A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets†

Silvio Bicciato1,*, Roberta Spinelli2, Mattia Zampieri3, Eleonora Mangano4, Francesco Ferrari5, Luca Beltrame2, Ingrid Cifola2, Clelia Peano2, Aldo Solari6 and Cristina Battaglia4

1Department of Biomedical Sciences, University of Modena and Reggio Emilia, via G. Campi 287, Modena 41100, 2Institute for Biomedical Technologies (ITB), National Research Council (CNR), via Fantoli 16/15, Milano, 3SISSA-ISAS, International School for Advanced Studies, via Beirut 2-4, Trieste, 4Department of Biomedical Science and Technologies and CISI, University of Milan, via Fantoli 16/15, Milano, 5Department of Biology, University of Padova, via U. Bassi 58/B and 6Department of Chemical Engineering Processes, University of Padova, via F. Marzolo 9, Padova, Italy

Received February 24, 2009; Accepted June 1, 2009

ABSTRACT

The integration of high-throughput genomic data represents an opportunity for deciphering the interplay between structural and functional organization of genomes and for discovering novel biomarkers. However, the development of integrative approaches to complement gene expression (GE) data with other types of gene information, such as copy number (CN) and chromosomal localization, still represents a computational challenge in the genomic arena. This work presents a computational procedure that directly integrates CN and GE profiles at genome-wide level. When applied to DNA/RNA paired data, this approach leads to the identification of Significant Overlaps of Differentially Expressed and Genomic Imbalanced Regions (SODEGIR). This goal is accomplished in three steps. The first step extends to CN a method for detecting regional imbalances in GE. The second part provides the integration of CN and GE data and identifies chromosomal regions with concordantly altered genomic and transcriptional status in a tumor sample. The last step elevates the single-sample analysis to an entire dataset of tumor specimens. When applied to study chromosomal aberrations in a collection of astrocytoma and renal carcinoma samples, the procedure proved to be effective in identifying discrete chromosomal regions of coordinated CN alterations and changes in transcriptional levels.

INTRODUCTION

Most tumor cells are characterized by genomics alterations, as polyploidies, imbalances of entire chromosomes and regional amplifications/deletions, which reflect in changes of the DNA copy number (CN) status (1,2). Another genomic aberration typical of tumors is loss of heterozygosity (LOH), i.e. the change from a heterozygous genotype in a normal sample to a homozygous one in a tumor specimen. LOH is due to hemizygous deletion or mitotic recombination and thus can occur with or without associated changes of the CN status (3). The presence of altered DNA CN and LOH may confer growth advantages to cells, which will be selected in descendant cells and contribute to cancer formation. Therefore, the pattern of genomic modifications in a tumor represents a structural fingerprint that may influence the transcriptional control mechanisms and locally impact the gene expression (GE) levels.

Several studies evidenced that, in tumors, there is a correlation between CN and average global expression levels of genes contained in the imbalanced chromosomal region. In their pioneering study, Phillips and coworkers (4) showed that the acquisition of tumorigenicity of an immortalized prostate epithelial cell line resulted in chromosomal gains and losses statistically correlated with increase and decrease in the average expression level of

†This work is dedicated to the memory of Stefano Ferrari.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
involved genes. In particular, 51% of up-regulated genes mapped to regions of DNA gain and 42% of downregulated genes mapped to chromosomal areas of DNA loss. The findings from Phillips et al. were confirmed by the study of Pollack et al. (5) who provided further evidences that widespread DNA CN alteration can lead directly to a global deregulation of GE. In particular, Pollack et al. reported that, in breast tumors and cell lines, DNA CN influences GE across a wide range of DNA CN alterations, with 62% of highly amplified genes showing moderately or highly elevated expression. These observations were further confirmed by Hyman (6) who illustrated a considerable influence of CN on GE patterns. Similar results were later reported in several other tumor types (7–12).

In all of these studies, CN data have been obtained using array-CGH (aCGH) technology. The integration of expression and aCGH data is relatively straightforward, since genes are directly interrogated by specific gene probes and the gene CN is readily available for the same entity interrogated by the expression array. On the contrary, using high-density single nucleotide polymorphism (SNP) arrays, the CN value refers to a SNP marker and the gene CN must be estimated. To date, only two computational procedures have been developed to calculate the gene CN directly from SNP mapping data (i.e. FASeg and dChip) and none directly integrates gene CN data with GE levels.

Linear regression models and statistical analysis of the correlation coefficients between DNA CN and mRNA expression data are normally used to estimate the fraction of all variations measured in mRNA levels that could be attributed to underlying changes in the DNA CN. However, the relationship between CN imbalances and GE changes in the complex genomic environment of a tumor cell may not be effectively captured by the simple, direct correlation of the signal levels. Since gains/losses in the DNA CN may not directly translate to the same quantity of expression change, computational approaches for the integrative analysis of gene dosage and expression data are needed for deciphering how the structural organization of genomes influences their functional utilization. This integrative approach is exemplified by the study of Garraway and colleagues (13), in which the analysis of CN data obtained by SNP mapping arrays drives the investigation of pre-existing GE profiles. Specifically, CN data were used to organize cancer samples into subgroups characterized by specific chromosomal aberrations associated to contiguous SNP chromosomal clusters. This genomic-based sub-grouping constituted the new phenotypic labeling of the samples in the GE analysis, i.e. the NC160 tissues were re-grouped into two new classes based on the presence or absence of the amplification at chromosome 3p14–p13 before performing the supervised analysis. The differential expression profiles, inside the SNP cluster characterizing the CN amplification at 3p14–p13, revealed the presence of a novel melanoma-specific oncogene. Integrated analytical approaches to identify chromosomal regions with significant co-occurrences of genomic imbalances and differential expression may thus represent an opportunity for upgrading the information content of genomic data and for discovering novel cancer biomarkers.

The purpose of this work is to present a bioinformatics procedure that allows the integration of CN, obtained from SNP mapping arrays, with transcriptional data, the identification of genome-wide, concurrent alterations of CN and regional GE in single tumor samples, and the extension of the integrative analysis to entire cancer data-sets. These two issues are achieved in three steps, i.e. (i) the statistical estimation of CN and transcriptional scores at common gene positions from microarray probe-data; (ii) the identification of sets of consecutive genes along the genome characterized by an unusually large number of concurrently altered CN and GE across a single-sample (thereof called Significant Overlap of Differentially Expressed and Genomic Imbalanced Regions, SODEGIR); and (iii) the aggregation of SODEGIRs from different samples to obtain global signatures of tumor types. The first step extends the Locally Adaptive Statistical procedure ([LAP, (14)] to SNP CN data and detects regional alterations of CN at gene level ([Lokern Smoothing Copy Number (LSCN)]. The second part provides the integration of CN and GE data statistically assessing gene dosage and transcriptional statuses on common genomic positions (e.g. Entrez Gene IDs) and identifies the SODEGIRs. The last step combines the various single-sample SODEGIRs into a unique, cancer specific SODEGIR signature. The whole methodology was applied to SNP mapping and GE data obtained by Affymetrix arrays from normal samples (Affymetrix reference), a renal cancer cell line (Caki-1), astrocytoma samples (15), and clear cell renal carcinoma samples (16). All results are available at the Companion Web Site (CWS) http://www.xlab.unimo.it/SODEGIR.

MATERIALS AND METHODS

Genome wide microarray technology and array datasets

CN and GE data have been obtained from public microarray repositories and are fully described in Supplementary Data. Briefly, the datasets comprise Caki-1 (a tumor cell line and reference RNA samples), Astro (a collection of astrocytoma specimens), RCC (tissue samples from renal carcinoma patients) and reference DNA (AffyRef, normal individuals), for a total of 263 Affymetrix Human Mapping SNP and 66 GeneChip HG-U133 Plus 2.0 arrays. Simulated data have been generated to test the performances of all computational steps, as described in Supplementary Data.

Gene expression, CN and LOH data processing

GE values have been quantified using robust multi-array average procedure [RMA, (17)] starting from .CEL files. Chromosome Copy Number Analysis Tool 4.01 (CNAT 4.01, Affymetrix, 2007) and Copy Number Analyzer for GeneChip 2.0 (CNAG 2.0) (18) were used to calculate SNP CN and LOH profiles from mapping arrays (Supplementary Table 1). The forward–backward Fragment Assembling Segmentation algorithm (FASeg) (19) was used to quantify gene CN from CNAT 4.01
SNP CN data, as described in the Supplementary Data. In all datasets, CN and LOH were determined through an un-paired analysis using HapMap samples as normal genotype reference. In addition, a paired analysis was carried out to quantify CN and LOH for the RCC dataset (RCCp) where pairs of tumor tissue and blood samples were available.

The SODEGIR method

The SODEGIR method is a three-step bioinformatics procedure for the identification of genome-wide, concomitant alterations of CN and GE in single samples and in complete datasets (Figure 1 and Supplementary Data). The first step stems from the Locally Adaptive Statistical procedure (LAP) (14), a statistical approach for the identification of imbalances in regional GE. LAP is here extended to SNP CN data, with the aim to detect alterations of regional CN at gene level. The second part statistically assesses the CN and GE statuses on common genomic positions (e.g. Entrez Gene IDs) and identifies the SODEGIRs, i.e. those chromosomal regions where the CN and GE statuses are concordant at a given statistical threshold, in a single sample. In the last step, the various single-sample SODEGIRs are statistically combined to assess a unique SODEGIR signature for an entire dataset. The entire procedure is coded in a set of R functions which are available in the CWS along with documentation and sample data.

Step 1. The first step transforms SNP CN and expression data into CN and GE scores and integrates them with structural information (i.e. chromosomal coordinates), using a kernel regression estimator with an adaptive bandwidth. Resembling LAP (14), the kernel smoothing allows estimating CN and GE scores at the chromosomal locations of Entrez Gene IDs from the probe set data of the microarrays. This first step can be applied separately to SNP CN and GE data. In the former case, the procedure is named LSCN, while the latter represents a revised version of LAP.

CN score. In the LSCN part of the procedure, CN data are transformed into a score $\Delta N_{ij}^{SNP}$ which quantifies, for each SNP $i$ in any sample $j$, the amplitude of the CN variation from the diploid status. Since several evidences questioned the assumption that normal samples have CN equal to 2 everywhere (20,21), the CN value of the diploid status is not set to 2 (i.e. $\log_2$ ratio = 0), but is estimated from the median CN calculated over all $i$ SNP probes ($i = 1, \ldots, L$) of the array. As such, the CN score $\Delta N_{ij}^{SNP}$ can be defined as follows:

$$\Delta N_{ij}^{SNP} = N_{ij}^{SNP} - \min \left( \tilde{N}_{ij}^{SNP}, thrN \right)$$  

where $N_{ij}^{SNP}$ is the CN of SNP $i$ in sample $j$, $\tilde{N}_{ij}^{SNP}$ is the median CN calculated over all the $i$ SNP probes of array $j$ and $\text{thrN} = 0.05$ is a control threshold to cope with potential outlying samples (see Supplementary Data).

GE score. In its original version, LAP calculates a statistic for ranking probes in order of strength of differential expression in two or more populations (14). Here, considering a single sample $j$ from a population of $m$ pathological samples with normalized expression level $x_{ij}$ for probe set $i$ ($i = 1, \ldots, P$) and a population of $n$ normal specimens with average GE $\bar{x}_i^{\text{norm}}$, the GE score $\Delta E_{ij}^{probe}$ can be defined as

$$\Delta E_{ij}^{probe} = \frac{x_{ij} - \bar{x}_i^{\text{norm}}}{s_i + s_0}$$

where the standard deviation $s_i$ for each probe set $i$ is estimated using all pathological and normal samples and is stabilized by the factor $s_0$ as in SAM (22):

$$s_i = \left\{ a \left[ \frac{1}{m + n} \sum_{j=1}^{m} (x_{ij} - \bar{x}_i^{\text{norm}})^2 + \frac{n}{m+n} \left( \sum_{k=1}^{n} (x_{ik} - \bar{x}_i^{\text{norm}})^2 \right) \right] \right\}^{1/2}$$

$$a = \frac{m+n}{m \cdot n \cdot (m+n-2)}$$

Estimation of scores at gene positions: lokerh smoothing. CN and GE scores are estimated at gene positions integrating probe set data and structural information using a kernel regression estimator with an automatically adapted local plug-in bandwidth. Specifically, CN and GE values are estimated at the same gene physical position from the signals of SNP and transcripts probes. As described in ref. (14,23), the integration of variational scores and structural information corresponds to estimate the value of a score at a given chromosomal coordinate, e.g. the Entrez Gene physical position of a gene in base pairs. This integration can be formally stated as a non-parametric regression problem where the score is to be estimated over fixed chromosomal coordinates using a smoothing function. In this particular case, CN and GE scores are integrated with structural information using the lokern functions, a set of kernel regression estimators with adaptive smoothing bandwidth (24). Specifically, for each sample $j$, the regression model specifies:

$$\Delta N_{ij}^{SNP} = \eta_j(M_{bi}) + \xi_{ij}$$

$$\Delta E_{ij}^{probe} = \tau_j(M_{bi}) + \epsilon_{ij}$$

where $M_{bi}$ is the physical position of SNP (probe) $i$, $\eta_j(M_{bi})$ and $\tau_j(M_{bi})$ are arbitrary functions of $M_{bi}$ and $\xi_{ij}$ and $\epsilon_{ij}$ are independent and identically distributed (i.i.d.) errors with zero mean. In these non-parametric models, the systematic part of the variation, i.e. the dependence of $\Delta N_{ij}^{SNP}$ ($\Delta E_{ij}^{probe}$) on the physical position $M_{bi}$ is left as an arbitrary function $\eta_j(M_{bi})$ [or $\tau_j(M_{bi})$], while the random part is specified by assuming that the error components are uncorrelated with zero mean and constant variance. Considering for instance CN values, the regression model takes as input the pairs $(M_{bi}, \Delta N_{ij}^{SNP})$ with $i = 1, \ldots, L$, estimates $E(\Delta N_{ij}^{SNP}) = \eta_j(M_{bi})$ by extracting a curve from the data, and returns the values $\Delta N_{gen}^{SNP}(M_{bi})$ of $\eta_j(M_{bi})$ at given design points $M_{bi}$ (e.g. the $g$ physical position of Entrez Genes).

Thus, the lokern functions take as input a vector of $\Delta N_{ij}^{SNP}$ and/or $\Delta E_{ij}^{probe}$ scores ordered with respect to a vector of $i$ spatial coordinates (i.e. the physical positions
of SNP/transcript probes in the mapping/GE array) and return as output the scores \( \Delta N_{\text{gene}}^{\text{gain,b}} \) and \( \Delta N_{\text{gene}}^{\text{loss,b}} \) estimated at the \( g \) physical position of \( G = 16,395 \) annotated Entrez Genes. The adaptive smoothing bandwidth accounts for the non-uniform distribution and density of genes along the genome and the smoothing function performs a local averaging of the observations when estimating the regression function. The \( \text{lokerns} \) package contains functions that calculate the regression with an automatically chosen local (\( \text{lokerns} \)) or global (\( \text{glikerns} \)) bandwidth.

**Step 2.** In the second step, the goal is to assess the statistical significance of CN and GE variations and to define regions with concomitant alterations of gene CN and GE, respectively: for CN and GE, respectively:

\[
\begin{align*}
H_{\text{gain}}^{N,j} & : \eta_j(M_{bg,j}) = 0 \\
K_{\text{gain}}^{N,j} & : \eta_j(M_{bg,j}) \neq 0 \\
H_{\text{down}}^{E,j} & : \tau_j(M_{bg,j}) = 0 \\
K_{\text{down}}^{E,j} & : \tau_j(M_{bg,j}) \neq 0
\end{align*}
\]

When no alterations of CN (GE) are present along the genome, i.e. when \( \bigcap_{g=1}^{G} H_{\text{gain}}^{N,j} \cap \bigcap_{g=1}^{G} H_{\text{down}}^{E,j} \) is true, the observed data values \( \Delta N_{ij}^{\text{SNP}} = \xi_{ij} \) (\( \Delta E_{ij}^{\text{probe}} = \epsilon_{ij} \)) are i.i.d. realizations and thus are exchangeable:

\[
\begin{align*}
\Delta N_{\text{SNP}}^{\text{SNP}}(\pi(1), \ldots, \pi(L)) & \overset{d}{=} \Delta N_{\text{SNP}}^{\text{SNP}}(\pi(1), \ldots, \pi(L)) \\
\Delta E_{\text{probe}}^{\text{SNP}}(\pi(1), \ldots, \pi(P)) & \overset{d}{=} \Delta E_{\text{probe}}^{\text{SNP}}(\pi(1), \ldots, \pi(P))
\end{align*}
\]

where \( \{\pi(1), \ldots, \pi(L)\} \) and \( \{\pi(1), \ldots, \pi(P)\} \) represent arbitrary permutations of \( \{1, \ldots, L\} \) and \( \{1, \ldots, P\} \), respectively and \( \overset{d}{=} \) denotes equality in distribution. This implies that, starting from the original data, all \( L! (P!) \) permutations of the data are equally likely and that a permutation scheme can be used to identify chromosomal regions with statistically significant CN and GE imbalances. Specifically, at each permutation, \( \Delta N_{ij}^{\text{SNP}} \) and \( \Delta E_{ij}^{\text{probe}} \) scores are randomly assigned to chromosomal locations and \( \Delta N_{ij}^{\text{gene}} \) and \( \Delta N_{ij}^{\text{gene,b}} \) re-estimated using the \( \text{lokerns} \) function (permuted scores \( \Delta N_{ij}^{\text{gene},b} \) and \( \Delta E_{ij}^{\text{gene},b} \)). The permutation process, over \( B \) random assignments, defines the distribution of the null scores for any output design position. Since the observed and expected gene CN and GE scores are estimated using the same function over the same input and output design points, the significance of CN and transcriptional imbalances can be computed testing \( H_{\text{gain}}^{N,j} \) and \( H_{\text{down}}^{E,j} \) on the estimated scores \( \Delta N_{\text{gene}}^{\text{gain}}, \Delta N_{\text{gene}}^{\text{loss}}, \Delta E_{\text{gene}}^{\text{gain}}, \) and \( \Delta E_{\text{gene}}^{\text{loss}} \) as test statistic, respectively. The significance \( p_{g,j}^{N} \) (or \( p_{g,j}^{E} \)) that the expected score \( \Delta N_{\text{gene}}^{\text{gain,b}} \) (or \( \Delta E_{\text{gene}}^{\text{gain,b}} \)) exceeds the observed one \( \Delta N_{\text{gene,b}}^{g,j} \) (or \( \Delta E_{\text{gene,b}}^{g,j} \)), over \( B \) permutations, can be then computed as follows:

\[
\begin{align*}
p_{g,j}^{N} & = \frac{\sum_{b=1}^{B} I\left( \left| \Delta N_{\text{gene,b}}^{g,j} \right| \geq \left| \Delta N_{\text{gene,b}}^{g,j} \right| \right)}{B} \\
p_{g,j}^{E} & = \frac{\sum_{b=1}^{B} I\left( \left| \Delta E_{\text{gene,b}}^{g,j} \right| \geq \left| \Delta E_{\text{gene,b}}^{g,j} \right| \right)}{B}
\end{align*}
\]

where \( I\{\cdot\} \) is an indicator function that takes the value 1 if the argument is true and 0 otherwise.

These \( p \)-values \( p_{g,j}^{N} \) and \( p_{g,j}^{E} \) have the peculiarity to be local, since the observed scores are compared only with the expected ones estimated on the same neighborhood of gene position \( g \). Indeed, during the permutation process, the chromosomal position is conserved while the scores are randomly shuffled. Once the distributions of empirical \( p \)-values have been generated, the \( q \)-value is used to correct the measure of significance for multiple testing.

**Status quantification and SODEGIR definition.** When the null hypothesis \( H_{g,j}^{N} \) is rejected, the CN (or GE) status of a gene \( g \) in a sample \( j \) is decided basing on whether \( \eta_{j}(M_{bg,j}) \) (or \( \tau_{j}(M_{bg,j}) \)) is smaller or greater than zero. The two-sided hypotheses of Equation (5) are equivalent to the simultaneous testing of the following pair of one-sided hypotheses:

\[
\begin{align*}
H_{g,j}^{\text{Gain}} & : \eta_j(M_{bg,j}) \geq 0 \quad \text{against} \quad K_{g,j}^{\text{Loss}} : \eta_j(M_{bg,j}) < 0 \\
H_{g,j}^{\text{Loss}} & : \eta_j(M_{bg,j}) \leq 0 \quad \text{against} \quad K_{g,j}^{\text{Gain}} : \eta_j(M_{bg,j}) > 0 \\
H_{g,j}^{\text{Exp}} & : \tau_j(M_{bg,j}) \geq 0 \quad \text{against} \quad K_{g,j}^{\text{Down}} : \tau_j(M_{bg,j}) < 0 \\
H_{g,j}^{\text{Down}} & : \tau_j(M_{bg,j}) \leq 0 \quad \text{against} \quad K_{g,j}^{\text{Exp}} : \tau_j(M_{bg,j}) > 0
\end{align*}
\]

Considering for instance CN, the rejection of either \( H_{g,j}^{\text{Gain}} \) or \( H_{g,j}^{\text{Loss}} \) is equivalent to the rejection of \( H_{g,j}^{N} \). Although Equations (5) and (8) are equivalent ways of formulating the same hypothesis testing problem, there is some advantage in using the formulation of Equation (8). Indeed, when the action to take in the event of rejection of \( H_{g,j}^{N} \) (or of \( H_{g,j}^{E} \)) depends upon which tail brought about the rejection, \( K_{g,j}^{\text{Loss}} \) or \( K_{g,j}^{\text{Gain}} \) (and \( K_{g,j}^{\text{Exp}} \) or \( K_{g,j}^{\text{Down}} \)) can be associated with the two courses of action.

In particular, the null hypothesis \( H_{g,j}^{N} \) (or \( H_{g,j}^{E} \)) is rejected according to thresholds on the \( q \)-value and on the scores. The \( q \)-value and score thresholds for CN and GE may be set to different values, depending on the desired stringency.
of the analysis (Supplementary Data). The CN q-value and score thresholds have been optimized based on the analysis of the AffyRef Reference DNA dataset, $\Delta q_{g,j}$ thresholds have been selected according to the criteria used for the CN ones, and GE q-value threshold has been set to the value commonly used with GE data (14).

CN and GE statuses are coded as 1 (CN loss, GE down-regulation) when the q-value is below the q-value threshold (e.g. 0 or 0.05) and the score is smaller than the low score threshold (e.g. the 10th quantile of scores), 3 (CN gain, GE up-regulation) when the q-value is below the q-value threshold and the score is larger than the high score threshold (e.g. the 90th quantile of scores, Supplementary Table 2), and 2 (CN and GE neutral) in all other cases. Given the quantification of CN and GE statuses in a single sample, a SODEGIR corresponds to a region of the genome where the CN and GE statuses are concordant (see Supplementary Data for details on the hypothesis intersection-union formulation). In particular, if both CN and GE statuses are equal to 1, the SODEGIR indicates deletion (SODEGIR status 1), while, if CN and GE statuses are both 3, the SODEGIR indicates amplification (SODEGIR status 3).

Step 3. The third step provides a statistical method to elevate the analysis from the single to the multiple-sample level and to detect the presence of a common SODEGIR signature across an entire dataset. In details, let $S_{g,j}$ be the SODEGIR status for sample $j$ at the gene $g$ and assume that $S_{g,j}$ follows a multinomial distribution with $Pr(S_{g,j} = 1) = \theta^1_g$, $Pr(S_{g,j} = 2) = \theta^2_g$, $Pr(S_{g,j} = 3) = \theta^3_g$ and $\theta^1_g + \theta^2_g + \theta^3_g = 1$. Under the null hypothesis that there are no real imbalanced regions, these probabilities are independent from $g$, i.e. $\theta^s_g = \theta^s$, $s = 1, 2, 3$. Then for each gene $g$, the following hypotheses are tested:

$$H^1_g : \theta^1_g = \theta^1 \quad \text{against} \quad K^1_g : \theta^1_g > \theta^1$$

$$H^3_g : \theta^3_g = \theta^3 \quad \text{against} \quad K^3_g : \theta^3_g > \theta^3$$

When there are no real imbalanced regions, i.e. when $\bigcap_{g=1}^{G} H^s_g$ is true, a reasonable estimator of $\theta^s$ is given by $\hat{\theta}^s = \left( \sum_{j=1}^{J} \sum_{g=1}^{G} \frac{S_{g,j}}{J_g} \right) / (J_g)$. The test statistic $T_g = \sum_{j=1}^{J} I[S_{g,j} = s]$, which is distributed as Binomial($J_g, \theta^s$) when $H^s_g$ is true, can be used to test each $H^s_g$. Hence, the p-value is given by

$$p^s_g = Pr(T_g \geq t_g) = \sum_{r=t_g}^{J_g} \binom{J_g}{r} \left( \hat{\theta}^s \right)^r (1 - \hat{\theta}^s)^{J_g-r},$$

where $t_g$ is the observed frequency of SODEGIR status $s$ at gene $g$ across the $J$ samples.

Once computed the $p$-values for each gene, the q-value is used to assign a measure of significance to each of the many tests simultaneously performed and is adopted as a summary score for deletions or amplifications.

RESULTS

The SODEGIR procedure (Figure 1) has been optimized and tested on CN and GE data obtained using Affymetrix Human Mapping 100K or 250K and HG-U133 Plus 2.0 arrays, respectively. In particular, the datasets comprise normal samples (DNA Affymetrix reference, AffyRef, http://www.affymetrix.com/support/datasets.affx), a renal tumor cell line (Caki-I), a subset of 12 astrocytomas samples (Astro) obtained from a public dataset of gliomas (15) and 12 clear cell renal carcinomas (RCCp and RCC) combined in paired normal/tumor specimens (16). All samples were firstly used to tune the parameters of the LSCN part of the procedure, i.e. to define an appropriate CN score and to verify the performances of the kernel-based estimator in calculating gene CN values. Given its definition, the CN score required to quantify the CN of the diploid status. Since several evidences questioned the assumption that normal samples have CN equal to 2 (20,21), the CN value corresponding to the diploid status was not set to 2 (i.e. log2 ratio = 0), but estimated directly from the data. In particular, as shown in Supplementary Figure 2, the median values of the various arrays, although not equal to zero, are tightly distributed around zero (log2 ratio = 0, CN = 2), irrespectively that the data represent normal (AffyRef and Blood) or pathological samples (Astro, RCCp and RCC). Given this evidence, the CN value of the diploid status was quantified as the median CN calculated over all SNP probe sets of a mapping array.

The efficacy of the locally adaptive approach (i.e. the lokerns function) in smoothing GE scores has been already shown in ref. (14), while its performance with
CN scores has been tested through the analysis of all normal and tumor samples. In particular, lokerns estimates CN scores at gene positions preserving the pattern of CN scores of SNP probes (Supplementary Figure 3, Panel A and CWS, AffyRef). Indeed, the estimated gene CN scores of 16 395 annotated Entrez Gene IDs perfectly reproduce SNP CN scores in CN neutral samples (Supplementary Figure 3, Panel A, chromosome 11 in AffyRef NA17203) and in samples presenting broad gains and losses (Supplementary Figure 3, Panel B, chromosome 7 in RCCp 27CG; Panel C, chromosome 10 in Astro HF1232). Moreover, the lokerns function is able to detect mixed patterns of CN changes as spikes and simultaneous loss and gain of chromosomal arms (Supplementary Figure 3, Panel D, chromosome 7 in Astro HF1232; Panel E, chromosome 5 in RCCp 50PC; Panel F, chromosome 3 in RCCp 27CG). Finally, lokerns regresses efficiently the CN score irrespectively of the array density (50K, 100K and 250K sets), although denser arrays allow a finer smoothing of the data using smaller bandwidths (Supplementary Figure 4). Thus, consistently with the GE analysis by LAP, the locally adaptive approach implemented by the lokerns function has been adopted also for regressing CN scores.

Since the true status of genes is unknown in real data sets, the performance of the proposed procedure was assessed on synthetic data through a simulation analysis. Differently from real data, in an artificial data set the true status and the test result of each gene are known. Since the true status of genes is unknown in real data (Supplementary Table 2) lead to a mean sensitivity of 0.91, 0.94 and 0.87 and a mean FDR of 0.014, 0.019 and 0.005 for LSCN, LAP and SODEGIR procedures, respectively.

The entire SODEGIR procedure was then applied to the various cancer data sets to identify single sample and dataset signatures. All results, including chromosome and genome views for single samples and for entire datasets, as well as tables with the characteristics of all CN, GE and SODEGIR regions, are available at CWS. The analysis of the tumor cell line Caki-1 allowed the detection of more

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>CN</th>
<th>GE</th>
<th>SODEGIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caki_100K</td>
<td>469.884</td>
<td>567.536</td>
<td>166.246</td>
</tr>
<tr>
<td>Caki_250K</td>
<td>488.047</td>
<td>567.536</td>
<td>172.702</td>
</tr>
<tr>
<td>HF0017</td>
<td>190.089</td>
<td>67.352</td>
<td>30.451</td>
</tr>
<tr>
<td>HF0108</td>
<td>318.065</td>
<td>164.59</td>
<td>89.831</td>
</tr>
<tr>
<td>HF0152</td>
<td>185.689</td>
<td>218.398</td>
<td>44.591</td>
</tr>
<tr>
<td>HF0491</td>
<td>289.853</td>
<td>184.265</td>
<td>11.172</td>
</tr>
<tr>
<td>HF0608</td>
<td>180.63</td>
<td>356.129</td>
<td>42.516</td>
</tr>
<tr>
<td>HF1139</td>
<td>302.704</td>
<td>349.615</td>
<td>228.719</td>
</tr>
<tr>
<td>HF1232</td>
<td>380.549</td>
<td>301.414</td>
<td>217.407</td>
</tr>
<tr>
<td>HF1269</td>
<td>459.684</td>
<td>432.209</td>
<td>140.952</td>
</tr>
<tr>
<td>HF1344</td>
<td>382.209</td>
<td>393.79</td>
<td>262.443</td>
</tr>
<tr>
<td>HF1442</td>
<td>247.563</td>
<td>176.247</td>
<td>100.145</td>
</tr>
<tr>
<td>HF1469</td>
<td>100.576</td>
<td>67.683</td>
<td>1.784</td>
</tr>
<tr>
<td>HF1511</td>
<td>255.157</td>
<td>39.215</td>
<td>13.809</td>
</tr>
<tr>
<td>RCCp</td>
<td>27CG</td>
<td>452.867</td>
<td>477.841</td>
</tr>
<tr>
<td>28RA</td>
<td>443.696</td>
<td>506.478</td>
<td>291.786</td>
</tr>
<tr>
<td>33BV</td>
<td>509.906</td>
<td>171.066</td>
<td>150.24</td>
</tr>
<tr>
<td>36MML</td>
<td>538.907</td>
<td>572.542</td>
<td>141.616</td>
</tr>
<tr>
<td>37BA</td>
<td>295.591</td>
<td>375.085</td>
<td>36.954</td>
</tr>
<tr>
<td>40RR</td>
<td>208.409</td>
<td>72.964</td>
<td>14.09</td>
</tr>
<tr>
<td>45DM</td>
<td>177.191</td>
<td>399.721</td>
<td>91.845</td>
</tr>
<tr>
<td>46SA</td>
<td>551.945</td>
<td>582.158</td>
<td>194.366</td>
</tr>
<tr>
<td>47CA</td>
<td>380.862</td>
<td>394.464</td>
<td>169.46</td>
</tr>
<tr>
<td>49CA</td>
<td>472.772</td>
<td>336.733</td>
<td>242.828</td>
</tr>
<tr>
<td>50PC</td>
<td>341.269</td>
<td>410.785</td>
<td>159.763</td>
</tr>
<tr>
<td>51MI</td>
<td>416.003</td>
<td>456.929</td>
<td>72.311</td>
</tr>
</tbody>
</table>

Values have been derived from .SDG_Table files of each sample, as deposited in CWS.
at CWS). Figure 2 displays the genome view of Caki-1 with the regions of CN gain/loss, GE up-/down-regulation and deleted (CN loss and GE down-regulation) and amplified SODEGIRs (CN gain and GE up-regulation) are shown as boxes on each chromosome. As in the cPlot view of R geneplotter package, horizontal lines represent chromosomes and grey bars indicate gene positions. Three lines per chromosome and shades of red and green are used to display CN gain/loss, GE up-/down, and SODEGIRs amplified and deleted. (B) Chromosome view of chromosome 3: CN status (N_AB) and LOH status as estimated by the CNAG HMM on each SNP probe, CN, GE and SODEGIR statuses as determined by the SODEGIR procedure on gene positions of chromosome 3. The grey bars indicate SNP probes (in N_AB and LOH lanes) or Entrez Gene ID positions (for CN, GE and SODEGIR lanes). Red and green bars in the N_AB lane indicate N_AB > 3 and < 1, respectively. Blue bars in the LOH lane highlight SNP probes with an inferred LOH value > 20. Green bars in CN, GE, and SODEGIR lanes indicate loss, down-, or deletion (i.e. a status of 1). Red bars in CN, GE, and SODEGIR lanes indicate gain, up-, or amplification (i.e. a status of 3). (C) Genome boxplot: distribution of gene GE scores according to gene CN scores on all SODEGIRs of the entire genome. CN levels are categorized into five bins highlighting two ranges of loss (green boxes, gene CN score < -0.1), one range of diploidy (white box, gene CN score between -0.1 and 0.1) and two ranges of gain (red boxes, gene CN score > 0.1). (D) Chromosome box plot for chromosome 3: distribution of GE scores according to gene CN scores on all the SODEGIRs of a specific chromosome (e.g. chromosome 3).
Gene transcriptional activity is quantitatively correlated to the gene dosage. Indeed, as shown in Figure 2C and D, the CN and expression levels of genes sharing the same status are tightly related both considering the whole-genome or a single chromosome. Similar results have been obtained when Caki-1 DNA was profiled using a Human Mapping 250K Nsp array (Table 1 and Caki250k in CWS).

The application of the entire procedure to astrocytoma and renal carcinoma data allowed determining SODEGIRs for single samples and the definition of dataset SODEGIR signatures. In particular, the analysis of Astro samples highlighted CN regions spanning from 100 to 460 Mb and affecting from 3% to 15% of the human genome. GE regions spanned from 40 to 430 Mb and accounted from 1% to 14% of the genome (Table 1). SODEGIRs are distributed over few chromosomes (from 1 to 3 chromosomes) and span from few Mb up to a maximum of 270 Mb (sample HT1344, for detailed results see CWS). Most samples present broad deleted or amplified SODEGIRs on chromosomes 10 and 7, respectively (as exemplified in Figure 3A by sample HT1139), and these regions determine the strong correlation between gene CN and transcriptional activity at the whole genome level (Figure 3B). Similarly, CN regions ranging from 180 to 540 Mb and GE regions covering from 50 to 580 Mb were determined in RCCp samples (Table 1). Again, SODEGIRs are distributed over few chromosomes (from 1 to 5 chromosomes) and span from 14 (sample 40RR) to 290 Mb (sample 28RA). Deleted and amplified SODEGIRs are mostly localized on chromosomes 3 and 5, respectively (e.g. samples 50PC; Figure 4A) and genes contained in these regions are characterized by correlated levels of gene dosage and expression (Figure 4B).

A deeper analysis of SODEGIRs identified in the chromosomes of the various tumor samples (Caki-1, Astro and RCCp) allows defining four main chromosomal patterns (Figure 5):

(i) Deletion of an entire chromosome characterized by the combination of complete CN loss, LOH and GE down-regulation (Figure 5A);
(ii) Deletion of part of a chromosome affected by CN loss and GE down-regulation in absence of LOH (Figure 5B);
(iii) Amplification of an entire chromosome characterized by the combination of complete CN gain and GE up-regulation (Figure 5C);
(iv) Amplification of part of a chromosome affected by CN gain and GE up-regulation (Figure 5D).

As reported in the analysis of Caki-1, amplified SODEGIRs are not usually associated to LOH. However, its worthwhile noting that, although the GE status is normally associated to the CN one, there are chromosomal area where the CN gain does not impact the transcriptional activity (Figure 5C; q11.21–q21.3) or vice versa, the GE up-regulation is not associated to a corresponding CN gain (Figure 5D; p12–q23.1).

The aggregation of the single sample SODEGIRs allowed identifying unique SODEGIR signatures for the astrocytoma and renal carcinoma datasets. The Astro signature is composed by three amplified regions distributed along chromosome 7, containing genes such as EGFR, CAV1, MET and NOS3, and by two major deleted regions on chromosome 10, containing GATA3 and PTEN tumor suppressors (Figure 6, Table 2). Noticeably, a recurrent amplified block comprising 7q22.3–q31.2 is shared by nine patients and three patients are characterized by a common deleted SODEGIR on chromosome 10, spanning from q22.1 to the telomer. Moreover, patients can be grouped according to their SODEGIR pattern into those presenting both deletion

---

Figure 3. Visualization of SODEGIR results for the analysis of an Astro single sample (e.g. HT1139). (A) genome view and (B) genome box plot.
and amplification (3 samples), those having only the amplification (6 patients) and those without any SODEGIR (3 patients). Similarly, the RCC signature is composed of an amplified region on chromosome 5 and a deleted one on chromosome 3 (Figure 7, Table 2). The amplified SODEGIR, located at 5q21.3–q35.3 and shared by eight samples, contains APC and PDGFB oncogenes while the deleted SODEGIR (on 3p14.1–p22.3 in eleven samples) hosts the well-known FHIT fragile site and RASSF1 tumor suppressor gene.

**DISCUSSION**

The integration of multiple sources of information represents a promising approach to deepen the resolution and
Figure 6. Results of the aggregation of SODEGIRs in the analysis of the *Astro* dataset. (A) *q*_plot: The statistical significance for the aggregation of amplifications/deletions is displayed as *q*-value. Chromosome positions are indicated along the y-axis with the centromere positions identified by yellow dotted lines. Amplifications (red lines) and deletions (green lines) that are shared by a statistically relevant number of samples surpass the significance threshold (blue dotted line, *q*-value ≤ 0.05). (B) SDG chromosome view for chromosome 10 in all astrocytoma samples. (C) SDG chromosome view for chromosome 7 in all astrocytoma samples.

Table 2. Summary of the SODEGIR signatures for the astrocytoma and renal carcinoma datasets

<table>
<thead>
<tr>
<th>Chr</th>
<th>Cytoband</th>
<th>Start (Mb)</th>
<th>End (Mb)</th>
<th>Length (Mb)</th>
<th>No of genes</th>
<th>Relevant cancer genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Astro dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amplification signature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td><em>q</em>22.1–<em>q</em>11.22</td>
<td>4.7</td>
<td>70.9</td>
<td>66.2</td>
<td>244</td>
<td>EGFR, IL6, RAC1, SFRP4, IGFBP3, PMS2</td>
</tr>
<tr>
<td></td>
<td>q21.13–q35</td>
<td>89.6</td>
<td>144.0</td>
<td>54.4</td>
<td>344</td>
<td>CDK6, MAF, CAV2, CAV1, CASP2, FLNC, WNT16, WNT2, MET, PIK3CG, PON1, NOS3, CHRS, ABP1, RHEB</td>
</tr>
<tr>
<td></td>
<td>q36.1–q36.1</td>
<td>149.0</td>
<td>151.0</td>
<td>2.0</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deletion signature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10q15.1–10q12.2</td>
<td>4.9</td>
<td>23.6</td>
<td>18.8</td>
<td>82</td>
<td>GATA3, IL2Rα, IL15Rα, STAM, CACNB2, MLLT10, FRAT1, CASP7, CHUK, SARI1A, FAS, PTEN, BTRC, HK1, MMP21, CYP2C9, MGMT, SUFU, DBMT1, LGII, MXII</td>
</tr>
<tr>
<td></td>
<td>10q21.3–10q26.3</td>
<td>70.2</td>
<td>132.0</td>
<td>61.8</td>
<td>387</td>
<td>HK1, MMP21, CYP2C9, MGMT, SUFU, DBMT1, LGII, MXII</td>
</tr>
</tbody>
</table>

| **RCCp dataset** | | | | | | |
| | Amplification signature | | | | | |
| 5 | 5q21.1–5q21.2 | 99.9 | 103.0 | 3.1 | 9 | IL9, IL4, IL5, GM-CS, IL13, MCC, NPM1, FGFR4, SPINK1, APC, IRF1, ACSL6, CXCL14, PDGFB |
| | 5q21.3–5q35.3 | 108.0 | 179.0 | 71.0 | 419 | |
| | Deletion signature | | | | | |
| 3 | 3p22.3–3p14.1 | 35.7 | 65.3 | 29.6 | 279 | FHIT, MLH1, RASSF1, GPRX1, ARMET, CCR3, CCR2, CXCR6, CCR1, SAMA3F, PLXNB1, RHOA, SMARCC1, TLR9 |

Amplified and deleted SODEGIRs are described in terms of cytoband, chromosomal region, and total number of annotated genes. Values have been derived from SDGset_Table files of datasets, as deposited in CWS.
enhance the interpretation of gene dosage and expression profiles alone. This strategy can be generalized to identify and prioritize targets for functional studies which are expected to hasten the translation of basic research findings into clinical applications. The proof of principle of this approach comes from the study by Garraway et al., who, combining genome-wide, SNP-based CN maps and GE profiles, identified an amplified area containing the transcription factor MITF, a novel, potential tissue-specific oncogene (13,25). However, an efficient integration of GE profiling data with structural information requires appropriate datasets, i.e. paired GE and CN signals for the same sample and the development of computational approaches to overcome the limits of simple correlation. Although the number of studies combining CN and GE measurements has been constantly increasing since the development of high-throughput technologies (aCGH and SNP mapping arrays), still the availability of paired data sets with GE and CN from the same patient is limited. Moreover, the use of genomic and transcriptional arrays from the same manufacturer, in which probes are linked to precise chromosomal positions and are annotated in the same format, is crucial for the implementation of integrative methods. From a bioinformatics standpoint, the integration of CN and GE levels is mostly achieved using linear regression models and correlation coefficients between DNA CN and mRNA expression (5–12). Given the complex genomic environment of a tumor cell, this approach may result inefficient in capturing wide-range relationships between CN imbalances and GE changes. Instead of focusing on the local correlation between the two types of data, we developed a computational framework which directly integrates CN and GE profiles at genome-wide level, by statistically assessing the gene dosage and transcription statuses on common genomic positions. When applied to DNA/RNA paired data, this procedure allows the identification of SODEGIRs and the definition of tissue-specific SODEGIR signatures.

The method is based on estimating both CN and GE scores at the same chromosomal coordinate, e.g. the Entrez Gene physical position of a gene in base pairs. In general, CN data can be obtained from aCGH or

Figure 7. Results of the aggregation of SODEGIRs in the analysis of the RCCp dataset. (A) q_plot: The statistical significance for the aggregation of amplifications/deletions is displayed as q-value. Chromosome positions are indicated along the y-axis with the centromere positions identified by yellow dotted lines. Amplifications (red lines) and deletions (green lines) that are shared by a statistically relevant number of samples surpass the significance threshold (blue dotted line, q-value ≤ 0.05). (B) SDG chromosome view for chromosome 5 in all RCCp samples. (C) SDG chromosome view for chromosome 3 in all RCCp samples.
SNP microarrays using methods based on Hidden Markov Models or segmentation algorithms (18,26–32). When using aCGH technology, genes are directly interrogated by specific gene probes and therefore the gene CN is readily available for the same entity interrogated by the expression array. Instead, using SNP mapping arrays, the CN value refers to a SNP marker and the gene CN must be estimated. To date, only two computational procedures, i.e. dChip and FASeg, have been developed to calculate the gene CN directly from SNP mapping data (33). dChip infers the gene CN value by averaging the signals of the SNP probes annotated in the chromosomal region of the gene (31), while the fragment reduction algorithm of FASeg produces fitted-CN data which can be annotated at gene level (19). Differently from both dChip and FASeg, the SODEGIR approach uses a kernel regression estimator with automatically adapted local plug-in bandwidth to estimate both CN and GE at the same gene physical position from signals of SNP and transcripts probes. The estimation process is a non-parametric regression where the signal, acquired by a probe designed to interrogate a given chromosomal position, is estimated at another chromosomal coordinate using a smoothing function. Specifically, when estimating the regression function, lokerns transforms CN and GE values of 115 561 SNP and 41 192 expression probes, respectively, into CN and GE levels for 16 395 annotated Entrez genes through a local averaging of the observations. A major advantage of this kernel regression estimator is the possibility to automatically adapt the smoothing bandwidth to account for the non-uniform distribution and density of genes along the genome. As such, the method automatically sets the optimal bandwidth according to the underlying structure of the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations). The efficacy of the locally adaptive approach in estimating GE levels has been already shown in ref. (14), while its application to the CN signals (LSCN) allows detecting gene dosage on gene transcriptional activity, both at the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations). The efficacy of the locally adaptive approach in estimating GE levels has been already shown in ref. (14), while its application to the CN signals (LSCN) allows detecting gene dosage on gene transcriptional activity, both at the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations). The efficacy of the locally adaptive approach in estimating GE levels has been already shown in ref. (14), while its application to the CN signals (LSCN) allows detecting gene dosage on gene transcriptional activity, both at the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations). The efficacy of the locally adaptive approach in estimating GE levels has been already shown in ref. (14), while its application to the CN signals (LSCN) allows detecting gene dosage on gene transcriptional activity, both at the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations). The efficacy of the locally adaptive approach in estimating GE levels has been already shown in ref. (14), while its application to the CN signals (LSCN) allows detecting gene dosage on gene transcriptional activity, both at the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations) levels for 16 395 annotated Entrez genes through a local averaging of the observations. A major advantage of this kernel regression estimator is the possibility to automatically adapt the smoothing bandwidth to account for the non-uniform distribution and density of genes along the genome. As such, the method automatically sets the optimal bandwidth according to the underlying structure of the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations). The efficacy of the locally adaptive approach in estimating GE levels has been already shown in ref. (14), while its application to the CN signals (LSCN) allows detecting gene dosage on gene transcriptional activity, both at the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations). The efficacy of the locally adaptive approach in estimating GE levels has been already shown in ref. (14), while its application to the CN signals (LSCN) allows detecting gene dosage on gene transcriptional activity, both at the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations). The efficacy of the locally adaptive approach in estimating GE levels has been already shown in ref. (14), while its application to the CN signals (LSCN) allows detecting gene dosage on gene transcriptional activity, both at the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations). The efficacy of the locally adaptive approach in estimating GE levels has been already shown in ref. (14), while its application to the CN signals (LSCN) allows detecting gene dosage on gene transcriptional activity, both at the genome thus avoiding both too small bandwidths, which would lead to wiggly regression curves and noisy estimations, and too large ones, which could smooth away important details (i.e. CN spikes or small local variations).
whole-genome and at the chromosome levels. In accordance with (6), CN gain seems to have a stronger influence on regional transcriptional activity than CN loss. The fact that gene amplification greatly enhances GE while there are many aneuploidy-independent mechanisms leading to down-modulation of transcriptional activity (Figures 2C and D, 3B, and 4B) is supported by several evidences from mammalian cell lines and tumors (37,38) and should be taken into consideration when chromosomal instability is inferred from transcriptional profiles.

The SODEGIR approach is robust both with respect to the algorithm used to generate CN data and to the experimental design. Specifically, using the renal carcinoma dataset, the performance of LSCN was evaluated on CN values generated by CNAT and CNAG using paired normal specimens (i.e. the matched blood samples of the tumor tissues, $RCC_p$) and HapMap samples as the reference set ($RCC_h$). LSCN performed similarly irrespectively of the type of method used to generate the CN and of the type of experimental scheme. However, LOH likelihood calculation was more efficient using matched normal samples as reference (data not shown).

In conclusion, SODEGIR represents a bioinformatics procedure for the integrative, gene-position based analysis of CN and GE data that allows the identification of discrete chromosomal regions of coordinated DNA CN alterations and changes in transcriptional levels. These imbalanced regions may constitute a valuable resource for discovering novel diagnostic, prognostic, and therapeutic markers, although deciphering the mechanisms of transcriptional regulation of genes associated with chromosomal aberrations will likely require the integration of additional information (e.g. microRNA expression levels, fluorescence in situ hybridization, mutations, methylation, chromatin immunoprecipitation, post-transcriptional regulation) and the development of statistical approaches able to handle different types of genomic data (39).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Fondazione CARIPARO (Progetti Eccellenza 2006); MIUR (FIRB RBLA03ER38, FIRB Idee Progettuali RBIP064CRT and PRIN 2007Y84HTJ); University of Modena (Finanziamento Linee Strategiche di Sviluppo dell’Ateneo, Medicina Molecolare e Rigenerativa, 2008); Fondazione Cassa di Risparmio di Modena (Bando ricerca 2007) and University of Milano (funds to CISI and Department of Biomedical Science and Technologies); fellowship of the University of Padova (CPDR074285/07) (to F.F.). Funding for open access charge: Fondazione CARIPARO (Progetti Eccellenza 2006).

Conflict of interest statement. None declared.

REFERENCES


