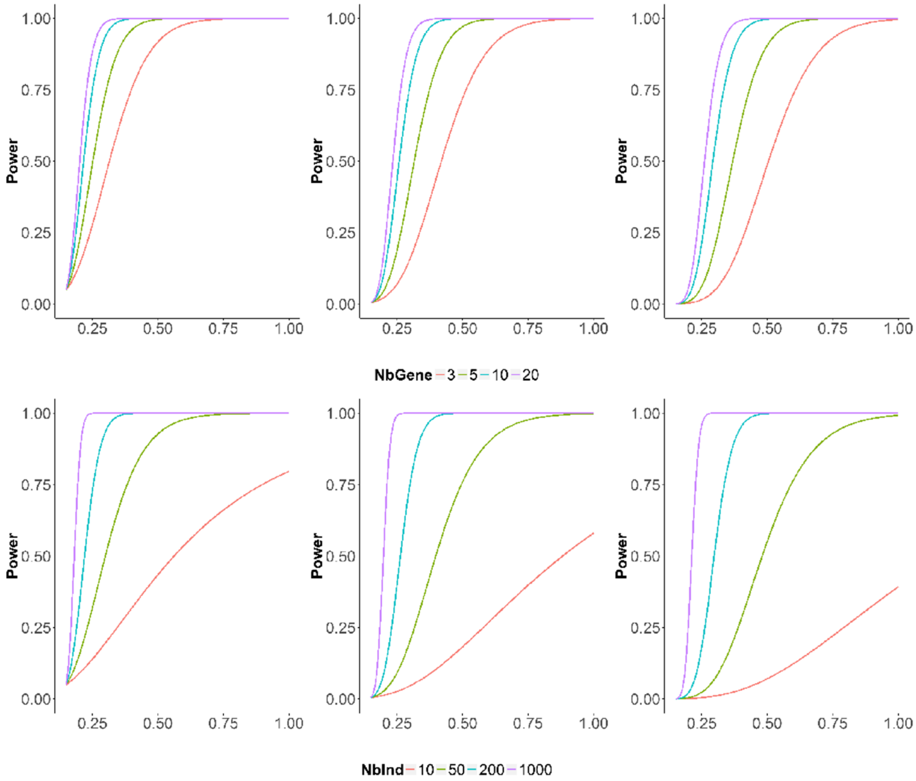

Fig. 1 Theoretical Power. *Top*: Power curves as a function of ρ, for a fixed cohort size $n = 100$ and varying region width $p_0 = 3, 5, 10, 20$. *Bottom*: Same graphs for a region of fixed width $p_0 = 5$ but varying cohort sizes $n = 10, 50, 200, 1000$. In all graphs ρ_0 is fixed at 0.15. The nominal level α of the test is set to 5% (*left*), 0.5% (*center*), 0.05% (*right*)

not shown). We also study the robustness of our procedure to a scheme where the within-region correlation is variable.

Simulation design

Scenario 1 (Easy case): the regions are defined as in [16]: each patient has one chromosome containing $p = 500$ genes and 4 regions with respective lengths $p_k = 5, 10, 20, 40$. Three values are considered for ρ_0 : $.08, .18, .28$. These values are inspired by the distribution (displayed in Fig. 2) of ρ_0 from Scenario 2. $\rho_0 = .28$ is higher than observed in [34], making the detection problem more difficult. ρ_1 varies between .3 and .9.

Scenario 2 (Realistic case – constant correlation on H_1 regions): each patient has 22 chromosomes. The length of the chromosomes, the number of regions within each chromosome and their respective sizes are the same as in the results from [34]. ρ_0 is specific to each chromosome and estimated on the same dataset. ρ_1 varies between .3 and .9.

Scenario 3 (Realistic case – variable correlation on H_1 regions): the design is the same as in Scenario 2, except that ρ_0 is fixed to .18. Furthermore, for each H_1 region covariance matrix is drawn from a p_k-variate Wishart distribution $\mathrm{W}_{p_k}(S, \nu)$ where the entries of the matrix S are one on the diagonal and $\rho_1 = .5$ elsewhere and ν is the number of degrees of freedom. Small values of ν, result in a higher variance, making the detection more difficult. Because ν has to be greater or equal to p_k, we took $\nu = p_k \times 2^{\beta}$, where $\beta = (0.5, 1, 1.5, \ldots, 5)$. So the variability decreases as β increases.

For each scenario, samples of n = 50 and 100 patients were considered and, for each combination (n, ρ_0, ρ_1) the simulation was replicated 100 and 20 times, for the first and the last two scenarios respectively.

Quality of the ρ_0 estimator

For this study, we consider Scenario 2. Figure 3 illustrates the estimation accuracy of ρ_0 under different levels of both H_0 and H_1 correlations on chromosome 5. Estimator (7) yields over-estimated values of the true background correlation level. One observes that the overestimation does not depend on the correlation level in H_1 regions, thanks to the use of the median. Still, as expected, it is linked to the proportion of pairs of adjacent genes with H_1

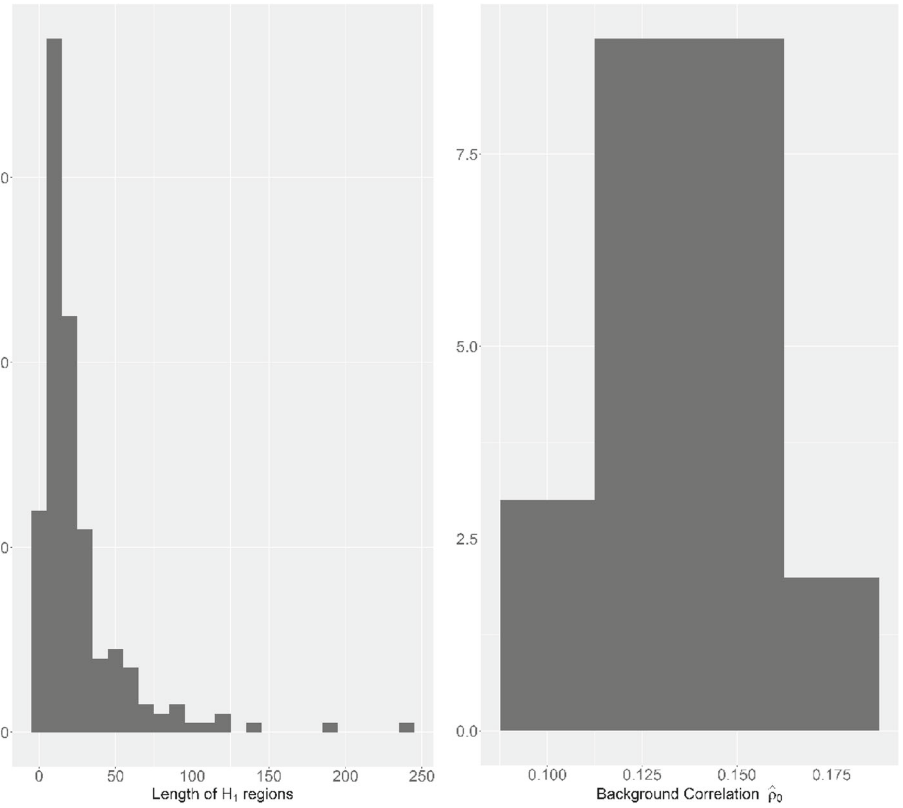

Fig. 2 Simulation Design. *Left*: Length of H_1 regions in the reference dataset. *Right*: Distribution of the background correlation $\hat{\rho}_0$ obtained from the reference data according to the segmentation obtained in [34]

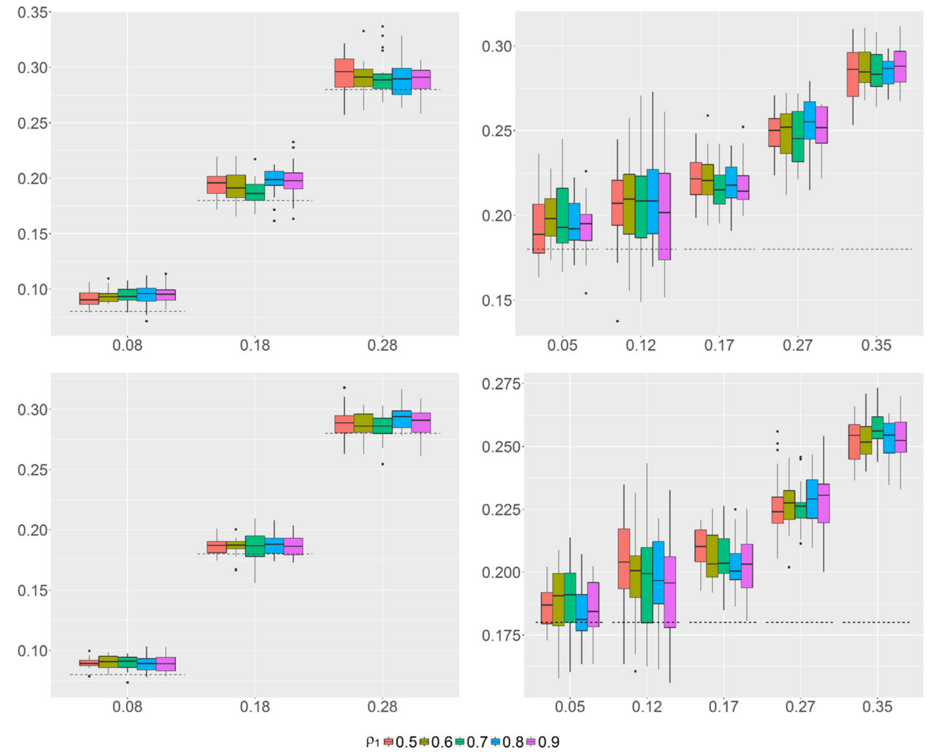

Fig. 3 ρ_0 estimator. *Left*: estimation of ρ_0 for chromosome 5 under different levels of both H_0 and H_1 correlations ($\rho_0 = 0.08, 0.18$ and 0.28). *Dashed lines* indicate the true ρ_0. *Right*: estimation of ρ_0 for $\rho_0 = 0.18$ and different levels of H_1 correlations according to the fraction of H_1 correlations (the results are showed for five typical chromosomes only). *Top*: $n = 50$. *Bottom*:$n = 100$

correlations, as showed in Fig. 3. Importantly, while overestimation of ρ_0 will result in a decrease of power, it will not increase the false positive rate (FDR or FWER).

Performance evaluation

To assess the performance of SegCorr, the true positive rate (TPR = sensitivity), false positive rate (FPR = 1− specificity) and area under the ROC curve (AUC) were considered. These criteria were first computed at the gene level. However, as the goal is to identify correlated regions, a definition of TPR and FPR at the region level was adopted. We considered the intersection between the true and the estimated segmentations and computed the number of true/false positive/negative regions. This amounts at classifying each gene into one of four status (true/false × positive/negative) and then to merge neighboring genes sharing a same status into regions. The status of a region is given by the status of its genes. Consequently, criteria computed at the region level are more stringent as they measure the precision of region boundary estimation.

Figure 4 (top) shows the AUC for Scenario 1 under various configurations, with ρ_1 fixed at 0.5. When ρ_0 is between 0.08 and 0.18, most regions are correctly detected. For $\rho_0 = 0.28$ (a value higher than what is observed on the reference dataset, see Fig. 2), the task becomes difficult and the performance deteriorates.

For Scenario 2, the behavior of SegCorr was explored under different ρ_1. Obviously the task becomes easier when ρ_1 gets larger. Figure 4 shows that SegCorr performs well when $0.5 \leq \rho_1 \leq 0.9$. When $\rho_1 \leq 0.5$, (remind that the background correlation can be as high as 0.2, see Fig. 2) although the performances remain good at the gene level, the boundaries of the regions are detected less accurately.

Comparison with the TCM algorithm

SegCorr was compared with the TCM algorithm introduced by [28] for the detection of regional correlations. The choice of the TCM as a competing method was based on the availability of the code. Indeed, the code of CluGene [13] is not currently available and this of G-NEST [20] relies on obsolete linux packages. Figure 5 displays the AUC achieved by SegCorr and TCM under Scenario 2 for $\rho_1 = 0.5$. When ρ_0 is large ($\rho_0 = 0.28$), one observes that the mean performance of both methods are comparable with higher variability for SegCorr at the gene level and at the region level for TCM. Since the aim is to detect regions rather than genes, the SegCorr procedure seems more appropriate. For small or medium values of background

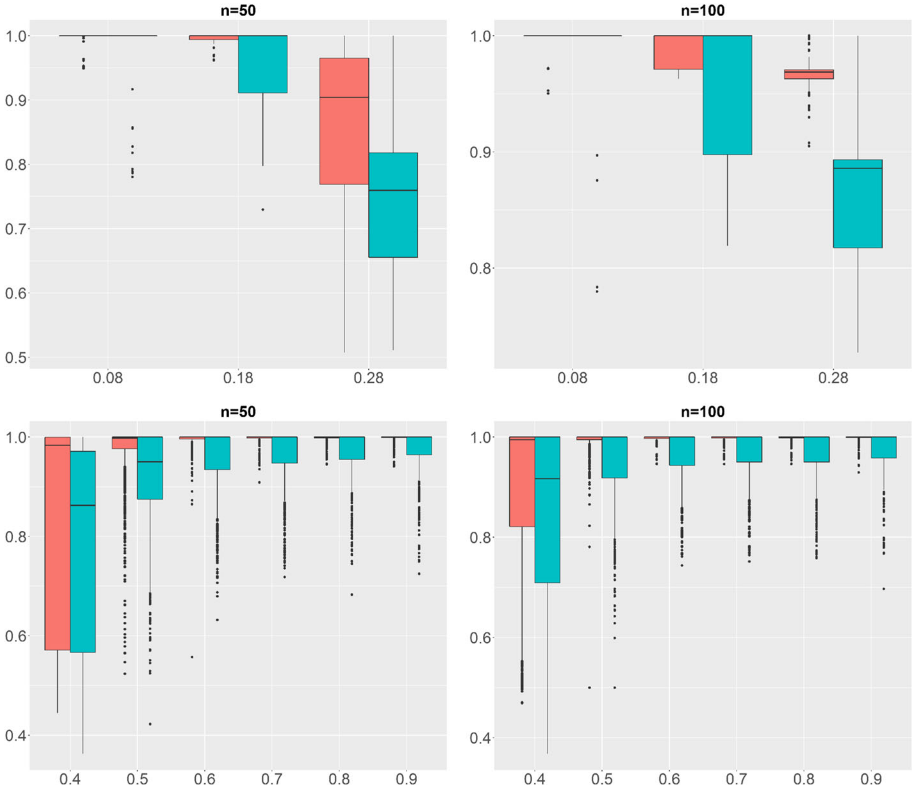

Fig. 4 AUC for Simulation Design 1 and 2. AUC at the gene level (*red*) and region level (*blue*). The higher the AUC the better. *Top*: Simulation design 1 with fixed $\rho_1 = .5$ (*x*-axis: ρ_0). *Bottom*: Simulation design 2 (*x*-axis: ρ_1)

correlations ($\rho_0 = 0.08, 0.18$) SegCorr achieves better AUC than TCM at both the gene and the region levels. As a conclusion, SegCorr appears to be a more consistent and efficient procedure to detect correlated regions. Similar performance between SegCorr and TCM can be observed for other values of ρ_1, results not included here.

Figure 6 illustrates the performance of SegCorr and TCM under Scenario 3. As in the previous case, SegCorr outperforms TCM both on the gene and region level.

We observe that the performance of both algorithms remains unchanged between the different values of β. Further investigations (results not shown) show that classification errors predominantly occur in small regions with or without variability. The simulation shows that only the mean correlation within the blocks matters and that the proposed method is robust to intra-region variability of correlations.

On an Intel i7-4790 CPU processor at 3.60GHz, the CPU times is 74s for SegCorr and 61s for TCM for the bladder cancer dataset. However, in practice TCM must be executed many times in order to manually tune its input parameters (such as the window size and the threshold). On the contrary, SegCorr has to be run only once.

Bladder cancer data

In this section, we apply SegCorr on a bladder cancer dataset described in Section 'Data presentation' below. It is now well known that copy number variation (CNV) impacts gene expression [29]. Here our goal is to detect regions where the correlation is not due to CNV occuring in cancer. Therefore we correct the expression signal for CNV variation according to the strategy described in Sections 'Accounting for known sources of regulation' and 'Procedure for CNV correction'. The effect of this correction is investigated in Section 'CNV-dependent regions'. Lastly, Section 'CNV-independent regions' illustrates the biological results obtained after correction for CNV.

Data presentation

The dataset consists of $n = 403$ bladder tumors. Gene expression have been measured using RNA-seq. The number of genes per chromosome ranges from 293 to 1695 (with average 702). Additionally CNV data have been obtained with Affymetrix Genome wide SNP 6.0 arrays and methylation data with Illumina Human methylation 450k arrays. All RNA-seq, SNP and methylation data were dowloaded from the TCGA open-access HTTP

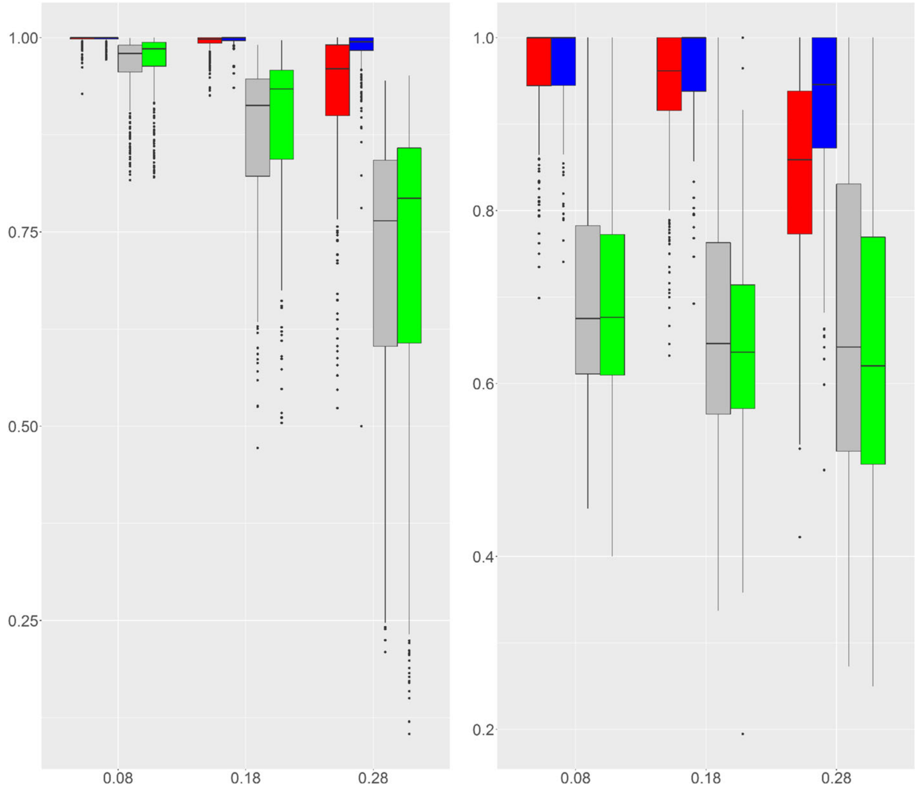

Fig. 5 AUC for SegCorr and TCM (Scenario 2). AUC of the SegCorr ($n = 50$-*red*, $n = 100$-*blue*) and TCM ($n = 50$-*grey*, $n = 100$-*green*) algorithms for Scenario 2 as a function of ρ_0. *Left*: gene level. *Right*: region level

directory (https://portal.gdc.cancer.gov/projects/TCGA-BLCA) and are level 3 data.

Study of the model selection threshold *S*

For the model selection criterion, the threshold S (defined in Section 'Inference of correlated regions', Eq. (6)) must be tuned in such a way to avoid under/over-segmentation. The smaller the value of S the higher the number of segments. As stated in Section 'Model selection', S was fixed to 0.7 as advocated in [17]. Figure 7 shows the evolution of the number and location of H_1 regions detected by SegCorr according to S on a typical chromosome (chromosome 3). One can see that most of these H_1 regions are stable for values of S between 0.6 and 0.9. Still, the value of S may need to be adapted when applied to other data-type or to another dataset. The choice of S can be parametrized in the SegCorr R package, with default value 0.7.

Procedure for CNV correction

To correct the expression signal from CNV, one first needs to detect the CNV regions from the SNP array signal. To this aim, we consider the segmentation method proposed by [26] implemented in the R package cghseg. Denote SNP_{it} the SNP signal of patient i at position t, the model writes

$$SNP_{it} = \mu_{ik} + E_{it} \text{ if } t \in I_k^i = \left[t_{k-1}^i + 1, t_k^i\right]. \quad (8)$$

where the E_{it} are i.i.d centered Gaussian with variance σ^2. The method estimates the number of regions, the boundaries of the regions, denoted $\hat{t}_k^i$ and the signal mean within each region k in patient i, denoted $\hat{\mu}_{ik}$. This procedure may be adapted to count data such as provided by DNAseq data, for which dedicated segmentation tools exist (see e.g. [8]).

We then use the regression model (2) to make the correction where x_{ij} is the mean $\hat{\mu}_{ik}$ obtained previously if the SNP position t corresponds to gene j of the expression signal in patient i. The TCGA expression data arise from RNAseq but are provided as read counts or normalized read counts (RSEM). Then the dataset was normalized using the log $(x + 1)$ method as provided in https://genome-cancer.ucsc.edu/. Finally, we directly applied Model (2) to the normalized RNAseq data.

Still, as often in RNAseq, an important proportion of zero is observed. Genes with null expression in all samples were removed. For the remaining zeros, we either left them when fitting the regression model, or removed them and then set the corresponding residual $\widetilde{Y}_{ij}$ to 0 (note that, in the last option, these observations do not contribute to the estimation of the between-gene correlation,

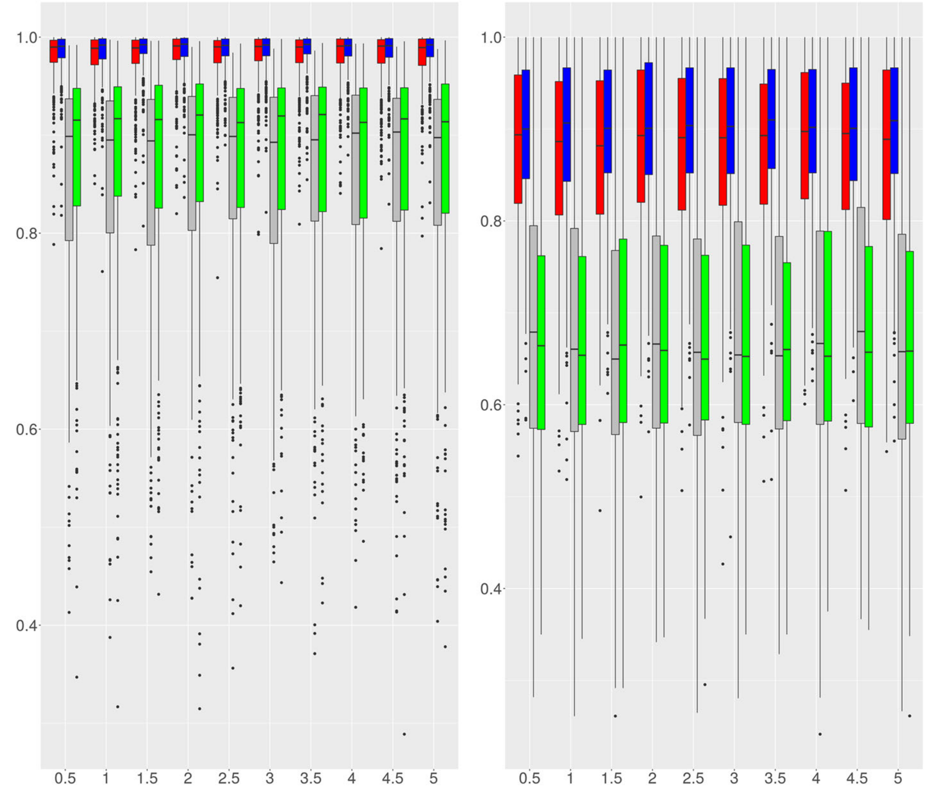

Fig. 6 AUC for SegCorr and TCM (Scenario 3). AUC of the SegCorr ($n = 50$-*red*, $n = 100$-*blue*) and TCM ($n = 50$-*grey*, $n = 100$-*green*) algorithms for Scenario 3 as a function of β. *Left*: gene level. *Right*: region level

as the mean of the residuals is 0 by construction). Both options were found to provide similar results, so only the ones obtained with the first option are displayed in the following.

Since the SNP and expression signals are not aligned, there might be either one, many or no SNP probes that belong to the corresponding gene region. We then propose to define x_{ij} as follows : if one or many probes are related to gene j, mean $\hat{\mu}_{ik}$ or the average of the different means is considered respectively; if there is no probe, a linear interpolation is performed.

CNV-dependent regions

We first investigate the effect of CNV correction (described in Section 'Procedure for CNV correction') by comparing the results obtained on the raw and corrected signals. Figure 8 displays the number of significant H_1 regions as a function of the test level α for both the raw and corrected signals. For small values of α (which are typically used for testing significance), the number of detected regions are quite similar. However, only one third of the detected genes are common, meaning that the regions detected with the two signals are quite different. Furthermore, as the correction removes all effects due to CNV, the estimated background correlation is lower in the corrected signal than in the raw signal (mean decrease across all chromosomes of .07). This makes the test we propose more powerful and explains why, while CNV-due regions are removed, the number of detected regions for a given α remains about the same.

To illustrate this phenomenon more precisely, we considered a set of four regions in chromosomes 3, 8, 10 and 12 known to be associated with CNV in bladder cancer [31, 35, 39]. These regions, given in Table 1, are detected by SegCorr when applied to the raw expression data. When considering the corrected signal, these regions are not detected any more. For the region in chromosome 10, the background correlation was $\widehat{\rho}_0 = 0.221$ and the correlation within this region was $\widehat{\rho}_k = 0.405$, resulting in a highly significant p-value: 8.25e-06. After correction we get $\widehat{\rho}_0 = 0.152$ and $\widehat{\rho}_k = 0.134$, which results in a non-significant p-value: 0.623.

More generally, over the 119 regions solely detected on the raw signal with p-value smaller than 5% (before multiple testing correction), one third (44) get non significant when considering the corrected signal. This explains

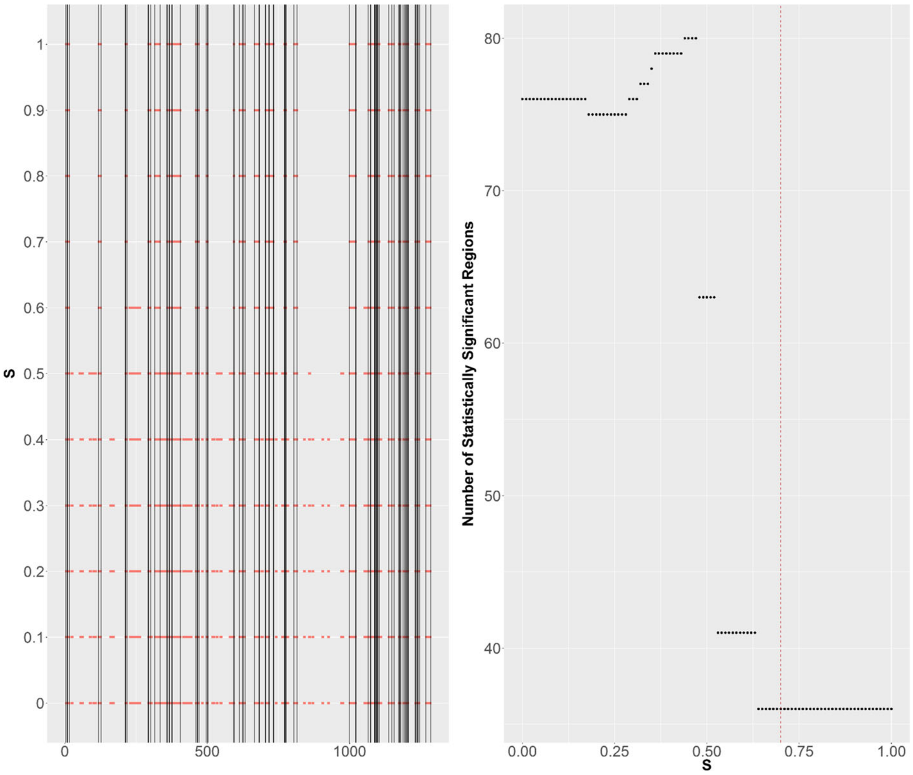

Fig. 7 Choice of *S*. *Left*: statistically significant regions in *red* obtained for different values of *S*. The *vertical lines* correspond to the ones obtained with the default value of *S* we considered ($S = 0.7$). *Right*: number of statistically significant regions for different values of *S*. The *dotted vertical red line* corresponds to $S = 0.7$

a substantial part of the difference between the regions detected on raw and corrected signals. This also shows that the proposed CNV correction strategy performs reasonably well.

CNV-independent regions

General description When applied to the CNV-corrected expression signal, SegCorr detected 588 significant regions (adjusted p-value $\leq$ 0.05) which are distributed throughout the genome (an average of 25 regions per chromosome). Among these regions, 135 regions contained well known gene family clusters such as the HOXA, HOXB, HOXD clusters, several KRT clusters, the epidermal differentiation complex, and HLA gene families clusters whose expression is known to be co-regulated [33]. We next undertook a Gene Ontology terms analysis with genes contained in the significant regions and identified an enrichment of genes belonging to the keratinization pathway (p-value 4.09E-19 and FDR q-value 9.01E-16). The expression of this pathway is strongly associated with a subgroup of bladder cancer called basal-like bladder cancer [27].

Epigenetic regions Apart for CNV, DNA methylation is one of the possible explanations for expression correlation. We first investigated whether the correlation between gene expression and DNA methylation is higher in significant regions when CNV correction is applied. The mean correlation varies marginally when considering the significant regions altogether. This suggests either that methylation is not a systematic cause of expression correlation or that the available signal is too noisy to detect methylation effect.

Still SegCorr allowed us to detect regions where DNA methylation is associated with expression correlation. More specifically, we now present one such region where the observed correlation is not due to CNV and can be associated with an epigenetic mark. This region located on chromosome 17 contains seven genes (*HOXB2, HOXB3, HOXB4, HOXB5, HOXB6, HOXB7, HOXB8*: $\widehat{\rho}_k =$ 0.717, p-value = 7.94e-62). Three genes from this regions have already been studied by [37] and has been referred to as 17-7.

Figure 9 (top) shows a clear pattern detected in both the expression data and the DNA methylation data. When classifying the patients into three groups, the right-most

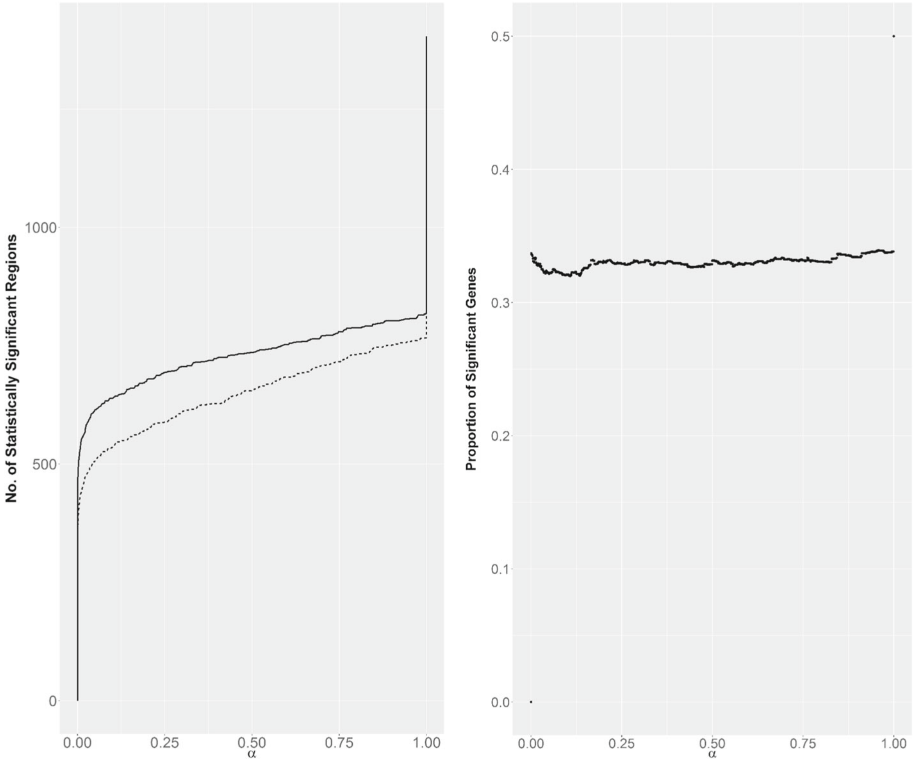

Fig. 8 Bladder Regions. *Left*: Number of statistically significant regions as a function of α (*solid line*: corrected signal, *dotted line*: raw signal). *Right*: proportion of significant genes common in the two signal as a function of α

group displays an over-expression of the genes and a low methylation signal. The methylation of the DNA is one of several epigenetic mechanisms used by the cell to silence the expression of a gene. The tumors that expressed the HOXB gene family present an hypomethylation of the DNA and the tumors which did not express these genes have an hyper methylation of the DNA. This suggests that this region is silenced by an epigenetic mechanism associated with DNA methylation.

Discussion

The identification of co-regulated chromosomal regions is important to fully understand the gene transcription network and to identify new mechanisms of gene regulation and their deregulations in pathological states such as cancer. In this paper, we developed a method to identify these regions and we applied it to cancer data. The method relies on a formal definition of what correlated regions are. It takes advantage of an efficient inference algorithm and a statistical testing procedure, which are both exact. We also proposed a correction strategy that allows one to investigate the possible causes of the observed correlations.

Using this method, we could identify copy number dependent and copy number independent correlated regions of expression. Copy number dependent regions correspond to genomic alterations; copy number independent regions could be due to different mechanisms, including epigenetic mechanism. We showed, for one region, which is part of the HOXB complex, that there is negative correlation between expression and DNA methylation. The detected regions should be further investigated to better understand the underlying mechanism. While the expression data used here were acquired using the RNA-seq technology, any other technology, including microarray technologies can be used as well.

In our analysis, we have assumed stretches of correlated contiguous neighboring genes. This is obviously a

Table 1 Four examples of CNV-dependent regions

Chrom.	Genes
3	*TSEN2, MKRN2, RAF1*
8	*ZNF703, ERLIN2, PROSC, GPR124, BRF2, RAB11, FIP1, ADRB3, EIF4EBP1, ASH2L, STAR, LSM1, BAG4, DDHD2, PPAPDC1B, WHSC1L1, LETM2, FGFR1, TACC1, PLEKHA2, HTRA4, TM2D2, ADAM9*
10	*ASB13, GDI2, ANKRD16, FBXO18*
12	*MDM1, RAP1B, NUP107, SLC35E3, MDM2, CPM, CPSF6, LYZ, YEATS4, FRS2, CCT2, BEST3, RAB3IP, CNOT2*

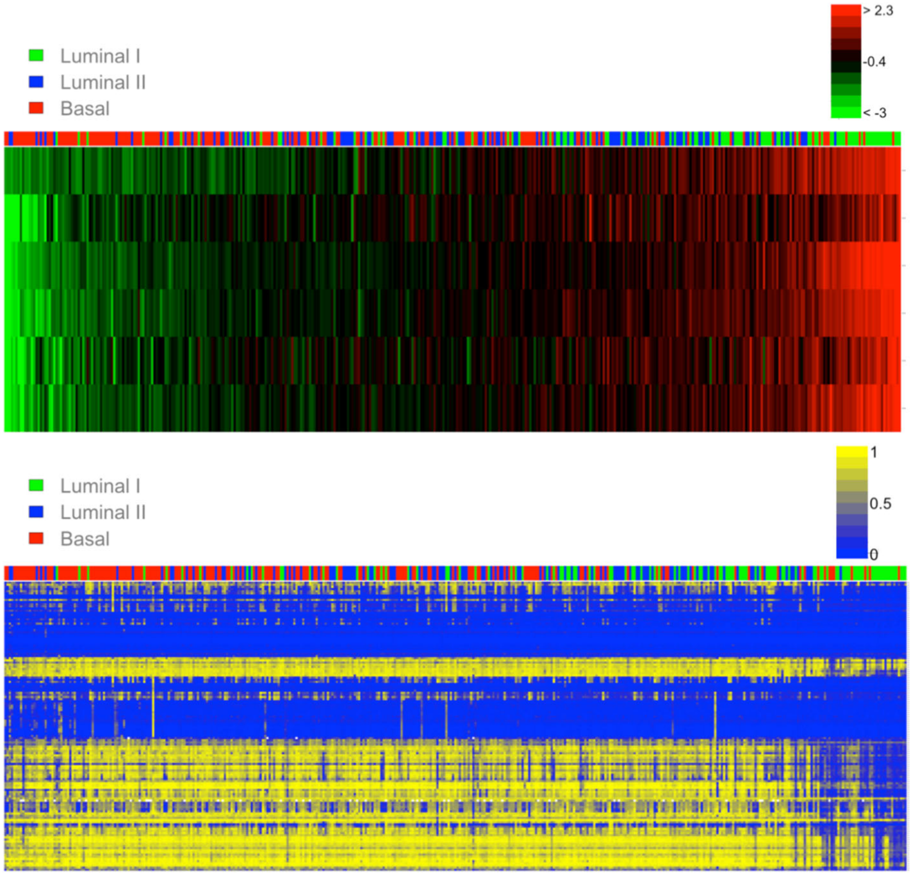

Fig. 9 Heatmaps for Region with Epigenetic Mark. Expression (*top*) and methylation (*bottom*) data from Region 17-7. The tumors have been ordered according to the average expression of the genes from region 17-7. The same ordering of the tumors (*x*-axis) was kept in the *bottom* plot

simplification. Within a correlated region, a gene (or a few genes) could exhibit a weak or even a negative correlation with the other genes. This could occur for different reasons: the gene can be not expressed; alternatively, the gene could be non affected by the regulation process that impacts the other ones; finally, the gene could be impacted in a opposite way compared with the other ones. Note that genes that exhibit no expression or no variation in the dataset can be detected and could be discarded before applying the analysis. While this preprocessing was not required in the present study, running the analysis without removing non-expressed genes would lower the performance of any method aimed at finding correlated (and reasonably homogeneous) regions. Alternatively, accounting for a variable number of uncorrelated genes in correlated regions is an obvious follow-up of the present work.

The proposed correction strategy could easily be generalized to more than one signal to correct for, as it does not rely on a joint modeling of all types of data at hand. Furthermore the segmentation used in the correction step enables one to deal with signals with different probe densities. Finally, this correction approach allowed us to keep all tumors in the study, as opposed to [34] were tumors with CNV in a given region were excluded when analysing this region.

Also, prior information on genes or regions could be accounted for in the segmentation step. Indeed, the likelihood $\widehat{\mathcal{L}}(\tau, \tau')$ associated with a given region can be reweighted or penalized, the dynamic programming algorithm then applies with the same computational complexity.

Conclusions

SegCorr is a novel statistical procedure build for the identification of adjacent co-expressed genes. Some of these regions could be attributed to copy number variation events. To this end, we propose a model to correct gene expression for CNV. This method can be extended for the correction of other data types. R package SegCorr is available on the CRAN.

Additional file

Additional file 1: Appendix file containing a table with all the competing methods, the proof of Lemma 1 and the distribution of the test statistic. (PDF 106 kb)

Abbreviations

AUC: Area under the curve; CNV: Copy number variation; FDR: False discovery rate; FPR: False positive rate; FWER: Family wise error rate; ROC: Receiver operating characteristic; TPR: True positive rate; TF: Transcription factors

Acknowledgments

The authors thank E. Chapeaublanc from Institut Curie for providing part of the data.

Funding

This work has been supported by the INCa_4382 research grant and by Fondation pour la Recherche Médicale (FDM20150633762).

Availability of data and materials

The R package SegCorr can be downloaded from CRAN: https://cran.r-project.org/web/packages/SegCorr/index.html. The results published here are in part based upon data generated by the TCGA Research Network: https://portal.gdc.cancer.gov/projects/TCGA-BLCA and are level 3 data.

Authors' contributions

ED, EL, TMH and SR developed the statistical methodology. ED created the R package and performed the data analysis together with JW. ED, FR and JW interpreted the results of the bladder cancer data. All authors contributed to the redaction of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]AgroParisTech UMR518, 75005 Paris, France. [2]INRA UMR518, 75005 Paris, France. [3]Institut Curie, PSL Research University, 75248 Cedex 05, Paris, France. [4]CNRS UMR144, Equipe Labellisee par La Ligue Nationale contre le Cancer, 75248 Cedex 05, Paris, France. [5]INRA, UMR 0320 - UMR 8120 Genetique Quantitative et Evolution-Le Moulon, F-91190 Gif-sur-Yvette, France. [6]Molecular Oncology Unit, Department of Biochemistry, Hospital Saint Louis, AP-HP, 75475, Cedex 10, Paris, France. [7]Université Paris Diderot, Sorbonne Paris Cité, CNRS UMR7212/INSERM U944, 75475, Cedex 10, Paris, France.

Received: 7 June 2016 Accepted: 26 June 2017
Published online: 11 July 2017

References

1. Aldred P, Hollox E, Armour J. Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. Hum Mol Genet. 2005;14(14):2045–52. doi:10.1093/hmg/ddi209.
2. Anderson T. An introduction to multivariate statistical analysis, 1st edn. New York: Series in Probability and Statistics, Wiley; 1958.
3. Auger I, Lawrence C. Algorithms for the optimal identification of segment neighborhoods. Bull Math Biol. 1989;51(1):39–54. doi:10.1007/BF02458835.
4. Bickmore WA. The Spatial Organization of the Human Genome In: Chakravarti A, Green E, editors. Annual Review of Genomics and Human Genetics, VOL 14, Annual Review of Genomics and Human Genetics, vol 14. Palo Alto: Annual Reviews; 2013. p. 67–84. doi:10.1146/annurev-genom-091212-153515.
5. Bien J, Tibshirani RJ. Sparse estimation of a covariance matrix. Biometrika. 2011;98(4):807–20. doi:10.1093/biomet/asr054.
6. Clark SJ. Action at a distance: epigenetic silencing of large chromosomal regions in carcinogenesis. Hum Mol Genet. 2007;16(1):R88—R95. doi:10.1093/hmg/ddm051.
7. Cleynen A, Dudoit S, Robin S. Comparing segmentation methods for genome annotation based on rna-seq data. J Agric Biol Environ Stat. 2014;19(1):101–18.
8. Cleynen A, Koskas M, Lebarbier E, Rigaill G, Robin S. Segmentor3IsBack: an R package for the fast and exact segmentation of Seq-data. Algorithms Mol Biol. 2014;9(1):1–11. doi:10.1186/1748-7188-9-6.
9. Cohen B, Mitra R, Hughes J, Church G. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet. 2000;26(2):183–6.
10. Coppe A, Danieli GA, Bortoluzzi S. REEF: searching REgionally Enriched Features in genomes. BMC Bioinforma. 2006;7(1):1–7. doi:10.1186/1471-2105-7-453.
11. De S, Babu MM. Genomic neighbourhood and the regulation of gene expression. Curr Opin Cell Biol. 2010;22(3):326–33. doi:10.1016/j.ceb.2010.04.004.
12. Dobson AJ. An introduction to generalized linear models. London: Chapman & Hall; 1990.
13. Dottorini T, Palladino P, Senin N, Persampieri T, Spaccapelo R, Crisanti A. CluGene: A Bioinformatics Framework for the Identification of Co-Localized, Co-Expressed and Co-Regulated Genes Aimed at the Investigation of Transcriptional Regulatory Networks from High-Throughput Expression Data. PLOS ONE. 2013;8(6):e66,196. doi:10.1371/journal.pone.0066196.
14. Frigola J, Song J, Stirzaker C, Hinshelwood R, Peinado M, Clark S. Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. Nat Genet. 2006;38(5): 540–9. doi:10.1038/ng1781.
15. Hilbe JM. Negative binomial regression. Cambridge: Cambridge University Press; 2011.
16. Lai W, Johnson M, Kucherlapati R, Park P. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics. 2005;21(19):3763–70. doi:10.1093/bioinformatics/bti611.
17. Lavielle M. Using penalized contrasts for the change-point problem. Signal Process. 2005;85(8):1501–10. doi:10.1016/j.sigpro.2005.01.012.
18. Lavielle M, Teyssière G. Detection of multiple change-points in multivariate time series. Lith Math J. 2006;46(3):287–306. doi:10.1007/s10986-006-0028-9.
19. Leday GG, van der Vaart AW, van Wieringen WN, van de Wiel MA, et al. Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines. Ann Appl Stat. 2013;7(2):823–45. doi:10.1214/12-AOAS605.
20. Lemay DG, Martin WF, Hinrichs AS, Rijnkels M, German JB, Korf I, Pollard KS. G-NEST: a gene neighborhood scoring tool to identify co-conserved, co-expressed genes. BMC Bioinforma. 2012;13:1–17. doi:10.1186/1471-2105-13-253.
21. Levina E, Rothman A, Zhu J, et al. Sparse estimation of large covariance matrices via a nested lasso penalty. Annals of Applied Statistics. 2008;2(1): 245–63. doi:10.1214/07-AOAS139.
22. Lu C, Feng J, Lin Z, Yan S. Correlation Adaptive Subspace Segmentation by Trace Lasso. In: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE; CVF; IEEE Comp Soc; APRS; Australiasn Natl Univ; NICTA; FACE++; Natl Robot Engn Ctr; Google; Disney Res; nVIDIA; Raytheon BBN Technologies; Facebook; Adobe; Kitware; OMRON, SRI Int, IEEE International Conference on Computer Vision; 2013. p. 1345–52. doi:10.1109/ICCV.2013.170, IEEE International Conference on Computer Vision (ICCV), Sydney, AUSTRALIA, DEC 01-08, 2013.
23. Menezes RX, Boetzer M, Sieswerda M, van Ommen GJB, Boer JM. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. BMC Bioinforma. 2009;10:1–15. doi:10.1186/1471-2105-10-203.
24. Nilsson B, Johansson M, Heyden A, Nelander S, Fioretos T. An improved method for detecting and delineating genomic regions with altered gene expression in cancer. Genome Biol. 2008;9(1):1–15. doi:10.1186/gb-2008-9-1-r13.

25. Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ. A statistical approach for array CGH data analysis. BMC Bioinforma. 2005;6:1–14. doi:10.1186/1471-2105-6-27.
26. Picard F, Lebarbier E, Hoebeke M, Rigaill G, Thiam B, Robin S. Joint segmentation,calling, and normalization of multiple CGH profiles. Biostatistics. 2011;12(3):413–28. doi:10.1093/biostatistics/kxq076.
27. Rebouissou S, Bernard-Pierrot I, de Reynies A, Lepage ML, Krucker C, Chapeaublanc E, Herault A, Kamoun A, Caillault A, Letouze E, Elarouci N, Neuzillet Y, Denoux Y, Molinie V, Vordos D, Laplanche A, Maille P, Soyeux P, Ofualuka K, Reyal F, Biton A, Sibony M, Paoletti X, Southgate J, Benhamou S, Lebret T, Allory Y, Radvanyi F. EGFR as a potential therapeutic target for a subset of muscle-invasive bladder cancers presenting a basal-like phenotype. Sci Transl Med. 2014;6(244):244ra91. doi:10.1126/scitranslmed.3008970.
28. Reyal F, Stransky N, Bernard-Pierrot I, Vincent-Salomon A, de Rycke Y, Elvin P, Cassidy A, Graham A, Spraggon C, Desille Y, Fourquet A, Nos C, Pouillart P, Magdelenat H, Stoppa-Lyonnet D, Couturier J, Sigal-Zafrani B, Asselain B, Sastre-Garau X, Delattre O, Thiery J, Radvanyi F. Visualizing chromosomes as transcriptome correlation maps: Evidence of chromosomal domains containing co-expressed genes - A study of 130 invasive ductal breast carcinomas. Cancer Res. 2005;65(4):1376–83. doi:10.1158/0008-5472.CAN-04-2706.
29. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam T, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525–8. doi:10.1126/science.1098918.
30. Seifert M, Abou-El-Ardat K, Friedrich B, Klink B, Deutsch A. Autoregressive Higher-Order Hidden Markov Models: Exploiting Local Chromosomal Dependencies in the Analysis of Tumor Expression Profiles. PLOS ONE. 2014;9(6):1–16. doi:10.1371/journal.pone.0100295.
31. Simon R, Richter J, Wagner U, Fijan A, Bruderer J, Schmid U, Ackermann D, Maurer R, Alund G, Knönagel H, et al. High-throughput tissue microarray analysis of 3p25 (raf1) and 8p12 (fgfr1) copy number alterations in urinary bladder cancer. Cancer Res. 2001;61(11):4514–9.
32. Spellman PT, Rubin GM. Evidence for large domains of similarly expressed genes in the drosophila genome. J Biol. 2002;1(1):1–8. doi:10.1186/1475-4924-1-5.
33. Sproul D, Gilbert N, Bickmore W. The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet. 2005;6(10): 775–81. doi:10.1038/nrg1688.
34. Stransky N, Vallot C, Reyal F, Bernard-Pierrot I, de Medina SGD, Segraves R, de Rycke Y, Elvin P, Cassidy A, Spraggon C, Graham A, Southgate J, Asselain B, Allory Y, Abbou CC, Albertson DG, Thiery JP, Chopin DK, Pinkel D, Radvanyi F. Regional copy number-independent deregulation of transcription in cancer. Nat Genet. 2006;38(12):1386–96. doi:10.1038/ng1923.
35. TCGA. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507(7492):315–22.
36. Tibshirani R, Wang P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. Biostatistics. 2008;9(1):18–29. doi:10.1093/biostatistics/kxm013.
37. Vallot C, Stransky N, Bernard-Pierrot I, Herault A, Zucman-Rossi J, Chapeaublanc E, Vordos D, Laplanche A, Benhamou S, Lebret T, Southgate J, Allory Y, Radvanyi F. A Novel Epigenetic Phenotype Associated With the Most Aggressive Pathway of Bladder Tumor Progression. J Natl Cancer Inst. 2011;103(1):47–60. doi:10.1093/jnci/djq470.
38. van Wieringen WN, Berkhof J, van de Wiel MA. A random coefficients model for regional co-expression associated with DNA copy number. Stat Appl Genet Mol Biol. 2010;9(1):. doi:10.2202/1544-6115.1531.
39. Williams SV, Platt FM, Hurst CD, Aveyard JS, Taylor CF, Pole JCM, Garcia MJ, Knowles MA. High-Resolution Analysis of Genomic Alteration on Chromosome Arm 8p in Urothelial Carcinoma. Genes Chromosomes Cancer. 2010;49(7):642–59. doi:10.1002/gcc.20775.
40. Xiao G, Reilly C, Khodursky AB. Improved Detection of Differentially Expressed Genes Through Incorporation of Gene Locations. Biometrics. 2009;65(3):805–14. doi:10.1111/j.1541-0420.2008.01161.x.
41. Yi Y, Mirosevich J, Shyr Y, Matusik R, George A. Coupled analysis of gene expression and chromosomal location. Genomics. 2005;85(3):401–12. doi:10.1016/j.ygeno.2004.11.011.
42. Zhang Q, Ding L, Larson DE, Koboldt DC, McLellan MD, Chen K, Shi X, Kraja A, Mardis ER, Wilson RK, et al. CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. Bioinformatics. 2010;26(4):464–9. doi:10.1093/bioinformatics/btp708.