BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 24 2009, pages 3228–3235
doi:10.1093/bioinformatics/btp592

Genome analysis

Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities

Martin Schäfer[1,2,*], Holger Schwender[1,3], Sylvia Merk[4], Claudia Haferlach[5], Katja Ickstadt[1,2] and Martin Dugas[6]

[1]Collaborative Research Center 475, [2]Department of Statistics, TU Dortmund University, Dortmund, Germany, [3]Department of Biostatistics, Johns Hopkins University, Baltimore, USA, [4]Galenus-G.H. AG, Basel, Switzerland, [5]MLL Munich Leukemia Laboratory, Munich and [6]Department of Medical Informatics and Biomathematics, University of Münster, Münster, Germany

Received on April 23, 2009; revised on October 9, 2009; accepted on October 12, 2009

Advance Access publication October 14, 2009

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: The analysis of a number of different genetic features like copy number (CN) variation, gene expression (GE) or loss of heterocygosity has considerably increased in recent years, as well as the number of available datasets. This is particularly due to the success of microarray technology. Thus, to understand mechanisms of disease pathogenesis on a molecular basis, e.g. in cancer research, the challenge of analyzing such different data types in an integrated way has become increasingly important. In order to tackle this problem, we propose a new procedure for an integrated analysis of two different data types that searches for genes and genetic regions which for both inputs display strong equally directed deviations from the reference median. We employ this approach, based on a modified correlation coefficient and an explorative Wilcoxon test, to find DNA regions of such abnormalities in GE and CN (e.g. underexpressed genes accompanied by a loss of DNA material).

Results: In an application to acute myeloid leukemia, our procedure is able to identify various regions on different chromosomes with characteristic abnormalities in GE and CN data and shows a higher sensitivity to differences in abnormalities than standard approaches. While the results support various findings of previous studies, some new interesting DNA regions can be identified. In a simulation study, our procedure also shows more reliable results than standard approaches.

Availability: Code and data available as R packages edira and ediraAMLdata from http://www.statistik.tu-dortmund.de/~schaefer/

Contact: martin.schaefer@udo.edu

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

1 INTRODUCTION

Abnormalities in the human genome are known to have a major impact on the susceptibility of developing tumors (Vogelstein and Kinzler, 2004). Examples of such abnormalities are changes in copy number (CN) and gene expression (GE), loss of heterocygosity or alternative splicing. CN alterations are among the structural genetic changes most investigated for their role in influencing cancer risk. It has been found that CN alterations of oncogenes, tumor suppressor genes and stability genes are responsible for cancer genesis at various steps (Pinkel and Albertson, 2005).

CN alterations can modify the function of genes in many ways, affect the dosage of the corresponding genes and can also influence the structure and regulation of genes located further away. Recent studies estimate that CN variants are responsible for >15% of heritable variation in GE (Stranger *et al.*, 2007), which in turn controls the functions of the human organism. For cancer cells, the impact of CN alterations on GE can be suspected to be even stronger. It may be assumed that the integration of CN and GE data can help to reveal 'driver' genes, i.e. genes causally involved in cancer pathogenesis, as opposed to 'passenger' genes who mutate during pathogenesis without favoring cancer (see, e.g. Haverty *et al.*, 2009).

While alterations of CN can sometimes be negatively associated with expression levels of such genes that lie outside the affected region (as described, e.g. in Lee *et al.*, 2006), an altered expression level of a gene is in general positively associated with the CN corresponding to its locus: losses of chromosomal material—provoked by monosomies, deletions or unbalanced translocations—tend to reduce expression, while gains of chromosomal material—provoked by trisomies, duplications or unbalanced translocations—are likely to increase it (Pinkel and Albertson, 2005). Based on this relation, we concentrate on finding the genes for which GE as well as CN are abnormal, i.e. present a notable deviation of individual patient values from the median of the references, toward the same direction. A presence of such equally directed abnormalities at a gene in DNA of cancer patients indicates that it may be a 'driver' gene.

Two issues are important in characterizing an integrated analysis of CN and GE. The first is whether CN and GE are considered simultaneously, i.e. in a bivariate approach. Often they are not,

*To whom correspondence should be addressed.

opting instead for a two-step procedure in which initially regions with predominant gains and losses of DNA material are identified (e.g. via assignment of the calls 'normal', 'gain' or 'loss'). Then genes in these regions are tested for differential expression, or over- and underexpressed genes are searched for separately, and eventually a list of genes is obtained for which CN and GE are both abnormal. Such two-step approaches have been proposed, e.g. by Bicciato *et al.* (2009), Heidenblad *et al.* (2005), Orozco *et al.* (2009), van Wieringen *et al.* (2006) and Yoshimoto *et al.* (2007). Many of these procedures at some point use an arbitrary threshold for CN or GE data, e.g. a quantile, to define abnormality. As a consequence, the information on abnormalities is partially discretized, leading to a loss of information and ignoring the amount of uncertainty involved in the assignment of CN calls. In testing for differential expression, a choice has to be made as to between which CN categories testing should be carried out, while testing may not be possible for some genes with an unfavorable call distribution of the patients. A two-step approach described by van Wieringen and van de Wiel (2009) employs call probabilities for CN data instead of calls to tackle some disadvantages of two-step procedures, but still potentially needs to exclude genes due to restrictions on the call probability distribution, and has to take a decision as to which kind of abnormality (gain or loss) should be considered for a specific gene. These disadvantages can be avoided in a bivariate approach that directly assesses the joint abnormality of CN and GE for genes (and genetic regions).

The type of relationship between CN and GE measured in an integrated analysis is the second important issue. Often integrated bivariate analyses of CN and GE focus on the dependence between the two variables, i.e. they look for pairs of data points for which the relation between the two variables is such that the behavior of one can be deduced from the behavior of the other. Typical examples are correlation (e.g. Gu *et al.*, 2008; Kotliarov *et al.*, 2009; Lee *et al.*, 2008) or regression analyses (e.g. Chu, 2007; Gu *et al.*, 2008; Menezes *et al.*, 2009; Stranger *et al.*, 2007). Berger *et al.* (2006) proposed a generalized singular value decomposition algorithm that iteratively selects genes with correlated patterns of variation for both variables. Dependence as attested by such approaches, e.g. by a significant Pearson's correlation coefficient, does not imply that the CN and GE values of the cases tend to differ considerably toward the same direction from those in healthy individuals. On the contrary, both variables may be very abnormal in this sense and still result in a correlation or regression coefficient near zero (cf. Section 2.2). Thus, analyzing the dependence between CN and GE can produce different results than assessing equally directed abnormalities of both CN and GE in comparison with reference data. We argue that the latter may be most suitable for the identification of DNA regions that play an important role in oncogenesis.

A few works have subsequently realized both a two-step analysis and an analysis of dependence (e.g. Järvinen *et al.*, 2006; Lipson *et al.*, 2004; Pollack *et al.*, 2002; Tsukamoto *et al.*, 2008).

To our best knowledge, the modified correlation approach presented in this study is the first one that combines a bivariate analysis with an assessment of the equally directed abnormality of CN and GE, avoiding the discussed shortcomings of other methods. We demonstrate the utility of the proposed methodology on a set of patients suffering from acute myeloid leukemia (AML) and in a simulation study, comparing it with two other standard methods.

The outline of this article is as follows. Section 2 introduces the methodology, while Section 3 gives a description of the AML data. In Section 4, the results of the application of our procedure to the AML data are discussed and compared with the results of two other methods. Finally, Section 5 presents conclusions and an outlook.

2 METHODS

2.1 Segmentation of CN data

CN and GE in general are measured on different platforms, i.e. possibly not at the same locations and based on DNA sections of different lengths. As a consequence of the differences in data structure, each GE value has to be assigned a CN value prior to an integrated analysis. Different approaches have been proposed for different platforms for this purpose: Myllykangas *et al.* (2008), e.g. assign to each GE data point the value of the closest available CN data point, whereas Järvinen *et al.* (2006) and Tsafrir *et al.* (2006) use linear interpolation procedures. Another possibility is to employ a segmentation algorithm on the raw CN data, denoising them (Lai *et al.*, 2005; for an integrated analysis of CN and GE, see, e.g. Haverty *et al.*, 2008). In estimating both the breakpoints at which the CN changes and the respective constant CN values between those breakpoints, the original nature of CN, i.e. integers that are constant over DNA segments, can be partly recovered: while the resulting values can take any rational number, they are a piecewise constant function of the location. In this way, CN values are obtained for each GE data point. We follow the latter approach and use the segmentation algorithm proposed by Huber *et al.* (2006).

Segmentation algorithms are generally carried out on $\log_2$-ratios of patient to reference data. This implies the strong assumption of equal breakpoints for patients and references, which we avoid by segmenting $\log_2$-transformed raw CN intensity data of both groups separately. Values of patients and controls are related afterwards by the methodology presented in Section 2.2. In Huber *et al.* (2006), the raw CN intensities of a given chromosome are modeled as a piecewise constant function of the data points $i=1,\ldots,n$, and the persons $j=1,\ldots,m$:

$$\breve{x}_{ij}=\mu_{js}+\varepsilon_{ij},\quad t_s\leq i<t_{s+1},$$

where $\breve{x}_{ij}$ is the $\log_2$-transformed raw CN intensity for person j at data point i, while $t_s, s=2,\ldots,S$, are the segment boundaries, $t_1=1$ and $t_{S+1}=n+1$, μ_{js} is the expected CN intensity for person j in segment s.

In each case, we choose the model that minimizes the residual sum of squares

$$\mathrm{RSS}(t_1,\ldots,t_S)=\sum_{s=1}^{S}\sum_{j=1}^{m}\sum_{i=t_s}^{t_{s+1}-1}(\breve{x}_{ij}-\hat{\mu}_{js})^2.$$

As an estimator $\hat{\mu}_{js}$ for the expected CN intensity of person j in segment s, Huber *et al.* (2006) use the arithmetic mean of its $\breve{x}_{ij}$ in segment s. To increase the robustness, we employ the median here and use only the median intensities $\hat{\mu}_{js}$ in the further analysis. In Figure 1, raw CN intensities and the corresponding $\hat{\mu}_{js}$ of one patient are exemplarily plotted for one chromosome.

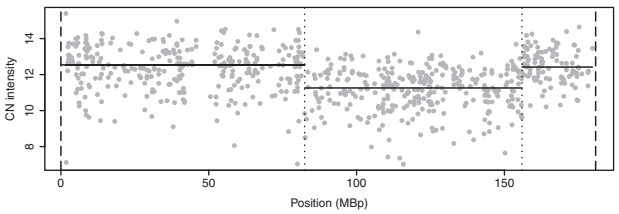

Fig. 1. Example for segmentation of a patient's raw CN intensity data for a chromosome. The vertical dotted lines indicate the segment breakpoints, the dashed lines the chromosome limits and the horizontal solid lines the estimated CN, i.e. the median of the raw values in the segment.

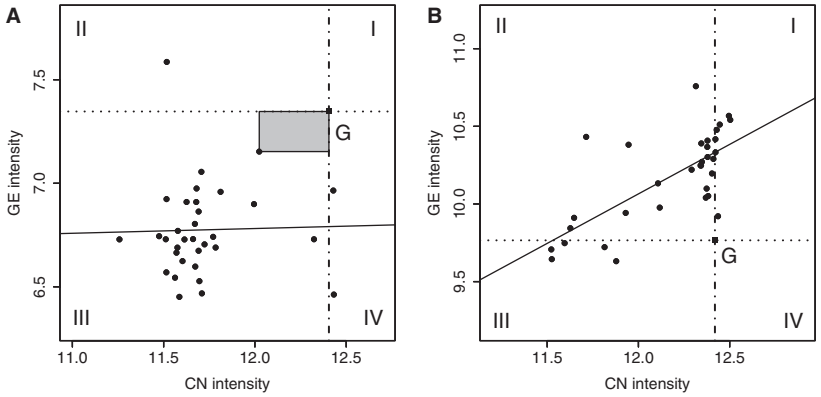

Fig. 2. The idea of the externally centered correlation coefficient. The CN and GE intensities of 33 patients are plotted for the 224963_at probe set (**A**) and the 54970_at probe set (**B**) from the Affymetrix U133 chip set. The dotted and dot-dashed lines represent the median CN and GE intensities in the reference dataset, respectively. G is their intersection. The solid line represents the line of a standard linear regression of GE intensity on CN intensity. The gray rectangle indicates deviation of CN and GE from the median of the references for a single patient.

The only free parameter that remains in the model is S, the number of segments. Huber *et al.* (2006) propose a penalized maximum likelihood approach to choose the value of S and discuss Akaike's information criterion (AIC) or the Bayesian information criterion (BIC) to control the overfitting due to too many segments, noting that the BIC works particularly well in simulations. Following this argumentation, we use the BIC (for an own comparison with the AIC, see Section 1 of the Supplementary Material).

2.2 The externally centered correlation coefficient

Popular bivariate measures for assessing the association between CN and GE are, e.g. Pearson's correlation coefficient and the related approach of regressing GE on CN. In our analysis, we modify Pearson's correlation coefficient in a way that permits to measure equally directed deviations of CN and GE in patients from the median of the references.

In Figure 2, $\log_2$-transformed GE intensities and $\log_2$-transformed and segmented CN intensities of two example genes are plotted for the patients. In Figure 2A, the values of both inputs are smaller than in the reference data for this locus, which therefore is an example for equally directed deviations as we want to detect them. As the regression coefficient is almost zero, though, a regression or correlation analysis in this case fails to indicate this abnormality. In Figure 2B, an example for the reverse situation is given: the regression coefficient is high but the deviations are not equally directed. Assessing the association of CN and GE via regression or correlation therefore may yield results contrary to those desired. Employing $\log_2$-ratios of patients to controls instead of the raw CN values would not change the significance of a correlation or regression coefficient.

Apparently, the degree of equally directed deviations from the median of the references for a gene depends on the location of the patients' values relative to G, the coordinate defined by the median values of the reference group (Fig. 2). If a patient's values are located in the first or third quadrant, they show deviations toward the same direction. If they are located in the second or fourth quadrant, they show deviations toward opposite directions, which means that we are not interested in detecting this data point. The deviation increases with both the distance from G and the distance from the two lines representing the medians of the references. Therefore, we propose to use the area of the rectangles defined by G and each point representing a patient to construct a deviation measure, as suggested in Figure 2.

Starting from Pearson's well-known correlation coefficient, we substitute the means over the cases' CN or GE values by the respective medians of the reference group to obtain a new measure, the externally centered correlation coefficient. For random vectors $\mathbf{X}=(X_1,\ldots,X_m)^T$ and $\mathbf{Y}=(Y_1,\ldots,Y_m)^T$, it is thus defined as

$$r_{EC} := \frac{\sum_{j=1}^{m}(X_j-A)(Y_j-B)}{\sqrt{\sum_{j=1}^{m}(X_j-A)^2}\cdot\sqrt{\sum_{j=1}^{m}(Y_j-B)^2}}. \quad (1)$$

Here, $Y_j, j=1,\ldots,m$, represent the $\log_2$-transformed GE signals of the patients and $X_j, j=1,\ldots,m$, their CN signals estimated from the $\log_2$-transformed CN intensities during the segmentation process, as described in Section 2.1. $\mathbf{U}=(U_1,\ldots,U_p)^T$ consists of reference GE intensities and $\mathbf{V}=(V_1,\ldots,V_t)^T$ of reference CN intensities, $m,p,t\in\mathbb{N}$. Ideally, both are available as paired values of a reference group $(U_h,V_h), h=1,\ldots,w, w=p=t$, but in other cases the analysis is possible as well. The patients' values are centered by the median value of the references for each variable. For consistency with general practice, the median is calculated on the original intensity scale, i.e. $A=\text{logMed}(\mathbf{U})$ and $B=\text{logMed}(\mathbf{V})$ with $\text{logMed}(\mathbf{U})=\log_2(\text{median}(2^{\mathbf{U}}))$. We call the deviations from the reference median in CN and GE, X_j-A and Y_j-B, equally directed if they possess the same sign, leading to a positive sign of r_{EC}.

r_{EC} yields values between -1 and 1 due to the Cauchy–Schwarz inequality, as does Pearson's correlation coefficient.

2.3 Wilcoxon test for symmetry

The externally centered correlation coefficient r_{EC} assesses the deviations of CN and GE intensities from the median of the references for single gene loci. In addition, it is desirable to have a measure of randomness for the degree to which the r_{EC} values exceed zero. To achieve this goal, we suggest to use the following test in an explorative way.

Consider the distribution of the summands $Z_1,\ldots,Z_m$ in (1), i.e.

$$Z_j = \frac{(X_j-A)(Y_j-B)}{\sqrt{\sum_{j=1}^{m}(X_j-A)^2}\cdot\sqrt{\sum_{j=1}^{m}(Y_j-B)^2}}, j=1,\ldots,m.$$

These summands in their absolute values correspond to the areas of the rectangles defined by G and the patients' values (Fig. 2).

Let the random variable Z represent the standardized rectangle areas in the patients' population, and view $Z_1,\ldots,Z_m$ as a random i.i.d. sample of Z. If equally directed deviations in CN and GE intensities are present, we expect the distribution of Z to be slanted toward positive values. This means that at any point, the value of its cumulative distribution function is greater than the one of the cumulative distribution function of $-Z$.

A permutation test, as used, e.g. in Tsukamoto *et al.* (2008) or van Wieringen and van de Wiel (2009) in this situation is potentially problematic as it implies that the empirical distribution of Z in the patients' genome

approximates the random case, which is uncertain for, e.g. study groups with a high frequency of aberrations. As the distribution of the Z_j for one gene is not known either, a classical non-parametric test is proposed.

We thus use the Wilcoxon signed rank test which is a suitable test for the null hypothesis

$$H_0 : P(Z \leq x) = P(-Z \leq x) \quad \forall x \in \mathbb{R}$$

against the alternative

$$H_1 : P(Z \leq x) > P(-Z \leq x) \quad \forall x \in \mathbb{R}$$

(Hollander, 2006). This test could be used to formally test the CN and GE deviations from the reference median in single loci for equally directed abnormalities. Since we, however, are interested in finding potentially abnormal genetic regions, we use the *P*-values as an explorative measure in the algorithm proposed in the next section.

2.4 Automatic detection of abnormal genetic regions

CN alterations and resulting GE changes can typically affect larger DNA sequences. It is therefore desirable to extend the detection of deviations in order to assess regions of neighboring genes. Lipson *et al.* (2004), e.g. use a moving window for this purpose, whereas Kingsley *et al.* (2006) divide the genome in 100 fixed bins. In the following, we discuss how results of the test from Section 2.3 can be used in a more flexible algorithm without fixed region widths to find regions on the DNA that show equally directed abnormalities for CN and GE. The following criteria should be met by such an algorithm:

- It should find those genetic regions that have a high concentration of small *P*-values.
- The more the data points are contained in a region, the higher is the evidence of abnormality in presence of a high proportion of small *P*-values. Therefore, the algorithm should take into account the number of data points in a region, i.e. of two regions showing similarly small *P*-values, the one containing the most data points should be preferred.
- The chance for any region to be identified should not depend on its length, i.e. short and long regions with a similar proportion of small *P*-values should have equal probabilities of being found.

In practice, such an algorithm can be divided into two tasks: iteratively searching for regions and checking whether found regions are sufficiently abnormal, e.g. exceed some prespecified threshold. Different criteria may be used for both tasks. Our algorithm is based on:

(a) the sum of the (previously transformed) *P*-values in a region as a measure for the concentration of small *P*-values.

(b) the proportion κ/ς as a filter criterion, where κ is the number of a region's *P*-values below a threshold τ, and ς is essentially the number of *P*-values in the whole genome. We require $\kappa/\varsigma \geq \pi$ for a region to be identified. τ and π represent the permeability of the filter through which we pass the regions in order to check for their abnormality. Speaking in images, by choosing τ, one can 'raise the bar', while π determines which proportion of the loci should pass it.

A detailed and formal description of the algorithm and the exact calculation of π and τ can be found in Sections 2 and 3 of the Supplementary Material.

3 DATA

In the following, the proposed method is applied to data from a collective of 33 patients suffering from AML that show a complex aberrant karyotype. AML patients often have characteristic cytogenetic abnormalities that are associated with specific disease subtypes and play an important role in survival prognosis. For this dataset, a prior analysis (Merk *et al.*, 2007) found a noticeable reduction of GE intensity as well as DNA losses on the long arm of chromosome 5. These changes are among the most frequent chromosomal aberrations in AML with a complex aberrant karyotype and are, like this karyotype as a whole, associated with a poor prognosis (Haferlach *et al.*, 2004; Schoch *et al.*, 2005).

Between 1999 and 2003, the 33 AML patients (21 males, mean age: 63.4) were assessed in German medical practices. The reference measurements for GE intensity were taken from 11 healthy individuals (5 males, mean age: 40.0, with age missing for one subject) between 2001 and 2003, while for CN intensity they were taken from 2 healthy individuals (1 male, mean age: 30.4) in 2005. All measurements were taken in the same laboratory. Differences between the datasets with respect to age and gender distribution are not considered harmful for CN intensity as it is associated with neither of the two. Possible increases in variance of GE intensity over age (Somel *et al.*, 2006) are eliminated during preprocessing. Genes may be differentially expressed between men and women, but studies indicate that only very few genes are likely to be affected, especially if the sex chromosomes are not considered (McRae *et al.*, 2007; Whitney *et al.*, 2003). Thus, a possible stratification for gender, drastically reducing the sample size, does not appear convenient. DNA was isolated from AML cells and subjected to the Mapping 10K 2.0 Array (CN) and the Human Genome U133 Array Set (GE), respectively. Details on the Affymetrix GeneChip technology can be found in Affymetrix (2003).

4 RESULTS

All calculations were carried out using the statistical software environment R, version 2.9.1 (R Development Core Team, 2009).

Background correction, normalization and summarization for GE data are conducted using RMA (Irizarry *et al.*, 2003) as implemented in the R package *affy*. The data points, i.e. probe sets that do not map uniquely to the genome or that map to one of the sex chromosomes or the mitochondrion genome are excluded. The positions of the remaining 12 811 probe sets from the Human Genome U133 Array Set that serve as the elementary units in our analysis are obtained from the BioMart data management system (December 2008 freeze; http://www.biomart.org/).

The CN intensity data are preprocessed using the CRMA method in the R package *aroma.affymetrix* (Bengtsson *et al.*, 2008a). We follow the procedure described in Bengtsson *et al.* (2008b) to perform allelic crosstalk calibration, summation of allele signals, robust probe-level modeling and PCR fragment-length normalization. Finally, the resulting intensity values are $\log_2$-transformed. Data of patients and references are processed separately in this way, then the segmentation algorithm described in Section 2.1 is carried out on both the datasets, choosing the BIC as penalization term (see Section 1 of the Supplementary Material for details). To each GE probe set and its corresponding GE intensity, we assign the CN intensity estimated for the middle between its beginning and end points. The SNP positions are retrieved from Affymetrix annotation files, additional annotation is obtained from the UCSC Genome Bioinformatics database (December 2008 freeze; http://genome.ucsc.edu/).

For the matched CN and GE values, the values of r_{EC} are computed. In Figure 3A, these values, considering the AML patients compared with healthy individuals, are shown. This figure reveals that chromosome 5 shows the highest values in the genome.

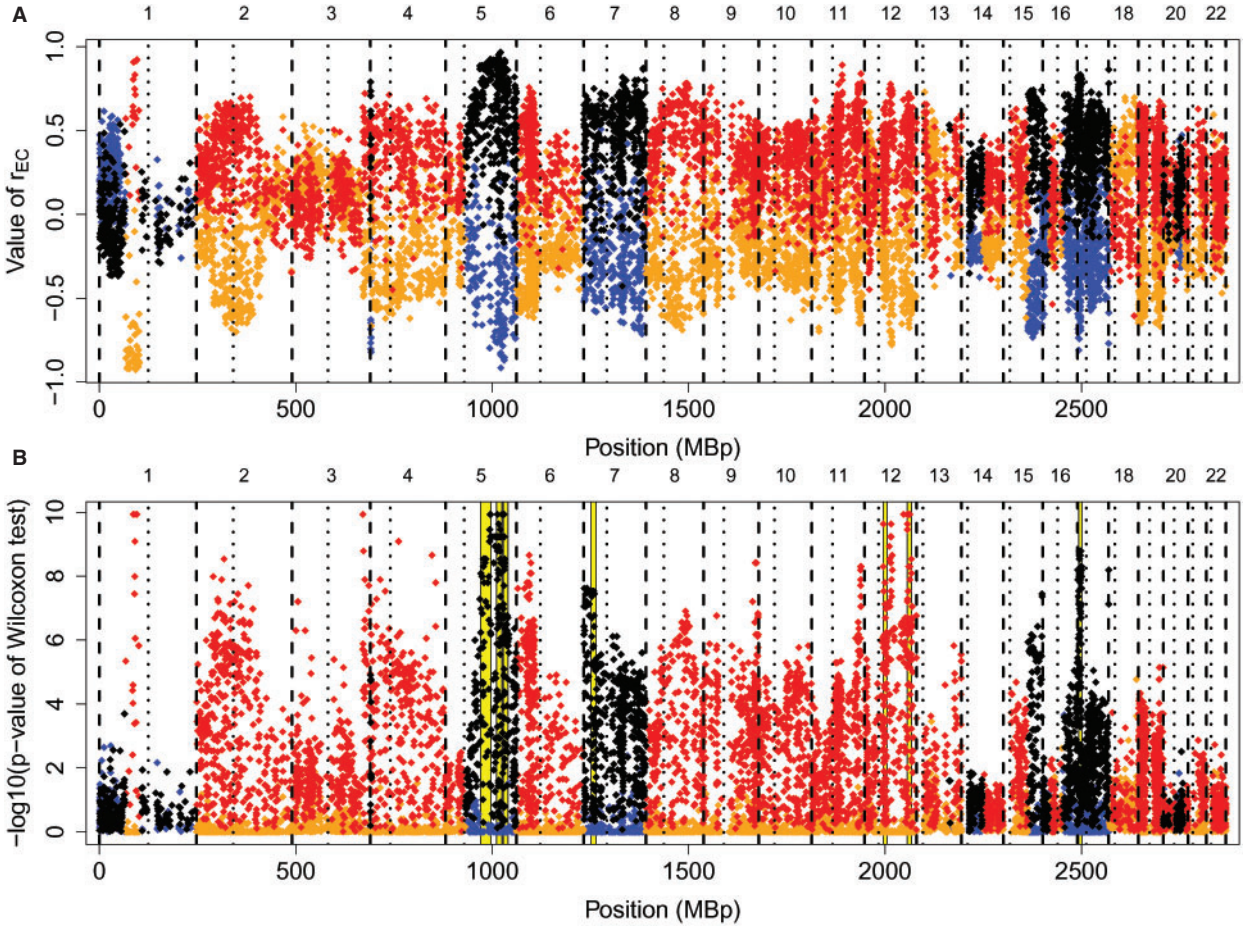

Fig. 3. Values of r_{EC} (**A**) and P-values of the Wilcoxon test (**B**), plotted on a $-\log_{10}$ scale. Dashed lines separate the chromosomes, dotted lines separate their respective p and q arms. The loci are represented by different colors according to the coordinate corresponding to their median CN and GE values. If it falls into the first of the quadrants defined by the median CN and GE values of references, the probe set is marked by a red point, indicating increased CN intensity combined with increased GE intensity. Black points indicate the third quadrant (both CN and GE intensity reduced), blue points indicate the second quadrant and orange points the fourth quadrant. The plotting positions on the x-axis are the middle positions of the probe sets. In (B), regions identified as abnormal toward equal directions ($\tau = 10^{-6}$, $\pi = 0.01$) are marked by a yellow background.

The probe sets on its q arm predominantly display a loss of DNA material accompanied by reduced GE, which is consistent with the findings of Merk *et al.* (2007). Large regions can be seen in which probe sets with either increased or reduced CN, but diverse GE behavior dominate. This makes sense as segmented CN intensities have a smaller variance than GE intensities. As we focus on equally directed deviations from the median of the references, we concentrate on the positive values of r_{EC}, i.e. concordant increase (red) or concordant reduction (black).

The values of r_{EC} are reflected in the P-values of the exploratively conducted Wilcoxon test, as shown in Figure 3B. Here, several deviations appear far more notable than suggested by the values of r_{EC}. The reason is that the denominator of r_{EC} shrinks the measure in its numerator more for probe sets for which the patients' values deviate highly from the median of the references. Since this effect is not relevant for the test's null hypothesis of symmetry, it is not reflected by the P-values. Thus, while r_{EC} is a well-interpretable coefficient to measure deviations of CN and GE intensities, the test is a valuable, further improved measure for assessing them with respect to equally directed deviations from the reference median. More detailed results for each chromosome are shown in Section 5 of the Supplementary Material.

We also apply a correlation procedure similar to the one described by Tsukamoto *et al.* (2008) and the two-step procedure of van Wieringen and van de Wiel (2009) to the AML dataset (for detailed descriptions and results, see Section 4 of the Supplementary Material). These procedures seem less able to separate regions that display a high versus low degree of equally directed deviations, as their results are less distinguished between the chromosomes (cf. Supplementary Figs 3 and 4). Their P-values are not only low for loci at which the majority of CN and GE values display equally directed abnormalities, but also in cases of abnormalities toward opposite directions, which is not desirable.

We conduct our algorithm for finding regions of equally directed abnormalities on the P-values of the Wilcoxon test (see Section 2.4 and Algorithm 1 in Section 2 of the Supplementary Material which also contains a detailed description of the approach). To show how the different degrees of abnormality of regions can be explored by

varying τ and π, we start with $\tau = 10^{-3}$ and $\pi = 0.01$, then reduce τ first to 10^{-4} and then to 10^{-6} (while keeping $\pi = 0.01$), gradually raising the degree of abnormality. For further adjustment, we raise π to 0.03 such that the abnormality of the found regions increases further. The found regions for each parameter specification are given in Table 1 (for the choice of further parameters, see Section 3 in the Supplementary Material). In Figure 3B, the found regions for $\tau = 10^{-6}$ and $\pi = 0.01$ are marked yellow.

These findings support various results of previous studies about CN changes in AML. On chromosome 5q, 7p and 17p, we find regions with strong concordant reductions of CN and GE. These regions have been found to show the most frequent abnormalities in AML with complex aberrant karyotype (see, e.g. Haferlach *et al.*, 2004; Schoch *et al.*, 2005). Gain of DNA material is observed less frequent in this karyotype, but still considerably often. In particular, it occurs on chromosomes 8 and 11 (Mrózek, 2008) on which we also detect regions of concordant increase of CN and GE. In summary, most of the regions we detect as representing a concordant reduction or increase of CN and GE intensities have been discussed in recent literature. Our approach is thus able to find abnormal regions shared by a considerable number of individuals in a study group.

However, there are also some differences between the general picture in literature and our findings. On the one hand, for example, many studies list DNA amplifications and increases of GE intensity on the q arms of chromosomes 20 and 21, which we do not identify. On the other hand, notable regions we find on the p arm of chromosome 7 (concordant reduction) and on the q arm of chromosome 12 (concordant increase) have not been previously discussed (Mrózek, 2008).

The algorithm's findings based on the other methods (see Section 4 of the Supplementary Material) match to some extent with the regions we find when applying it to the *P*-values resulting from testing r_{EC} with the Wilcoxon test. For example, the abnormality regions on chromosome 5 are also detected when considering the *P*-values from the other methods. However, there are also striking differences; the abnormalities on chromosomes 7p, 12q and 17p that are noticeable in our results are missing in those of both other methods.

We furtherly conduct a simulation study to compare our approach with the two other methods. The results (to be found in Section 4 of the Supplementary Material, together with a detailed description of the study) indicate that our approach is better able to measure equally directed deviations from the median of references in CN and GE data than the other methods. In particular, our approach is less prone to bias due to values that do not share the deviation pattern of a majority of patients for a locus.

Table 1. Abnormal regions as identified by Algorithm 1 (Supplementary Material)

Chromosome	Direction	Regions found with different parameter values			
		$\tau = 10^{-3}$ $\pi = 0.01$	$\tau = 10^{-4}$ $\pi = 0.01$	$\tau = 10^{-6}$ $\pi = 0.01$	$\tau = 10^{-6}$ $\pi = 0.03$
2	+	17–29			
	+	39–56	39–56		
	+	61–76	61–76		
5	−	50–116	50–116	89–116	
	−	130–144	130–144	130–144	130–144
	−	147–159	147–159	147–159	
	−	172–181	172–181		
6	+	25–38	25–38		
	+	41–54	41–54		
7	−	1–10	1–10		
	−	19–48	19–48	19–32	
	−	72–112	72–112		
	−	119–131	119–131		
	−	133–144	133–144		
	−	148–157			
8	+	95–106			
9	+	27–38			
	+	99–140	128–140		
10	+	69–78			
	+	94–106	94–106		
11	+	60–81	60–70		
	+	113–125	113–125		
12	+	47–57	47–57	47–57	
	+	108–120	108–120	108–120	
15	−	67–72			
16	−	65–73			
	−	83–89			
17	−	3–11	3–11	3–11	3–11
	−	25–32			
	−	34–49			
19	+	12–19			

The rows show the regions (in Mb), the columns show four different specifications of τ and π. The second column shows the direction of the abnormalities, i.e. concordant reduction (−) or concordant increase (+) of CN and GE.

5 DISCUSSION

The increase of data on different genetic features creates the need for concepts for their integrated analysis. These analyses might help to understand more completely the effects and interactions of different types of genomic variation with respect to disease risk. DNA CN and GE are among the most investigated types of variation in cancer research. An increased CN is one of the main causes that induces a higher expression of the corresponding gene, while a reduced CN can lower the GE. Equally directed abnormalities of CN and GE, i.e. a concordant increase or a concordant reduction of both variables, in patients compared with references may thus help to identify 'driver' genes that are causally related to disease development. We introduced an approach based on the externally centered correlation coefficient, a new measure that assesses the degree to which CN and GE are both altered toward the same direction for a given locus. We provide a Wilcoxon test to assess whether positive deviations from zero are random in the coefficient's values, as well as an algorithm to define abnormality for regions encompassing various genes in an explorative way.

Other methods in general either assess CN and GE abnormalities only consecutively for one variable at a time, losing information, or measure the dependence between CN and GE in a correlation or regression analysis, which does not inform about actual deviations between values from diseased and healthy persons. The ability of our approach to identify regions of concordant abnormality is demonstrated on a group of leukemia patients that were previously found to have characteristic losses of DNA material on chromosome 5. Our procedure identifies this region as abnormal, as well as further regions that have been addressed in other studies.

In addition, some regions are found that have not been discussed previously, which indicates that our procedure can provide new insights to the interplay between chromosomal mutations and GE in cancerogenesis and helps to develop new hypotheses.

On the AML dataset and in a simulation study, our method outperformed two standard approaches with respect to separating regions of strong equally directed abnormalities from the other regions (see Section 4 in the Supplementary Material).

It is strongly recommended to use paired data for the patients as well as for the control group; nevertheless, in case of data shortage CN and GE reference data can also come from two different groups for our approach as the reference values enter in an aggregated way. Unless more standardized protocols are widely used in laboratories, it is advisable to use reference data from the same platform and laboratory to avoid bias (see, e.g. Sherlock, 2005). Ideally, a greater number of reference samples than we used here should be incorporated.

Since many collectives—in particular of patients with a complex aberrant karyotype—are very heterogenous, a useful extension of our method would be to enable the search of abnormalities that are prevalent only in different (unknown) subgroups of patients.

While we focused on assessing equally directed abnormalities of CN and GE at the same loci, the externally centered correlation coefficient can also be used for assessing abnormality of one gene's CN with all other genes' expression, as was done, e.g. in Lee *et al.* (2008) or Stranger *et al.* (2007). This may lead to a fruitful investigation of CN's impact on gene structure and regulation as well as interaction between genes.

Our approach could also be used to integrate other types of data, responding to an increasing general need to integrate various data sources. It is not limited to a special platform and also suited for microarrays of higher resolution than those used here.

We used a Wilcoxon test based on the externally centered correlation coefficient for explorative search of DNA regions with equally directed deviations from the reference median in CN and GE, which is feasible for relatively small sample sizes. The test could also be used to investigate confirmatory hypotheses with respect to the abnormality of certain genes (or genetic regions). In this case, it would be necessary to adjust the *P*-values for multiple comparisons.

Funding: Deutsche Forschungsgemeinschaft (SFB 475, 'Reduction of Complexity in Multivariate Data Structures'); the European Leukemia Network of Excellence (LSHC-CT-2004); COST action BM0801 Translating genomic and epigenetic studies of MDS and AML (EuGESMA).

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2003) *GeneChip® Expression Analysis*. Affymetrix, Santa Clara, CA.

Bengtsson,H. *et al.* (2008a) aroma.affymetrix: a generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Technical report Nr. 745*, Department of Statistics, University of California, Berkeley.

Bengtsson,H. *et al.* (2008b) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**, 759–767.

Berger,J.A. *et al.* (2006) Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **3**, 2–16.

Bicciato,S. *et al.* (2009) A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. *Nucleic Acids Res.*, **37**, 5057–5070.

Chu,J. (2007) Bayesian function estimation using overcomplete dictionaries with application in genomics. PhD Thesis, Department of Statistical Science, Duke University.

Gu,W. *et al.* (2008) Global associations between copy number and transcript mRNA microarray data: an empirical study. *Cancer Inform.*, **6**, 17–23.

Haferlach,T. *et al.* (2004) Genetic classification of acute myeloid leukemia (AML). *Ann. Hematol.*, **83**, S97–S100.

Haverty,P.M. *et al.* (2008) High-resolution genomic and expression analyses of copy number alterations in breast tumors. *Genes Chromosomes Cancer*, **47**, 530–542.

Haverty,P.M. *et al.* (2009) High-resolution analysis of copy number alterations and associated expression changes in ovarian tumors. *BMC Med. Genomics*, **2**, 21.

Heidenblad,M. *et al.* (2005) Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene*, **24**, 1794–1801.

Hollander,M. (2006) *Encyclopedia of Statistical Sciences*. Wiley, New York, NY, pp. 8579–8583.

Huber,W. *et al.* (2006) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, **47**, 1963–1970.

Irizarry,R.A. *et al.* (2003) Summaries of Affymetrix genechip probe level data. *Nucleic Acids Res.*, **31**, e15.

Järvinen,A.-K. *et al.* (2006) Identification of target genes in laryngeal squamous cell carcinoma by high-resolution copy number and gene expression microarray analyses. *Oncogene*, **25**, 6997–7008.

Kingsley,C.B. *et al.* (2006) Magellan: a web based system for the integrated analysis of heterogeneous biological data and annotations; application to DNA copy number and expression data in ovarian cancer. *Cancer Inform.*, **1**, 10–21.

Kotliarov,Y. *et al.* (2009) Correlation analysis between single-nucleotide polymorphism and expression arrays in gliomas identifies potentially relevant target genes. *Cancer Res.*, **69**, 1596–1603.

Lai,W.R. *et al.* (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Lee,H. *et al.* (2008) Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics*, **24**, 889–896.

Lee,J.A. *et al.* (2006) Spastic paraplegia type 2 associated with axonal neuropathy and apparent PLP1 position effect. *Ann. Neurol.*, **59**, 398–403.

Lipson,D. *et al.* (2004) Joint analysis of DNA copy numbers and gene expression levels. In Jonassen,I. and Kim,J. (eds) *Algorithms in Bioinformatics, 4th International Workshop, WABI 2004, Bergen, Norway, September 17-21, 2004, Proceedings*. Springer, Berlin.

McRae,A.F. *et al.* (2007) Replicated effects of sex and genotype on gene expression in human lymphoblastoid cell lines. *Hum. Mol. Genet.*, **16**, 364–373.

Menezes,R.X. *et al.* (2009) Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics*, **10**, 203–217.

Merk,S. *et al.* (2007) Visualization and combined analysis of SNP and gene expression data with Rcnat. *Poster, CAMDA Critical Assessment of Microarray Data Analysis 2007 Conference*, Valencia.

Mrózek,K. (2008) Cytogenetic, molecular genetic, and clinical characteristics of acute myeloid leukemia with a complex karyotype. *Semin. Oncol.*, **35**, 365–377.

Myllykangas,S. *et al.* (2008) Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes. *Int. J. Cancer*, **123**, 817–825.

Orozco,L.D. *et al.* (2009) Copy number variation influences gene expression and metabolic traits in mice. *Hum. Mol. Genet.*, **18**, 4118–4129.

Pinkel,D. and Albertson,D.G. (2005) Array comparative hybridization and its applications in cancer. *Nat. Genet.*, **37**, S11–S17.

Pollack,J.R. *et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

R Development Core Team (2009) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Schoch,C. *et al.* (2005) Acute myeloid leukemia with a complex aberrant karyotype is a distinct biological entity characterized by genomic imbalances and a specific gene expression profile. *Genes Chromosomes Cancer*, **43**, 227–238.

Sherlock,G. (2005) Of fish and chips. *Nat. Methods*, **2**, 329–330.

Somel,M. *et al.* (2006) Gene expression becomes heterogeneous with age. *Curr. Biol.*, **16**, R359–R360.

Stranger,B.E. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Tsafrir,M. *et al.* (2006) Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res.*, **66**, 2129–2137.

Tsukamoto,Y. *et al.* (2008) Genome-wide analysis of DNA copy number alterations and gene expression in gastric cancer. *J. Pathol.*, **216**, 471–482.

van Wieringen,W.N. and van de Wiel,M.A. (2009) Non-parametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, **1**, 19–29.

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

Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789–799.

Whitney,A.R. *et al.* (2003) Individuality and variation in gene expression patterns in human blood. *Proc. Natl Acad. Sci. USA*, **100**, 1896–1901.

Yoshimoto,T. *et al.* (2007) High-resolution analysis of DNA copy number alterations and gene expression in renal clear cell carcinoma. *J. Pathol.*, **213**, 392–401.