Supplementary Information

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

Silvio Bicciato[1*], Roberta Spinelli[2], Mattia Zampieri[3], Eleonora Mangano[4], Francesco Ferrari[5], Luca Beltrame[2], Ingrid Cifola[2], Clelia Peano[2], Aldo Solari[6], Cristina Battaglia[4]

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

* Corresponding author

Silvio Bicciato
Dept. of Biomedical Sciences
University of Modena and Reggio Emilia
via G. Campi 287
41100 Modena, Italy
Email: silvio.bicciato@unimore.it
Tel.: +39-059-2055219
Fax.: +39-059-2055410

Contents

Genome wide microarray technology and array datasets

The proposed methodology has been applied to public and proprietary copy number and gene expression data obtained using Affymetrix Human Mapping 100K or 250K and HG-U133 Plus 2.0 arrays, respectively. The Affymetrix Human Mapping 100K set comprises a set of two arrays using Xba and Hind GeneChip® Mapping assays and provides the genotyping analysis of 116,204 SNPs, designed to map a total of 9,899 unique Entrez Gene IDs. The Human Mapping 250K Nsp array is part of the 500K set and allows monitoring approximately 262,000 SNPs mapped to a total of 10,514 unique Entrez Gene IDs. Affymetrix Human Genome U133 Plus 2.0 array allows the gene expression profiling of 47,401 human transcripts and variants, corresponding to 19,872 unique Entrez Gene IDs.
The computational procedure was tested using data derived from four major datasets:

- *Caki-1* (a tumor cell line);
- *Astro* (a collection of astrocytoma samples);
- *RCC* (renal carcinoma samples);
- *reference DNA* (normal individuals),

comprising a total of 263 Human Mapping arrays and 66 HG-U133 Plus 2.0 gene chips.

Caki-1: tumor cell line dataset

The *Caki-1* dataset represents the genomic and gene expression profiling of a tumor cell line and includes 2 mapping arrays (i.e. a complete 100K set) and 6 HG-U133 Plus 2.0 chips. The Caki-1 cell line is derived from the skin metastasis of a clear cell renal carcinoma, is included in the NCI-60 cell line collection, and is characterized by a nearly triploid chromosomal complement and a very complex karyotype (1,2). Gene expression profiling was carried out in triplicate hybridizing a commercial human reference RNA (*HRR*; Stratagene Universal Human Reference RNA) and the RNA extracted from Caki-1. Both genotyping and gene expression raw data are available on Array Express (E-MEXP-902 and E-MEXP-448). Additionally, Caki-1 DNA was also profiled using a Human Mapping 250K Nsp array (raw data are available upon request).

Astro: astrocytoma dataset

The *Astro* dataset comprises paired genotyping and gene expression data derived from 12 astrocytoma samples (HF0017, HF0108, HF0152, HF0491, HF0608, HF1139, HF1232, HF1269, HF1344, HF1442, HF1469, and HF1511) selected from the collection of (3). The dataset accounts for 24 mapping arrays (Human Mapping 100K) and 12 HG-U133 Plus 2.0 gene chips. In the gene expression analysis, the non-tumor population is represented by a set of 34 brain samples derived from epileptic (*NT Brain*: HF0088, HF0120, HF0131, HF0137, HF0151, HF0163, HF0171, HF0211, HF0232, HF0295, HF0303, HF0377, HF0383, HF0467, HF0512, HF0523, HF0526, HF0533, HF0593, and HF0616) and normal individuals (*NT Brain*: NT1, NT2, NT3, NT4, NT5, NT6, and NT7). All raw data have been downloaded from the Repository of Molecular Brain Neoplasia Data (https://caintegrator.nci.nih.gov/rembrandt/) and from GEO (GSE6109 and GSE4290).

RCC: clear cell renal carcinoma dataset

The *RCC* dataset is described in (4) and comprises genotyping and gene expression data of 12 clear cell renal carcinoma patients (28RA, 33BV, 36MML, 37BA, 27CG, 40RR, 45DM, 46SA, 47CA, 49CA, 50PC, and 51MI). The genomic DNA of the 12 RCC tissues and of the corresponding blood samples (*Blood*) were analyzed using 48 Human Mapping 100K arrays. Gene expression profiling was performed on the RNA samples obtained from the 12 RCC tissues and from 11 renal cortex samples (*Cortex*: 28RA, 32GM, 33BV, 35PA, 36MMl, 37BA, 40RR, 41SG, 44DE, 50PC, and 51MI) using 23 HG-U133 Plus 2.0 arrays. Patient

descriptions and raw data are available at Array Express (E-TABM-284/E-TABM-283 and E-TABM-282).

Reference DNA dataset

The genotyping data of two sets of normal individuals have been downloaded from the Affymetrix Data Resource Center (http://www.affymetrix.com/support/datasets.affx) and used as a common reference for copy number analysis and SODEGIR tuning. Specifically, the first reference set comprises 10 samples (*AffyRef*: NA17201, NA17202, NA17203, NA17204, NA17205, NA17206, NA17207, NA17208, NA17210, NA17211), randomly selected from the Mapping 100K CCNT Reference Data, while the second is a subset of the Mapping 100K HapMap Trio Dataset (*HapMap*: 30 male and 30 female CEU founders). The *HapMap* samples were also used as the reference set in all un-paired genotyping analyses. Additionally, a subset of the Mapping 250K HapMap Trio Dataset (*HapMap250*, 48 samples, Affymetrix Data Resource Center) was used as reference in the analysis of Caki-1 with the Human Mapping 250K Nsp array.

Copy number and LOH analysis

Chromosome Copy Number Analysis Tool 4.01 (CNAT 4.01, Affymetrix, 2007) and Copy Number Analyzer for GeneChip 2.0 (CNAG 2.0, (5)) were used to calculate SNP copy number (CN) and loss of heterozygosity (LOH) from mapping arrays genotyping data. Specifically, in all dataset, 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 where tumor tissue and blood pairs were available (RCC_p).

In details, CNAT 4.01 was applied to quantify the log2 copy number (Log2Ratio), using Gaussian smoothing at a fix bandwidth of 2 Mb, and the copy number state (CNState), using the Hidden Markov Model (HMM) Transition Decay at 1 Mb. The *Astro* dataset could not be processed using CNAT 4.01 because DTT (or CAB) files were not available. CNAG 2.0 was applied to all datasets to calculate the raw total log2 copy number (Log2Ratio_AB), selecting the averaging mode performed over 10 SNPs, with the exclusion of min and max values. The total copy number N_AB (i.e., the SNP copy number state) was determined through an HMM estimation and LOH events were described by the LOH likelihood. In the case of the *Caki-1* dataset, the Log2Ratio_AB values have been calculated for the 50K Xba, the 50K Hind, the combined 100K array set and the 250K Nsp. All CNAT and CNAG plots were visually inspected to identify regions affected by copy number gain and loss and regions of inferred LOH. Table 1_SI highlights the different data input, data output, and settings of CNAT 4.01 and CNAG 2.0.

Table 1_SI: Different data input and settings of CNAT 4.01 and CNAG 2.0 software.

	CNAT 4.01	**CNAG 2.0**
Data input	.CAB or .DTT/.CHP	.CEL and .CHP
Data output	Probe set, Chromosome, Position, Log2Ratio, HmmMedianLog2Ratio, CNState, NegLog10PValue	AffymetrixSNPsID, rsID, Chromosome, Position, Log2Ratio_AB, N_AB, Call_test, LOH_likelihood
Paired/un-paired analysis	yes/yes	yes/yes
CN analysis	Gaussian smoothing	Average best fit
LOH analysis	LOH likelihood	LOH likelihood
CN state	HMM Transition Decay	HMM
Annotation	NCBIv36.1 (March 2006)	NCBIv35 (May 2004)

The SODEGIR method

Overview

SODEGIR is a bioinformatics procedure that allows:

i) identifying genome-wide, concomitant alterations of copy number (CN) and gene expression (GE) in single samples;
ii) extending the integrative analysis to entire datasets.

These two issues are addressed in three steps composing the method (Figure 1_SI):

1. statistical estimation of copy number and transcriptional scores at common genomic positions;
2. identification of significant overlap of differentially expressed and genomic imbalanced regions (SODEGIR) on a single-sample basis;
3. aggregation of SODEGIRs from different samples to obtain global signatures of tumor types.

Step 1 stems from the Locally Adaptive Statistical procedure (LAP), a statistical approach for the identification of imbalances in regional gene expression (6). LAP is based on a kernel regression analysis which allows integrating high-throughput data with the structural information of a genome without assuming any particular gene distribution. This methodology is here extended to SNP copy number data, with the aim to estimate CN values at gene positions and detect alterations of regional copy number.

Step 2 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 gene CN and GE statuses are concordant at a given statistical threshold, in a single sample. In particular, we define a SODEGIR as *deleted* (status 1) when the CN status is *loss* and the GE one is *down-regulation*, as *amplified* (status 3) when the CN is *gain* and the GE is *up-regulation*, and as *unchanged* (status 2) when either CN or GE are neutral.

Step 3 elevates the analysis from the single to the multiple-sample level, statistically combining the various single-sample SODEGIRs to assess a unique SODEGIR signature of the entire dataset.

The entire procedure is coded in several R functions which depend on R and Bioconductor packages and are available at the companion web site (http://www.xlab.unimo.it/SODEGIR/).

Input and output data

The inputs to the SODEGIR methods are as follows:

- a **.cn** file for each sample containing CN data either in terms of absolute values or log2 ratios for SNP probe sets. The .cn file is normally a data table in tab-delimited format, containing different columns depending on the output settings of the originating software. As an example, a CN data matrix obtained from CNAT 4.01 contains the Log2Ratio column while a .cn file derived from CNAG 2.0 has Log2Ratio_AB column as identifiers of the CN values for each SNP (Table 1_SI);
- a **.ge** file for each sample representing the gene expression data in terms of absolute values for transcript probe sets. Specifically, for Affymetrix arrays the .ge file is a .CEL file whose probe level data are converted to expression values using robust multi-array average procedure (RMA) (7);
- a **dna_annotation** file with all information about the SNP probes contained in the mapping array. The dna_annotation file allows annotating each probe identifier in the array in terms of chromosome and chromosomal position (in bp). In the case of multiple SNP arrays, the file is constructed merging the information contained in NetAffx single array annotation files (e.g., http://www.affymetrix.com/support/technical/byproduct.affx?product=100k).

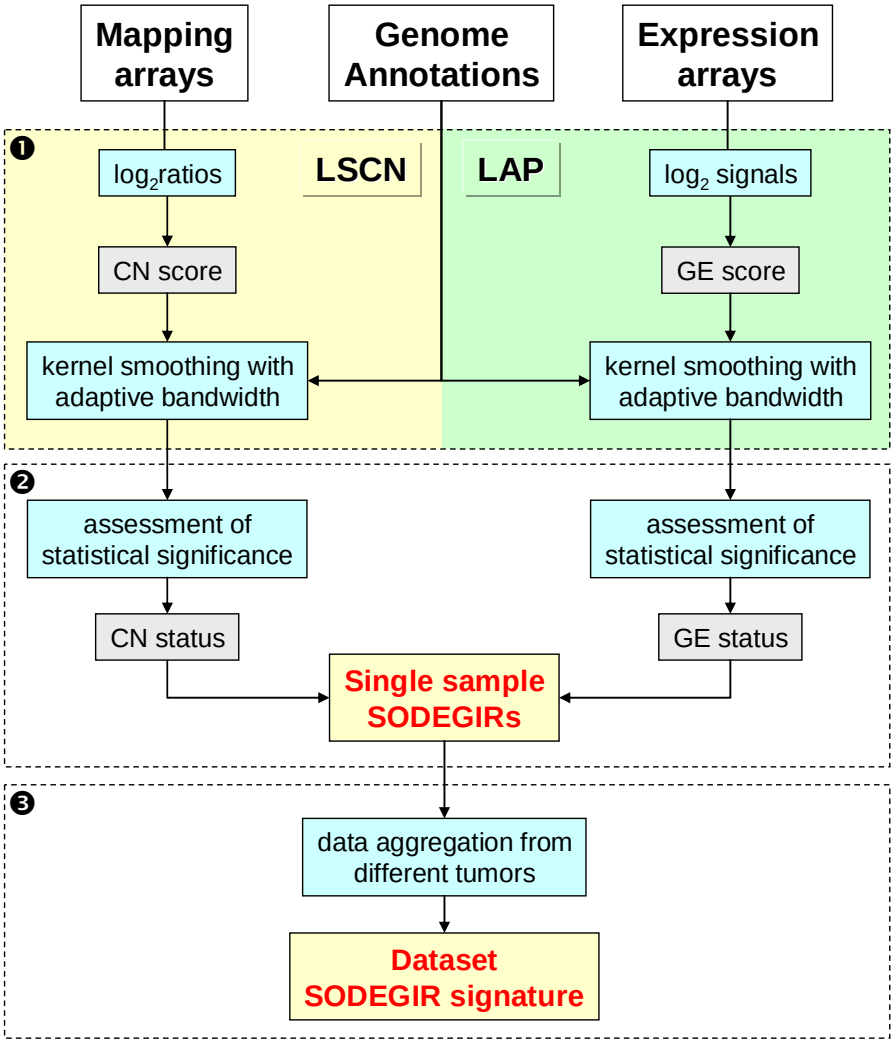

Figure 1_SI: The workflow of the SODEGIR procedure.

- an **rna_annotation** file with all the annotation about probe sets contained in the GE array. The rna_annotation file links each probe identifier in the array with Entrez Gene ID, gene symbol and description, chromosome, chromosomal position (in bp), strand, and cytoband. In the case of Affymetrix arrays, the file has been constructed using the Bioconductor annotation data package *hgu133plus2*. Probe sets without a unique chromosomal position and those referring to the X and Y chromosomes were filtered out. This re-annotation step resulted in the selection of 41,192 probe set IDs.
- a **sample_info** file which contains .cn and .ge filenames, the sample names and the sample phenotype (e.g., normal, tumor) to be used along the analysis. The sample_info file can be supplemented with any additional information that can be included in the R AnnotatedDataFrame class representing the sample;
- a **parameters** file containing default values for a series of parameters, as the position of the ID column, the position of CN data column, and the number of rows to skip (e.g. rows containing comments and/or column labels) in the .cn files, the number B of permutations, the type of CN data (absolute or log2ratios), and the threshold limit for the genome median (see the section on CN score).

Four different files for each sample constitute the outputs of the SODEGIR analysis:

- a **.lscn** file, i.e. a data table in .csv format, containing gene CN score, q-value, and complete annotations (Entrez Gene ID, chromosome, position, cytoband, strand) for any gene of the rna_annotation file;
- a **.lap** file, i.e. a data table in .csv format, containing GE scores, q-value, single probe and regional fold-changes, and complete annotations (Entrez Gene ID, chromosome, position, cytoband, strand) for any gene of the rna_annotation file;
- a **.SDG** file that combines all the information contained in the .lscn and .lap files, plus reports the CN, GE and SODEGIR statuses at given q-value and score thresholds for any gene of the rna_annotation file;
- a **.SDG_table** file specifying the characteristics of all CN, GE and SODEGIR clusters (i.e. gain and loss, up- and down-regulated, deleted and amplified regions) in terms of chromosome, start, end, length, cytoband, number of associated genes and SNPs and symbols of genes contained in the region.

Moreover, the SODEGIR procedure outputs four files for each dataset:

- a **.cnSDGset** file, i.e. a data table in tab-delimited format, containing the CN status of all samples (as derived from the various .sdg files), the p- and q-values of genes with CN status shared in a statistically relevant number of samples (common CN signature), and complete annotations (Entrez Gene ID, chromosome, position, cytoband, strand) for any gene as reported in the rna_annotation file;
- a **.geSDGset** file, i.e. a data table in tab-delimited format, containing the GE status of all samples (as derived from the various .sdg files), the p- and q-values of genes with GE status shared in a statistically relevant number of samples (common GE signature), and complete annotations (Entrez Gene ID, chromosome, position, cytoband, strand) for any gene of the rna_annotation file;
- a **.SDGset** file, i.e. a data table in tab-delimited format, containing the SODEGIR status of all samples (as derived from the various .sdg files), the p- and q-values of genes with SODEGIR status shared in a statistically relevant number of samples (dataset SODEGIR signature), and complete annotations (Entrez Gene ID, chromosome, position, cytoband, strand) for any gene of the rna_annotation file;
- a **.SDGset_table** file specifying the characteristics of all CN, GE and SODEGIR clusters (i.e. gain and loss, up- and down-regulated, deleted and amplified regions) shared, at a given q-value threshold, in a statistically relevant number of samples. Clusters are annotated in terms of chromosome, start, end, length, cytoband, number of associated genes and SNPs and symbols of genes contained in the region.

Finally, SODEGIR results are visualized at various levels of detail. In particular, the outputs can be represented in:

- a **genome view** where regions of CN gain/loss, GE up-/down-regulation and *deleted* (CN loss and GE down-regulation) and *amplified* (CN gain and GE up-regulation) SODEGIRs 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;
- a **chromosome view** displaying CN (N_AB) and LOH statuses as estimated by the CNAG HMM on each SNP probe, CN, GE, and SODEGIR statuses as determined by the SODEGIR procedure on gene positions for a given chromosome in a single sample . The grey bars indicate SNP probes (in N_AB and LOH lanes) or Entrez Gene ID positions (in CN, GE and SODEGIR lanes). Red and green bars in the N_AB lane indicate N_AB greater than 3 and less than 1, respectively. Blue bars in the LOH lane highlight SNP probes with an inferred LOH value greater than 20. Green bars in CN, GE, and

SODEGIR lanes indicate loss, down-regulation, or deletion (i.e. a status of 1). Red bars in CN, GE, and SODEGIR lanes indicate gain, up-regulation, or amplification (i.e. a status of 3).

- **a SDG chromosome view** which highlights the SODEGIRs on a given chromosome in all samples of a dataset. The grey bars indicate Entrez Gene ID positions and the color scheme is the same as in the SODEGIR lane of a single sample chromosome view (i.e. green bars for status 1, red bars for status 3);
- a **boxplot** of CN and GE relative levels in SODEGIRs. This plot allows quantifying the impact of gene copy number on global gene expression levels in the entire genome (genome boxplot) or, eventually, in a single chromosome (chromosome boxplot). CN levels are categorized into 5 bins highlighting 2 ranges of loss (green boxes, gene CN score less than -0.1), 1 range of diploidy (white box, gene CN score between -0.1 and 0.1) and 2 ranges of gain (red boxes, gene CN score greater than 0.1). The GE values in the y-axis correspond to GE scores;
- a **q_plot** reporting the aggregation of CN, GE and SODEGIR results for the analysis of an entire dataset. The statistical significance for the aggregation of gains/losses, up/down-regulations and amplifications/deletions is displayed as q-value. Chromosome positions are indicated along the y-axis with the centromere positions identified by yellow dotted lines. Gains, up-regulations and amplifications (red lines) and losses, down-regulations 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).

Step 1

The first step transforms SNP copy number and gene expression data into CN and GE scores and integrates them with structural information (i.e., chromosomal coordinate), using a kernel regression estimator with adaptive bandwidth. Resembling LAP (6), 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 on SNP copy number and on gene expression data. In the case the input are only SNP copy number data, the procedure is named *Lokern Smoothing Copy Number* (*LSCN*), while when only gene expression signals are available the procedure is a revised version of LAP.

CN score

In the LSCN part of the procedure, CN data are transformed into a score $\Delta N_{i,j}^{SNP}$ 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 copy number 2 everywhere (8-10), the CN value of the diploid status is not set to 2 (i.e., log2ratio=0), but is estimated from the median CN calculated over all *i* SNPs. Although not equal to zero, the CN medians calculated over all samples of an entire dataset are nevertheless tightly distributed around zero (CN=2), irrespectively that the dataset represents normal or pathological samples (Figure 2_SI). As such, the CN score $\Delta N_{i,j}^{SNP}$ can be defined as follows:

$$\Delta N_{i,j}^{SNP} = N_{i,j}^{SNP} - \min\left(\widetilde{N}_{j}^{SNP}, thrN\right) \quad (1)$$

where $N_{i,j}^{SNP}$ is the copy number of SNP *i* in sample *j*, $\widetilde{N}_{j}^{SNP}$ is the median CN calculated over all the *i* SNP probes of array *j* and $thrN$=0.05 is an upper threshold to cope with potential outlying samples (Figure 2_SI).

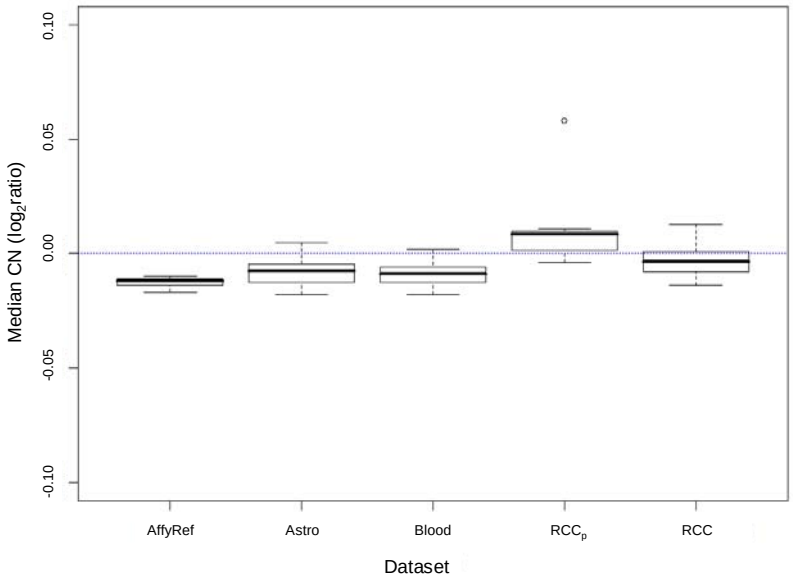

Figure 2_SI: Box plot of CN medians for all samples of all considered datasets. The upper dot in the *RCC*$_p$ dataset represents an outlying sample.

GE score

In its original version (6), LAP calculates a statistic for ranking probes in order of strength of differential expression in two or more populations. Specifically, given a matrix **X** of normalized expression levels x_{ij} for gene i in sample j and a response vector **Y** for n samples, the score ΔE_i^{probe} is defined according to Tusher et al. (11):

$$\Delta E_i^{probe} = \frac{r_i}{s_i + s_0} \tag{2}$$

where the quantities r_i (i.e., the change in gene expression) and s_i (i.e. the standard deviation of the data) of each probe set i assume different formulations in different experimental designs (e.g., two- and multi-class problems, paired data, quantitative responses, time course experiments, survival analyses) and the estimates of gene-specific variance over repeated measurements are stabilized by a factor s_0 (see Tusher et al. 2001 and SAM technical manual for details).

Considering the analysis of a single sample j from a population of m pathological samples with normalized expression level x_{ij} for probe set i and a population of n normal specimens with average gene expression $\bar{x}_i^{norm}$, the GE score $\Delta E_{i,j}^{probe}$ can be defined as:

$$\Delta E_{i,j}^{probe} = \frac{x_{i,j} - \bar{x}_i^{norm}}{s_i + s_0} \tag{3}$$

where the standard deviation s_i for each probe set i is estimated using all pathological and normal samples:

$$s_i = \left\{ a \left[\sum_{j=1}^{m} \left(x_{i,j} - \bar{x}_i^{patol} \right)^2 + \sum_{k=1}^{n} \left(x_{i,k} - \bar{x}_i^{norm} \right)^2 \right] \right\}^{1/2}$$
$$a = \frac{m+n}{m \cdot n} \cdot \frac{1}{m+n-2} \quad (4)$$

Estimation of scores at gene positions: *lokern* smoothing

CN and GE scores are estimated at gene positions integrating probe set data and structural information using a kernel regression estimator with automatically adapted local plug-in bandwidth. As described in (6,12), 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 bp. 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. Non-parametric regression problems can be approached using various methods, as kernel smoothing, orthogonal series, spline functions or wavelets. A critical issue in selecting the regression strategy is represented by the procedure for adapting the smoothing parameters (e.g., the bandwidth) which can be adapted globally or locally (13). CN and GE scores are integrated with structural information using the *lokern* set of functions adapted from the Gasser-Müller type estimator (13,14) which is available as part of the *lokern* package (http://cran.r-project.org/web/packages/lokern/index.html).
Specifically, for each sample *j*, the regression model specifies:

$$\Delta N_{i,j}^{SNP} = \eta_j(Mb_i) + \epsilon_{i,j}$$
$$\Delta E_{i,j}^{probe} = \tau_j(Mb_i) + \varepsilon_{i,j} \quad (5)$$

where Mb_i is the physical position of SNP (probe) i, $\eta_j(Mb_i)$ and $\tau_j(Mb_i)$ are arbitrary functions of Mb_i, and $\epsilon_{i,j}$ and $\varepsilon_{i,j}$ 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 $\Delta N_{i,j}^{SNP}$ ($\Delta E_{i,j}^{probe}$) on the physical position Mb_i, is left as an arbitrary function $\eta_j(Mb_i)$ (or $\tau_j(Mb_i)$), while the random part is specified by assuming that the error components are uncorrelated with zero mean and constant variance. Considering for instance copy number values, the kernel regression model takes as input the pairs (Mb_i, $\Delta N_{i,j}^{SNP}$ with i=1, ..., L), estimates $\mathbb{E}\left(\Delta N_{i,j}^{SNP}\right) = \eta_j(Mb_i)$ by extracting a curve from the data, and returns the values $\Delta N_{g,j}^{gene}$ of $\eta_j\left(Mb_g\right)$ at given design points Mb_g (e.g., the g physical position of Entrez Genes). Similarly for gene expression levels, the input pairs (Mb_i, $\Delta E_{i,j}^{probe}$ with i=1, ..., P) are used for estimating $\mathbb{E}\left(\Delta E_{i,j}^{probe}\right) = \tau_j(Mb_i)$ and returning the values $\Delta E_{g,j}^{gene}$ of $\tau_j\left(Mb_g\right)$ at, e.g. the g physical position of Entrez Genes.
The input positions of L=115,561 SNPs and M=41,192 expression probes are derived from the Human Mapping 100K NetAffx Annotation Files and from the HG-U133 Plus 2.0 Bioconductor annotation library, respectively (see dna_annotation and rna_annotation files, as described in the Input and Output Data section). The output vector contains estimates of CN and GE scores for G=16,395 annotated Entrez Genes, obtained filtering out genes in chromosomes X and Y. It's worthwhile noting that, in the case of multiple probes mapping to the same locus (probe set redundancy in Affymetrix GE arrays), the average GE score of multiple probes mapping to the same physical coordinate has been assigned to the

chromosomal position (6). CN and GE scores at Entrez Gene positions (i.e., $\Delta N_{g,j}^{gene}$ and $\Delta E_{g,j}^{gene}$ for g=1, …, G) have been obtained using a kernel regression estimator with adaptive smoothing bandwidth (13). The adaptive smoothing bandwidth accounts for the non-uniform distribution and density of genes along the genome. Briefly, the smoothing function performs a local averaging of the observations when estimating the regression function. Crucial for the kernel regression estimation is the choice of the smoothing bandwidth, since too small bandwidths will lead to a wiggly curve, while too large ones will smooth away important details. The *lokern* package contains functions that calculate the regression with an automatically chosen local (*lokerns*) or global (*glkerns*) bandwidth. The efficacy of the locally adaptive approach (i.e. the *lokerns* function) in smoothing GE scores has been already shown in (6), while its performance with CN scores are shown in Figure 3_SI. As shown in the various panels, both the local (red line) and the global (green line) approaches perform an efficient smoothing of the CN scores (black dots) and allow detecting broad as well as subtle changes.

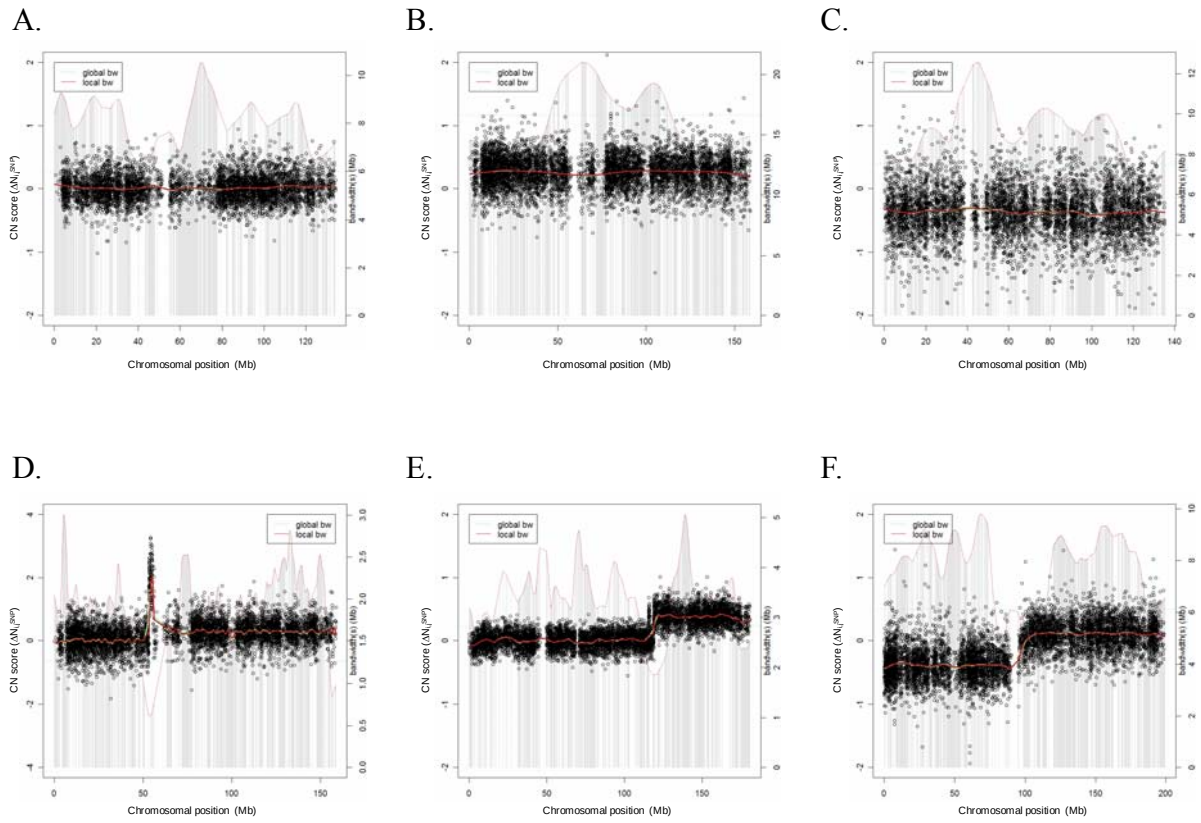

Figure 3_SI: Smoothing of CN scores $\Delta N_{i,j}^{SNP}$ (black dots) using kernel regression estimator with adaptive local (red line) and global (green line) bandwidth. Bandwidth amplitude is shown as gray bars (right y-axis). A. copy number neutral (chromosome 11 in *AffyRef* NA17203); B. copy number gain of an entire chromosome (chromosome 7 in RCC_p 27CG); C. copy number loss of an entire chromosome (chromosome 10 in *Astro* HF1232); D. copy number spike in the p-arm and gain of the entire q-arm (chromosome 7 in *Astro* HF1232); E. copy number gain of part of the q-arm (chromosome 5 in RCC_p 50PC); F. copy number loss of the entire p-arm and gain of the q-arm (chromosome 3 in RCC_p 27CG).

Moreover, both *lokerns* and *glkerns* regress 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 (Figure 4_SI). Thus, consistently with GE analysis, the locally adaptive approach (i.e. the *lokerns* function) has been applied also to regress CN scores in the LSCN part of the procedure depicted in Figure 1_SI.

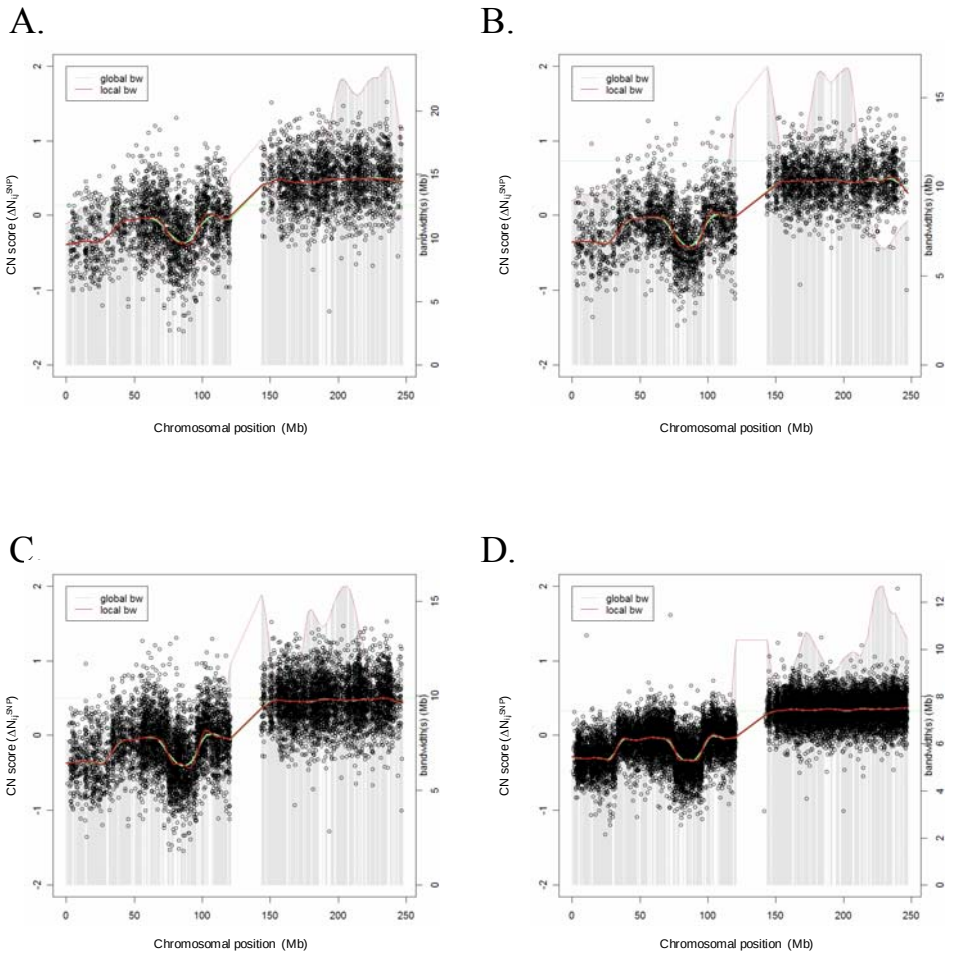

Figure 4_SI: Smoothing of CN scores $\Delta N_{i,j}^{SNP}$ (black dots) in chromosome 1 of Caki-1 using array with different SNP densities. The kernel regression estimator with adaptive local and global bandwidth is represented by red and green lines, respectively. Bandwidth amplitude is shown as gray bars (right y-axis). A. CN scores from 50K Hind GeneChip® Mapping assay; B. CN scores from 50K Xba GeneChip® Mapping assay; C. CN scores from Human Mapping 100K set; D. CN scores from Human Mapping 250K Nsp array.

Step 2

In the second step, the goal is to assess the statistical significance of copy number and gene expression variations and define regions with concomitant alterations of gene CN and GE in single samples. The procedure locally computes the significance levels (i.e., p-values and q-values) through a permutation scheme and estimates the CN and GE statuses of annotated genes. Finally, SODEGIRs are defined based on the copy number and transcriptional statuses.

Assessment of statistical significance

The scope is to make inferences about $\eta_j(Mb_g)$ (or $\tau_j(Mb_i)$) at each position g by testing the significance of a departure from the null form of $\eta_j(Mb_g)$ (or $\tau_j(Mb_i)$) corresponding to no

alterations of copy number (gene expression). This corresponds to test the following multiple hypotheses, for CN and GE respectively:

$$\begin{aligned} &H_{g,j}^{N}: \eta_j(Mb_g) = 0 \\ &K_{g,j}^{N}: \eta_j(Mb_g) \neq 0 \end{aligned} \quad g=1, \ldots, G$$

$$\begin{aligned} &H_{g,j}^{E}: \tau_j(Mb_g) = 0 \\ &K_{g,j}^{E}: \tau_j(Mb_g) \neq 0 \end{aligned} \quad g=1, \ldots, G \tag{6}$$

When no alterations of copy number (gene expression) are present along the genome, i.e. when $\bigcap_{g=1}^{G} H_{g,j}^{N}$ ($\bigcap_{g=1}^{G} H_{g,j}^{E}$) is true, the observed data values $\Delta N_{i,j}^{SNP} = \epsilon_{i,j}$ ($\Delta N_{i,j}^{probe} = \varepsilon_{i,j}$) are i.i.d. realizations and thus are exchangeable:

$$\begin{aligned} &\left(\Delta \mathrm{N}_1^{\mathrm{SNP}}, \ldots, \Delta \mathrm{N}_{\mathrm{L}}^{\mathrm{SNP}}\right) \stackrel{d}{=} \left(\Delta \mathrm{N}_{\pi(1)}^{\mathrm{SNP}}, \ldots, \Delta \mathrm{N}_{\pi(\mathrm{L})}^{\mathrm{SNP}}\right) \\ &\left(\Delta \mathrm{E}_1^{\mathrm{probe}}, \ldots, \Delta \mathrm{E}_{\mathrm{P}}^{\mathrm{probe}}\right) \stackrel{d}{=} \left(\Delta \mathrm{E}_{\pi(1)}^{\mathrm{probe}}, \ldots, \Delta \mathrm{E}_{\pi(\mathrm{P})}^{\mathrm{probe}}\right) \end{aligned} \tag{7}$$

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 $\stackrel{d}{=}$ 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, $\Delta N_{i,j}^{SNP}$ and $\Delta E_{i,j}^{probe}$ scores are randomly assigned to chromosomal locations and $\Delta N_{g,j}^{gene}$ and $\Delta E_{g,j}^{gene}$ re-estimated using the *lokerns* function (permuted scores $\Delta N_{g,j}^{gene,b}$ and $\Delta E_{g,j}^{gene,b}$). 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 $H_{g,j}^{N}$ and $H_{g,j}^{E}$ on the estimated scores $\Delta N_{g,j}^{gene}$ and $\Delta E_{g,j}^{gene}$ as test statistic, respectively. The significance $p_{g,j}^{N}$ (or $p_{g,j}^{E}$) that the expected score $\Delta N_{g,j}^{gene,b}$ (or $\Delta E_{g,j}^{gene,b}$) exceeds the observed one $\Delta N_{g,j}^{gene}$ (or $\Delta E_{g,j}^{gene}$), over B permutations, can be then computed as follows :

$$\begin{aligned} p_{g,j}^{N} &= \frac{\sum_{b=1}^{B} I\left\{\left|\Delta N_{g,j}^{gene,b}\right| \geq \left|\Delta N_{g,j}^{gene}\right|\right\}}{B} \\ p_{g,j}^{E} &= \frac{\sum_{b=1}^{B} I\left\{\left|\Delta E_{g,j}^{gene,b}\right| \geq \left|\Delta E_{g,j}^{gene}\right|\right\}}{B} \end{aligned} \tag{8}$$

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 (Figure 5_SI), the q-value is used to assign a measure of significance to each of

many tests performed simultaneously. Q-values $q^N_{g,j}$ and $q^E_{g,j}$ are estimated using R *qvalue* package (http://genomics.princeton.edu/storeylab/qvalue/).

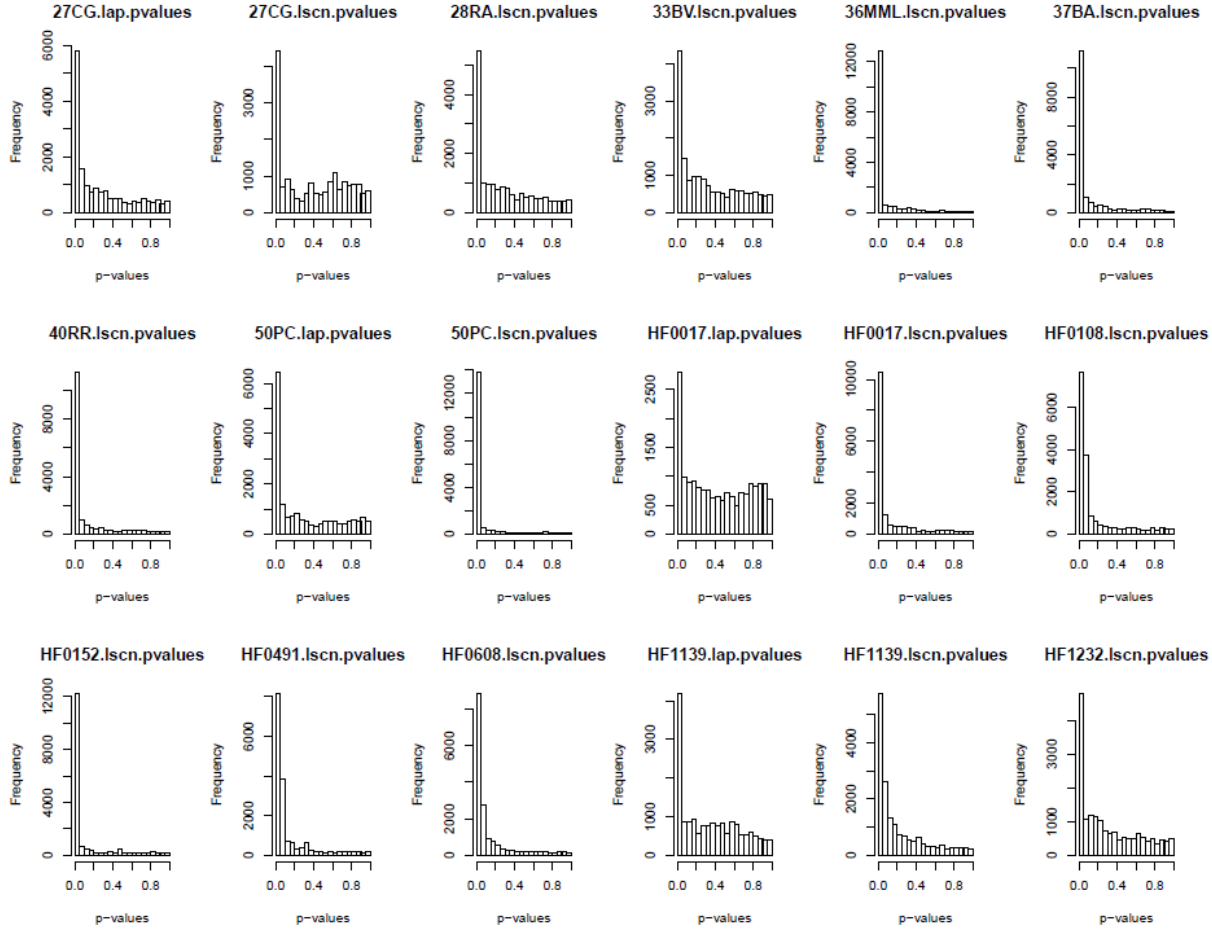

Figure 5_SI: Distribution of p-values $p^N_{g,j}$ (calculated by LSCN) and $p^E_{g,j}$ (calculated by LAP) for RCC samples and of p-values $p^N_{g,j}$ for astrocytoma samples.

Status quantification and SODEGIR definition

When the null hypothesis $H^N_{g,j}$ ($H^E_{g,j}$) is rejected, the copy number (or gene expression) status of a gene g in a sample j is decided basing on whether $\eta_j(Mb_g)$ (or $\tau_j(Mb_g)$) is smaller or greater than zero. The two-sided hypotheses of Eq. 6 are equivalent to the simultaneous testing of the following pair of one-sided hypotheses:

$$\begin{aligned} &H^{Ngain}_{g,j}: \eta_j(Mb_g) \geq 0 \quad against \quad K^{Nloss}_{g,j}: \eta_j(Mb_g) < 0 \\ &H^{Nloss}_{g,j}: \eta_j(Mb_g) \leq 0 \quad against \quad K^{Ngain}_{g,j}: \eta_j(Mb_g) > 0 \\ &H^{Eup}_{g,j}: \tau_j(Mb_g) \geq 0 \quad against \quad K^{Edown}_{g,j}: \tau_j(Mb_g) < 0 \\ &H^{Edown}_{g,j}: \tau_j(Mb_g) \leq 0 \quad against \quad K^{Eup}_{g,j}: \tau_j(Mb_g) > 0 \end{aligned} \tag{9}$$

Considering for instance copy number, the rejection of either $H^{Ngain}_{g,j}$ or $H^{Nloss}_{g,j}$ is equivalent to the rejection of $H^N_{g,j}$. Although Eq. 6 and Eq. 9 are equivalent ways of formulating the

same hypothesis testing problem, there is some advantage in using the formulation of Eq. 9. 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, $K_{g,j}^{Nloss}$ or $K_{g,j}^{Ngain}$ (and $K_{g,j}^{Edown}$ or $K_{g,j}^{Eup}$) can be associated with the two courses of action.

In particular, the null hypothesis $H_{g,j}^{N}$ ($H_{g,j}^{E}$) is rejected according to thresholds on the q-value (i.e., $thrq^{N}$ and $thrq^{E}$) and on the scores (i.e., $thr\Delta N$ and $thr\Delta E$). The q-value and score thresholds for CN and GE may be set to different values, depending on the desired stringency of the analysis. In this case $thrq^{N}$ and $thrq^{E}$ have been set to 0 and 0.05, respectively. For both CN and GE, low (e.g., $thr\Delta N_{low}$ and $thr\Delta E_{low}$) and high (e.g., $thr\Delta N_{high}$ and $thr\Delta E_{high}$) score thresholds have been defined from the distribution of the estimated scores $\Delta N_{g,j}^{gene}$ and $\Delta E_{g,j}^{gene}$. Specifically, the 10th and 90th quantile of $\Delta N_{g,j}^{gene}$ and $\Delta E_{g,j}^{gene}$ distributions have been selected as low and high score thresholds, respectively. The CN q-value and score thresholds have been optimized based on the analysis of the *AffyRef* Reference DNA dataset (data not shown), $\Delta E_{g,j}^{gene}$ thresholds have been selected according to the criteria used for the CN ones, and $thrq^{E}$ has been set to the value commonly used with gene expression data (6).

As such, CN and GE statuses are coded as (Table 2_SI):

- 1 (CN loss, GE down-regulation) when the q-value is below the q-value threshold and the score is smaller than the low score threshold, i.e. $K_{g,j}^{Nloss}$ ($K_{g,j}^{Edown}$) is true;
- 3 when the q-value is below the q-value threshold and the score is larger than the high score threshold (CN gain, GE up-regulation) i.e. $K_{g,j}^{Ngain}$ ($K_{g,j}^{Eup}$) is true;
- 2 (CN and GE neutral) in all other cases.

Table 2_SI: Quantification of CN and GE statuses based on q-value and score thresholds.

Status	**CN**		**GE**	
	$q_{g,j}^{N}$	$\Delta N_{g,j}^{gene}$	$q_{g,j}^{E}$	$\Delta E_{g,j}^{gene}$
1	*=0*	*≤quantile($\Delta N_{g,j}^{gene}$,0.1)*	*≤0.05*	*≤quantile($\Delta E_{g,j}^{gene}$,0.1)*
3	=0	*≥quantile($\Delta N_{g,j}^{gene}$,0.9)*	*≤0.05*	*≥quantile($\Delta E_{g,j}^{gene}$,0.9)*

Given the quantification of CN and GE statuses in a single sample, a significant overlap of differentially expressed and genomic imbalanced regions (SODEGIR) corresponds to a region of the genome where the CN and GE statuses are concordant. 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). The identification of the SODEGIRs corresponds to test the following hypothesis pair:

$$\begin{aligned} &H_{g,j}^{amp}: H_{g,j}^{Ngain} \cup H_{g,j}^{Eup} \quad against \quad K_{g,j}^{del}: K_{g,j}^{Nloss} \cap K_{g,j}^{Edown} \\ &H_{g,j}^{del}: H_{g,j}^{Nloss} \cup H_{g,j}^{Edown} \quad against \quad K_{g,j}^{amp}: K_{g,j}^{Ngain} \cap K_{g,j}^{Eup} \end{aligned} \tag{10}$$

In this case either $K_{g,j}^{del}$ or $K_{g,j}^{ampl}$ can be true and each testing problem is specified with the intersection–union formulation (16). As a consequence, the rejection of $H_{g,j}^{ampl}$ ($H_{g,j}^{del}$) and the acceptance of $K_{g,j}^{del}$ ($K_{g,j}^{amp}$) is accomplished only if both $H_{g,j}^{Ngain}$ and $H_{g,j}^{Eup}$ ($H_{g,j}^{Nloss}$ and $H_{g,j}^{Edown}$) are rejected, i.e. the SODEGIR status is equal to 1 (3) if both CN and GE statuses are equal to 1 (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 common SODEGIR signature across an entire dataset. Specifically, SOGEDIRs from all single sample analyses are aggregated to generate summary scores for amplifications and deletions using a binomial distribution test and the q-value to correct for multiple hypothesis testing.

Aggregation of single sample SODEGIRs

The procedure aims at determining where the regions of deletion and amplification shared by multiple samples are located and how likely it is that an observed shared region is due to chance. To this end, let $\boldsymbol{S_{g,j}}$ be the SODEGIR status for sample j at the gene g and assume that $\boldsymbol{S_{g,j}}$ follows a multinomial distribution with $\boldsymbol{Pr(S_{g,j}=1)=\theta_g^1}$, $\boldsymbol{Pr(S_{g,j}=2)=\theta_g^2}$, $\boldsymbol{Pr(S_{g,j}=3)=\theta_g^3}$, and $\boldsymbol{\theta_g^1+\theta_g^2+\theta_g^3=1}$. Under the null hypothesis that there are no real imbalanced regions, these probabilities are independent from g, i.e. $\boldsymbol{\theta_g^s=\theta^s, s=1,2,3}$. Then for each gene g, the following hypotheses are tested:

$$\begin{aligned} &H_g^1: \theta_g^1 = \theta^1 \ \ against \ \ \ K_g^1: \theta_g^1 > \theta^1 \\ &H_g^3: \theta_g^3 = \theta^3 \ \ against \ \ \ K_g^3: \theta_g^3 > \theta^3 \end{aligned} \tag{11}$$

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

$$p_g^s = Pr(T_g \geq t_g) = \sum_{r=t_g}^{J} \binom{J}{r} (\hat{\theta}^s)^r (1-\hat{\theta}^s)^{J-r} \tag{12}$$

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 summary score for deletions or amplifications. The same statistical approach has been used to aggregate, at dataset level, CN or GE statuses alone, thus computing dataset scores for the genomic regions with CN gain or loss, or with up-/down-regulation.

Supplementary results

All results from the single sample and the aggregation analyses of *Caki-1*, *Astro*, *RCC*, *RCC*$_p$, and reference DNA datasets are available at the web companion site http://www.xlab.unimo.it/SODEGIR/. Specifically, for any single sample, the supplementary files include:

- the characteristics of all CN, GE and SODEGIR clusters (*.SDG_table*);
- the boxplots of CN and GE relative levels in SODEGIRs (*.boxplot*);
- chromosome views displaying 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 for a given chromosome (*.chr.view*);

- genome views where 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 (*.genome.view*).

Moreover, for the *Astro* and *RCC* datasets, the supplementary files report:

- the physical characteristics of LOH regions as estimated by the CNAG HMM in all samples (*.LOH_table*)
- the characteristics of all CN, GE and SODEGIR clusters shared, at a given q-value threshold, in a statistically relevant number of samples (*.SDGset_table*)
- chromosome views which highlight LOH regions and SODEGIRs on a given chromosome in all samples of a dataset (*.chr.LOH_view* and *.chr.SDG_view*),
- q plots reporting the aggregation of CN, GE and SODEGIR results for the analysis of the entire dataset (*.q_plot.CN*, *.q_plot.GE*, and *.q_plot.SODEGIR*).

Simulation analysis

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. For sake of simplicity, all possible statuses of a gene (1, 2 or 3) have been summarized into a binary classification, where the neutral status is indicated by 2 while ≠ 2 denotes an altered condition. The performance of the gene discovery is best seen in a simple two-by-two table, where the genes are classified according to their true status and the test result (Table 3_SI).

Table 3_SI: Contingency table for assessing the performance on simulated data.

		True status	
		≠ 2	**2**
Test result	**≠ 2**	TP	FP
	2	FN	TN

In particular, the contingency table allows computing the following elements:

1. TP (true positives), i.e. the number of altered genes correctly identified as altered;
2. FP (false positives), i.e. the number of neutral genes wrongly identified as altered;
3. TN (true negatives), i.e. the number of neutral genes correctly identified as neutral;
4. FN (false negatives), i.e. the number of altered genes wrongly identified as neutral;
5. $FDR = FP/(TP + FP)$, i.e. the proportion of false positives among the genes identified as altered. This notation slightly departs from that of Benjamini and Hochberg (1995) because here we use FDR to denote the realized false discovery rate. What Benjamini and Hochberg called the false discovery rate is the expected proportion of false positives among the rejected hypotheses, and we denote it by E(FDR), which can be estimated by the mean of FDR realizations (hereafter mean FDR);
6. $Sensitivity = TP/(TP + FN)$, i.e. the proportion of altered genes which are correctly identified as such.

Since the processes generating GE and CN signals and their underlying probability distributions in real datasets are unknown, synthetic data have been generated directly from the gene expression and copy number values. Specifically, artificial CN and GE data mimicking samples with no alterations (gene status =2) have been obtained independently

permuting copy number and expression values within each chromosome *c* in each sample *j* derived from six out of 11 normal specimens of the RCC dataset (28RA, 33BV, 36MML, 37BA, 40RR and 50PC):

$$\begin{aligned}\left(\mathrm{N}_{\mathrm{j},1}^{\mathrm{c}},\ldots,\mathrm{N}_{\mathrm{j},\mathrm{L_c}}^{\mathrm{c}}\right) &\leftarrow \left(\mathrm{N}_{\mathrm{j},\pi(1)}^{\mathrm{c}},\ldots,\mathrm{N}_{\mathrm{j},\pi(\mathrm{L_c})}^{\mathrm{c}}\right)\\ \left(\mathrm{x}_{\mathrm{j},1}^{\mathrm{c}},\ldots,\mathrm{x}_{\mathrm{j},\mathrm{M_c}}^{\mathrm{c}}\right) &\leftarrow \left(\mathrm{x}_{\mathrm{j},\pi(1)}^{\mathrm{c}},\ldots,\mathrm{x}_{\mathrm{j},\pi(\mathrm{M_c})}^{\mathrm{c}}\right)\end{aligned} \quad (13)$$

where $\{\pi(1),\ldots,\pi(L_c)\}$ and $\{\pi(1),\ldots,\pi(M_c)\}$ are independent permutations of $\{1,\ldots,L_c\}$ SNP and $\{1,\ldots,M_c\}$ probe positions in chromosome *c*. CN and GE scores have been quantified directly from $N_{j,\pi}^{c}$ and $x_{j,\pi}^{c}$ according to Eq. 1 and Eq. 3. To verify the performances of the entire procedure (LSCN, LAP, and SODEGIR) under the null hypothesis, 30 random data generations (i.e., 30 data sets of CN and GE values derived permuting 5000 times copy number and expression values of six normal specimens from the RCC dataset) were analyzed and CN, GE and SODEGIR statuses quantified according to the thresholds reported in Table 2_SI. Table 4_SI reports the simulation results in terms of number of type I errors out of 16395 null hypotheses tested in each simulation.

Table 4_SI: Number of type I errors out of 16395 null hypotheses tested in each simulation.

Data type	Simulation number																													
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*
Data originated from sample 28RA																														
CN	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	78	0	32	0	0	0	0	0	0	0	0	0
GE	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	7	0	0
Data originated from sample 33BV																														
CN	0	0	0	0	0	8	10	0	0	0	0	0	0	0	0	0	65	49	0	0	0	0	0	0	0	0	0	33	0	0
GE	0	0	4	8	0	0	0	0	0	0	0	0	0	0	16	4	0	3	0	0	0	0	4	18	0	0	0	0	0	2
Data originated from sample 36MML																														
CN	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
GE	4	6	0	0	0	0	30	0	0	0	0	9	0	0	0	0	0	1	0	4	0	8	0	0	0	0	0	0	0	0
Data originated from sample 37BA																														
CN	0	0	0	0	0	0	0	0	0	0	0	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	2	0	0
GE	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	3
Data originated from sample 40RR																														
CN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GE	0	0	0	0	0	0	2	33	0	0	0	0	0	0	5	0	0	0	1	13	0	0	1	21	11	11	0	5	0	0
Data originated from sample 50PC																														
CN	0	0	0	0	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0
GE	11	0	0	6	0	0	0	0	0	0	0	0	0	10	1	0	1	0	0	0	0	0	0	0	3	0	0	8	2	34

The estimated E(FDR), i.e., in this case, the estimated probability of making one or more type I errors, (with 95% confidence intervals) are 0.083 (0.05÷0.13), 0.222 (0.16÷0.29) and 0 (0÷0.02) for CN (i.e., LSCN), GE (i.e., LAP) and SODEGIR, respectively, indicating that the overall SODEGIR procedure is highly conservative, although singularly LSCN and LAP tends to not entirely control the E(FDR) at α=0.05.

To test the performances of LSCN, LAP, and SODEGIR under the alternative hypothesis, CN and GE values of genes in a non-neutral status (i.e. $\neq 2$), within a specific chromosomal

region r, were generated adding (or subtracting) specific constants k^N and k^E to the data generated in Eq. 13:

$$\begin{aligned} N^c_{j,r} &\leftarrow N^c_{j,r} \pm k^N \qquad r \in \mathcal{R}^N \\ x^c_{j,r} &\leftarrow x^c_{j,r} \pm k^E \qquad r \in \mathcal{R}^E \end{aligned} \tag{14}$$

where $k^N, k^E \in \mathbb{R}$ and are calculated from the average standard deviations of the original data and $\mathcal{R}^N$ and $\mathcal{R}^E$ are regions within chromosome c. The average standard deviations of copy number and gene expression over all chromosomes and samples are approximately $\bar{\sigma}^N = 0.3$ and $\bar{\sigma}^E = 9/4$, respectively. As such, k^N has been set equal to $\bar{\sigma}^N$, $2/3\bar{\sigma}^N$, and $1/2\bar{\sigma}^N$ when simulating large, medium, and small effects, respectively. The corresponding values of k^E have been set equal to $2\bar{\sigma}^E$, $4/3\bar{\sigma}^E$, and $\bar{\sigma}^E$ considering the intrinsic differences between CN and GE data. Finally, CN and GE scores have been quantified directly from $N^c_{j,r}$ and $x^c_{j,r}$ according to Eq. 1 and Eq. 3. Figure 6_SI shows CN and GE simulated data with a medium amplification (up-regulation) effect, i.e. $k^N = 0.2$, $k^E = 3$ and a window of 20 Mb.

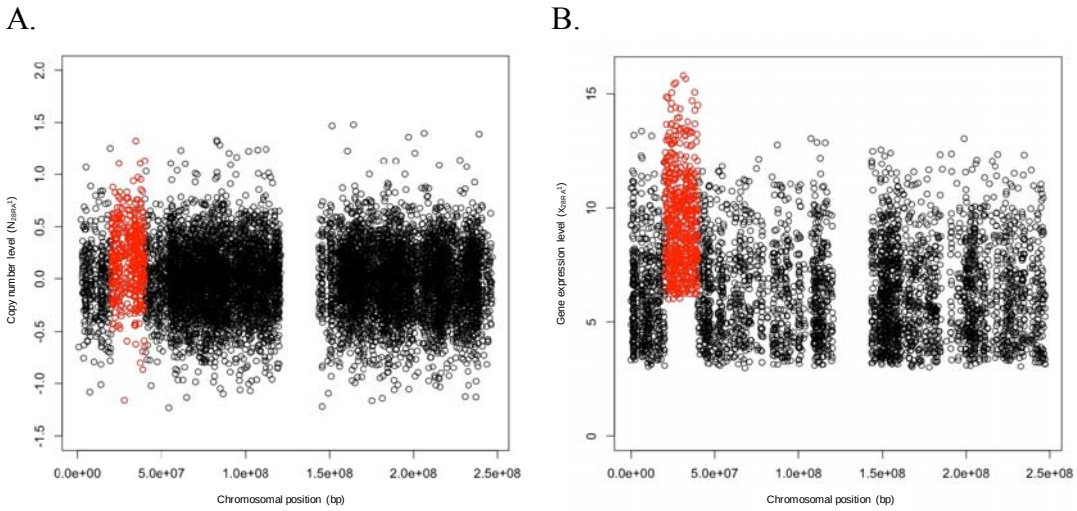

Figure 6_SI: CN (panel A.) and GE (panel B.) values of genes in a non-neutral status (i.e. ≠ 2) within a region of 20 Mb located between 20Mb and 40Mb in chromosome 1 of patient 28RA. $N^1_{28RA,(20\div40Mb)}$ and $x^1_{28RA,(20\div40Mb)}$ have been generated using with $k^N = 0.2$ and $k^E = 3$.

Non-neutral status of CN and GE signals has been simulated generating 10 non-neutral effects (named from A to L in Table 5_SI), differing in terms of affected chromosome (e.g., chromosomes 1 and 3), size of the affected regions (chromosomal segments or entire arms), and amplitude of the effect (small, medium, large) added or subtracted to CN and GE data. Figure 7_SI highlights the location of non-neutral regions and the type (i.e., gain/loss and up-/down-regulation) of each effect described in Table 5_SI.

Table 5_SI: Type of effects used to simulate non-neutral gene statuses. Effects differ in terms of affected chromosome, size of the affected regions, and amplitude of the effect added on CN and GE data (small, medium, large).

Effect name	Affected region						Effect amplitude	k^N	k^E
	chr	*from (Mb)*	*to (Mb)*	*length (Mb)*	*# of genes*	*% of genes*			
A	1	20	40	20	255	1.6	medium	0.2	3
B	1	110	115	5	60	0.4	large	0.3	9/2
C	1	120	190	70	464	2.8	small	0.15	9/4
D	1	entire p-arm		120	950	5.8	small	0.15	9/4
E	1	entire q-arm		155	806	4.9	small	0.15	9/4
F	3	10	20	10	62	0.4	medium	0.2	3
G	3	40	55	15	207	1.3	medium	0.2	3
H	3	120	165	45	248	1.5	small	0.15	9/4
I	3	entire p-arm		90	452	2.8	small	0.15	9/4
L	3	entire q-arm		109	495	3	small	0.15	9/4

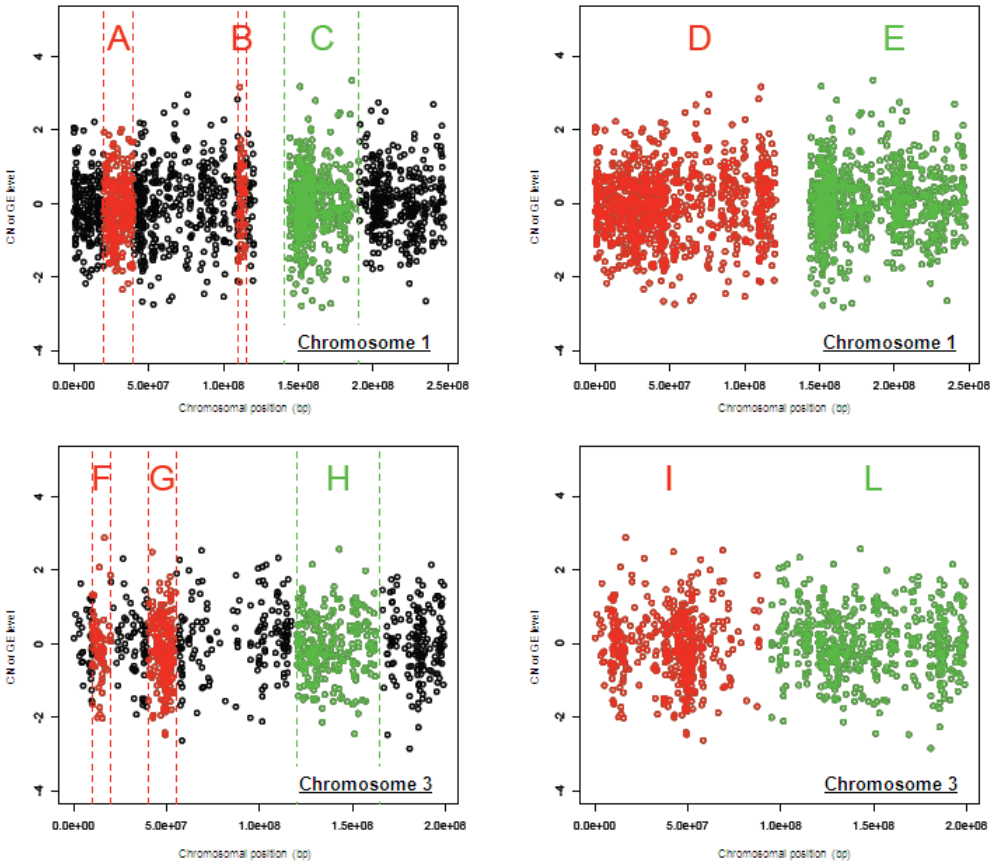

Figure 7_SI: Location of non-neutral regions and type (i.e., gain/loss and up-/down-regulation) of each effect described in Table 5_SI.

The 10 effects described in Table 5_SI have been mixed in two major scenarios, one named *small regions* and one named *large regions*, composed of 10 configurations each. Specifically, the *small regions* scenario simulates the presence of different CN and GE matched and un-matched effects (i.e. the existence or not of SODEGIRs) affecting relatively small regions of two chromosomes. Details of the 10 *small regions* configurations are as follows (Figure 8_SI):

Configuration #1: a region of 70 Mb in chromosome 1 and a region of 45 Mb in chromosome 3 both affected by a small effect CN loss and GE down-regulation (effects C and H);

Configuration #2: the same as configuration #1 plus a region of 5 Mb in chromosome 1 affected by a large effect CN gain and GE up-regulation (effects B, C and H);

Configuration #3: the same as configuration #2 plus a region of 10 Mb in chromosome 3 affected by a medium effect CN gain and GE up-regulation (effects B, C F, and H);

Configuration #4: the same as configuration #3 plus a region of 15 Mb in chromosome 3 affected by a medium effect CN gain and GE up-regulation (effects B, C F, G, and H);

Configuration #5: the same as configuration #4 plus a region of 20 Mb in chromosome 1 affected by a medium effect CN gain and GE up-regulation (effects A, B, C F, G, and H);

Configuration #1': the same as configuration #1 but the region of 45 Mb in chromosome 3 lacks of the matched GE down-regulation;

Configuration #2': the same as configuration #2 but the region of 5 Mb in chromosome 1 lacks of the matched GE up-regulation and the region of 45 Mb in chromosome 3 lacks of the matched GE down-regulation;

Configuration #3': the same as configuration #3 but the region of 5 Mb in chromosome 1 lacks of the matched GE up-regulation and the region of 45 Mb in chromosome 3 lacks of the matched GE down-regulation;

Configuration #4': the same as configuration #4 but the region of 5 Mb in chromosome 1 lacks of the matched GE up-regulation and the region of 45 Mb in chromosome 3 lacks of the matched GE down-regulation;

Configuration #5': the same as configuration #5 but the region of 5 Mb in chromosome 1 lacks of the matched GE up-regulation and the region of 45 Mb in chromosome 3 lacks of the matched GE down-regulation.

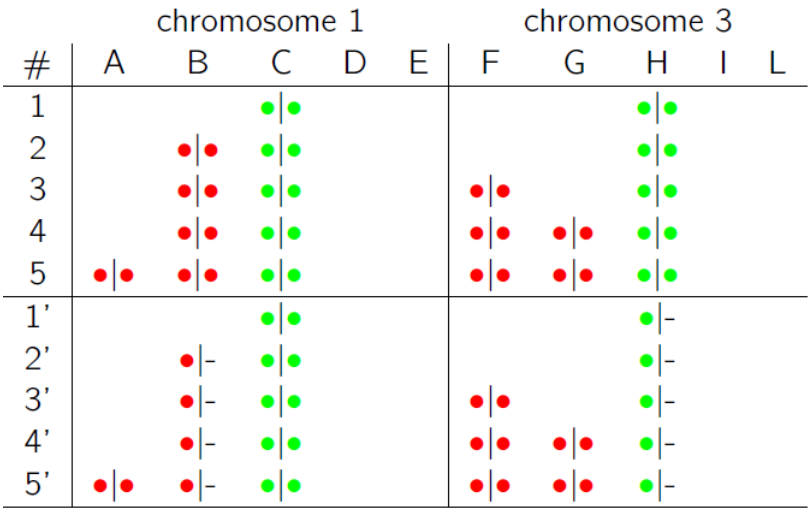

Figure 8_SI: Description of the 10 configurations of effects in the *small regions* scenario. The effect is indicated by roman letters (A to L) as described in Table 5_SI. The symbol ●|● indicates a CN gain and a concomitant GE up-regulation (amplification), ●|● a CN loss and a

concomitant GE down-regulation (deletion), and •|- a CN gain without a concomitant GE up-regulation (discordant region).

Similarly, the *large regions* scenario (Figure 9_SI) simulates the presence of:

Configuration #6: a small effect CN gain and GE up-regulation affecting the entire p arm of chromosome 1 (effect D);

Configuration #7: the same as configuration #6 plus a region of 70 Mb in the q arm of chromosome 1 affected by a small effect CN loss and GE down-regulation (effects C and D);

Configuration #8: the same as configuration #6 plus a small effect CN loss and GE down-regulation affecting the entire q arm of chromosome 1 (effects D and E);

Configuration #9: the same as configuration #8 plus a small effect CN gain and GE up-regulation affecting the entire p arm of chromosome 3 (effects D, E, and I);

Configuration #10: the same as configuration #9 plus a small effect CN loss and GE down-regulation affecting the entire q arm of chromosome 3 (effects D, E, I, and L);

Configuration #6': the same as configuration #6 lacking of the matched GE up-regulation;

Configuration #7': the same as configuration #7 lacking of the matched GE up-regulation of the p arm of chromosome 1;

Configuration #8': the same as configuration #8 lacking of the matched GE up-regulation of the p arm of chromosome 1;

Configuration #9': the same as configuration #9 lacking of the matched GE up-regulation of the p arm of chromosome 1;

Configuration #10': the same as configuration #10 lacking of the matched GE up-regulation of the p arm of chromosome 1.

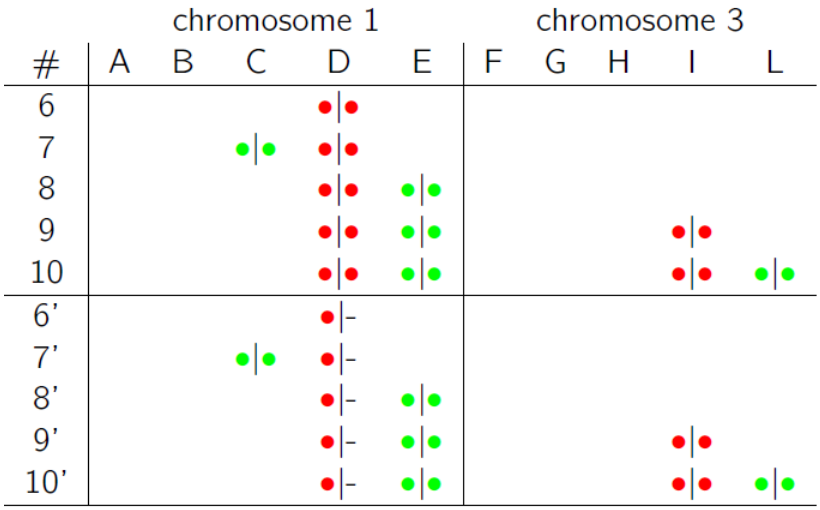

Figure 9_SI: Description of the 10 configurations of effects in the *large regions* scenario. The type of effect is indicated by roman letters (A to L) as described in Table 5_SI. The symbol •|• indicates a CN gain and a concomitant GE up-regulation (amplification), •|• a CN loss and a concomitant GE down-regulation (deletion), and •|- a CN gain without a concomitant GE up-regulation (discordant region).

The 20 different configurations have been applied to CN and GE data obtained from three different permutations of the original data signals of the 6 RCC samples previously described (Eq. 13) and then applying Eq. 14, according to the values of table 5_SI. CN and GE scores have been quantified directly from $N_{j,\pi}^{c}$ and $x_{j,\pi}^{c}$ according to Eq. 1 and Eq. 3 leading to a total of 360 different simulated sets. The analysis of the simulated data sets and the quantification of the observed CN, GE and SODEGIR statuses (according to the thresholds of Table 2_SI) 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. In Table 6_SI and Figures 10_SI and 11_SI are reported the mean sensitivity, the mean FDR and the respective 95% confidence intervals for LSCN, LAP and SODEGIR procedures when applied to the analysis of the 20 configurations of matched and un-matched alterations in the *small* and *large regions* scenarios.

Table 6_SI: Mean sensitivity, mean FDR and the respective 95% confidence intervals for LSCN, LAP and SODEGIR procedures when applied to the analysis of the 20 configurations of matched and un-matched alterations in the small and large regions scenarios (360 total sample simulations).

Configuration #	LSCN				LAP				SODEGIR			
	Sensitivity		FDR		Sensitivity		FDR		Sensitivity		FDR	
	Mean	*95% CI*	*Mean*	*95% CI*	*Mean*	*95% CI*	*Mean*	*95% CI*	*Mean*	*95% CI*	*Mean*	*95% CI*
***Small regions* scenario**												
1	0.86	0.51÷1	0	0÷0.01	0.86	0.82÷1	0.01	0÷0.05	0.77	0.49÷1	0	---
2	0.88	0.83÷0.96	0.02	0÷0.07	0.9	0.87÷0.94	0.01	0÷0.06	0.82	0.76÷0.92	0	0÷0.01
3	0.87	0.79÷0.98	0.03	0÷0.8	0.91	0.88÷0.94	0.04	0.01÷0.07	0.81	0.74÷0.92	0.02	0÷0.04
4	0.92	0.82÷0.99	0.04	0÷0.7	0.95	0.89÷0.97	0.04	0÷0.08	0.89	0.76÷0.95	0.02	0÷0.03
5	0.88	0.83÷0.94	0.03	0÷0.07	0.9	0.85÷0.95	0.04	0÷0.08	0.83	0.78÷0.90	0.01	0÷0.03
1'	0.82	0.72÷0.95	0.03	0÷0.08	0.89	0.85÷0.96	0	0÷0.05	0.75	0.6÷0.97	0	---
2'	0.92	0.83÷0.96	0.02	0÷0.05	0.9	0.87÷0.95	0	0÷0.06	0.88	0.79÷0.94	0	---
3'	0.89	0.79÷0.97	0.03	0÷0.06	0.91	0.77÷0.97	0.05	0÷0.09	0.82	0.66÷0.92	0.02	0÷0.04
4'	0.9	0.85÷0.97	0.03	0÷0.07	0.92	0.88÷0.95	0.04	0÷0.10	0.84	0.75÷0.93	0.01	0÷0.04
5'	0.89	0.82÷0.95	0.04	0÷0.07	0.87	0.79÷0.94	0.05	0÷0.10	0.83	0.72÷0.91	0.02	0÷0.04
***Large regions* scenario**												
6	0.92	0.83÷1	0	---	1	0.99÷1	0.01	0÷0.04	0.92	0.83÷1	0	---
7	0.96	0.91÷1	0	---	0.98	0.96÷0.99	0	0÷0.03	0.94	0.88÷0.98	0	---
8	0.93	0.86÷1	0	---	0.99	0.96÷1	0.01	0÷0.02	0.92	0.85÷1	0	---
9	0.96	0.87÷1	0	---	0.99	0.96÷1	0.01	0÷0.03	0.95	0.84÷1	0	---
10	0.96	0.9÷1	0	---	0.97	0.95÷0.99	0.01	0÷0.02	0.93	0.86÷0.98	0	---
6'	0.92	0.83÷1	0	---	---	---	---	---	---	---	---	---
7'	0.94	0.82÷1	0	0÷0.01	0.91	0.85÷0.98	0.01	0÷0.04	0.82	0.73÷0.95	0	---
8'	0.93	0.86÷1	0.01	0÷0.02	0.99	0.93÷1	0.01	0÷0.02	0.89	0.78÷1	0	---
9'	0.95	0.91÷1	0	---	0.99	0.94÷1	0.01	0÷0.04	0.94	0.88÷1	0	---
10'	0.95	0.91÷1	0	0÷0.01	0.95	0.89÷1	0.01	0÷0.02	0.90	0.83÷0.98	0	---

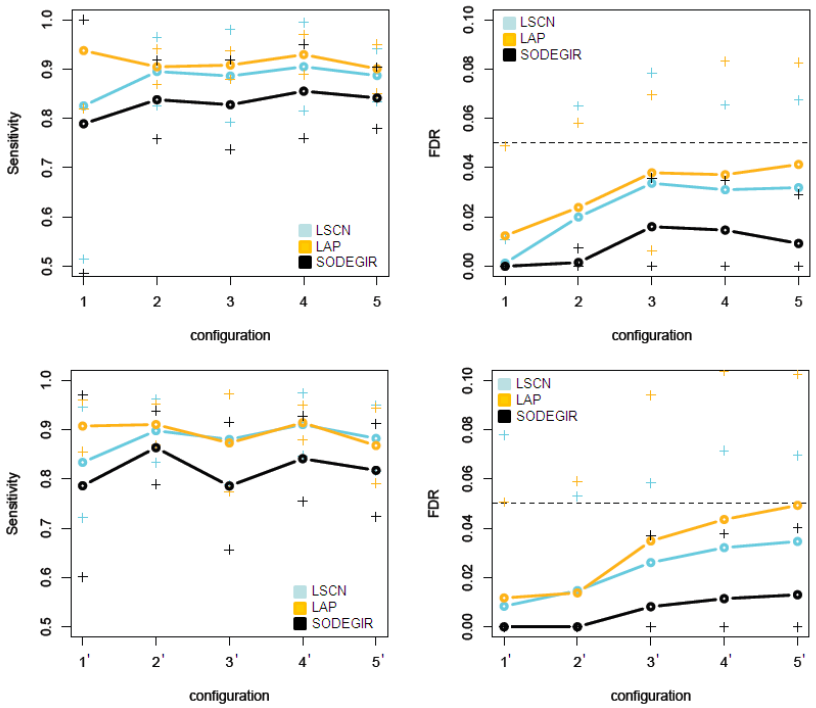

Figure 10_SI: Mean sensitivity (o), mean FDR (o) and the respective 95% confidence intervals (+) for LSCN, LAP and SODEGIR procedures when applied to the analysis of the 10 configurations of matched and un-matched alterations in the *small regions* scenario (180 total sample simulations). The dashed line indicates the 0.05 FDR threshold.

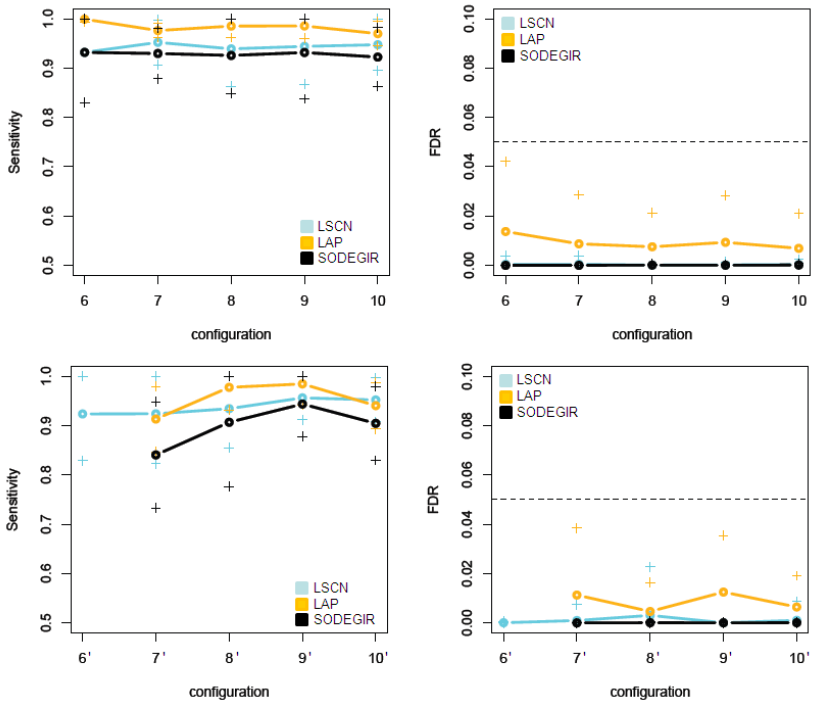

Figure 11_SI: Mean sensitivity (o), mean FDR (o) and the respective 95% confidence intervals (+) for LSCN, LAP and SODEGIR procedures when applied to the analysis of the 10 configurations of matched and un-matched alterations in the *large regions* scenario (180 total sample simulations) The dashed line indicates the 0.05 FDR threshold.

Comparison of LSCN with FASeg

The forward-backward Fragment Assembling Segmentation algorithm (FASeg) (17) was used to determine the gene copy number from genotyping data and compare results with the gene CN values estimated by LSCN at Entrez Gene IDs. In details, CN data of the *AffyRef* samples were quantified by CNAT 4.01 without any smoothing and loaded into FASeg. The significance cutoff value for the likelihood-ratio test in segmentation (*sig*) and the initial smoothing range (*smooth.range*) were set to $1e^{-5}$ and 75, respectively. FASeg returned a matrix with CN data for all SNP probes in all samples which was used to calculate the gene CN values for 24,535 gene accession numbers. After re-annotating gene accession numbers in terms of Entrez Gene IDs and filtering out duplicated identifiers, the FASeg gene CN matrix resulted in 15,702 Entrez Gene IDs, all represented in the LSCN gene CN matrix. As in LSCN, the CN status of a gene *g* in a sample *j* has been defined setting a low and a high threshold on the FASeg gene CN. Specifically, the 10^{th} and 90^{th} quantile of FASeg gene CN distributions have been selected as low and high thresholds, respectively. Similarly to the aggregation of SODEGIRs, a binomial distribution test with the q-value correction has been applied to identify regions of status concordance (amplifications and deletions) shared by a statistically relevant number of samples. Results obtained from FASeg gene CN of *AffyRef* dataset are available at the web companion site as *.SDG_tables* and *q plots* (http://www.xlab.unimo.it/SODEGIR/).

References

1. Roschke, A.V., Tonon, G., Gehlhaus, K.S., McTyre, N., Bussey, K.J., Lababidi, S., Scudiero, D.A., Weinstein, J.N. and Kirsch, I.R. (2003) Karyotypic complexity of the NCI-60 drug-screening panel. *Cancer Res*, **63**, 8634-8647.
2. Strefford, J.C., Stasevich, I., Lane, T.M., Lu, Y.J., Oliver, T. and Young, B.D. (2005) A combination of molecular cytogenetic analyses reveals complex genetic alterations in conventional renal cell carcinoma. *Cancer Genet Cytogenet*, **159**, 1-9.
3. Kotliarov, Y., Steed, M.E., Christopher, N., Walling, J., Su, Q., Center, A., Heiss, J., Rosenblum, M., Mikkelsen, T., Zenklusen, J.C. *et al.* (2006) High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. *Cancer Res*, **66**, 9428-9436.
4. Cifola, I., Spinelli, R., Beltrame, L., Peano, C., Fasoli, E., Ferrero, S., Bosari, S., Signorini, S., Rocco, F., Perego, R. *et al.* (2008) Genome-wide screening of copy number alterations and LOH events in renal cell carcinomas and integration with gene expression profile. *Mol Cancer*, **7**, 6.
5. Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D.K., Kennedy, G.C. *et al.* (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*, **65**, 6071-6079.
6. Callegaro, A., Basso, D. and Bicciato, S. (2006) A locally adaptive statistical procedure (LAP) to identify differentially expressed chromosomal regions. *Bioinformatics*, **22**, 2658-2666.
7. Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, **31**, e15.
8. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat Genet*, **36**, 949-951.

9. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525-528.
10. Sebat, J. (2007) Major changes in our DNA lead to major changes in our thinking. *Nat Genet*, **39**, S3-5.
11. Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, **98**, 5116-5121.
12. Toedling, J., Schmeier, S., Heinig, M., Georgi, B. and Roepcke, S. (2005) MACAT--microarray chromosome analysis tool. *Bioinformatics*, **21**, 2112-2113.
13. Herrmann, E. (1997) Local bandwidth choice in kernel regression estimation. *Journal of Graphical and Computational Statistics*, **6**, 35-54.
14. Gasser, T. and Müller, H.G. (1979) In Rosenblatt, G.a. (ed.), *Smoothing Techniques for Curve Estimation*. Springer Verlag, Heidelberg.
15. Shaffer, J.P. (2002) Multiplicity, directional (type III) errors, and the null hypothesis. *Psychol Methods*, **7**, 356-369.
16. Berger, R.L. (1982) Multi-Parameter Hypothesis-Testing and Acceptance Sampling. *Technometrics*, **24**, 295-300.
17. Yu, T., Ye, H., Sun, W., Li, K.C., Chen, Z., Jacobs, S., Bailey, D.K., Wong, D.T. and Zhou, X. (2007) A forward-backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (SNP) array. *BMC Bioinformatics*, **8**, 145.