Published online 10 January 2008

Nucleic Acids Research, 2008, Vol. 36, No. 2 ***e13***
doi:10.1093/nar/gkm1143

Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data

Christiaan Klijn[1,2], Henne Holstege[1], Jeroen de Ridder[1,2], Xiaoling Liu[1], Marcel Reinders[2], Jos Jonkers[1,*] and Lodewyk Wessels[1,2]

[1]Netherlands Cancer Institute, Division of Molecular Biology, Plesmanlaan 121 1066 CX Amsterdam and [2]Delft University of Technology, Information and Communication Theory group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft, The Netherlands

Received August 20, 2007; Revised December 7, 2007; Accepted December 10, 2007

ABSTRACT

Tumor formation is in part driven by DNA copy number alterations (CNAs), which can be measured using microarray-based Comparative Genomic Hybridization (aCGH). Multiexperiment analysis of aCGH data from tumors allows discovery of recurrent CNAs that are potentially causal to cancer development. Until now, multiexperiment aCGH data analysis has been dependent on discretization of measurement data to a gain, loss or no-change state. Valuable biological information is lost when a heterogeneous system such as a solid tumor is reduced to these states. We have developed a new approach which inputs nondiscretized aCGH data to identify regions that are significantly aberrant across an entire tumor set. Our method is based on kernel regression and accounts for the strength of a probe's signal, its local genomic environment and the signal distribution across multiple tumors. In an analysis of 89 human breast tumors, our method showed enrichment for known cancer genes in the detected regions and identified aberrations that are strongly associated with breast cancer subtypes and clinical parameters. Furthermore, we identified 18 recurrent aberrant regions in a new dataset of 19 p53-deficient mouse mammary tumors. These regions, combined with gene expression microarray data, point to known cancer genes and novel candidate cancer genes.

INTRODUCTION

Malignant transformation of normal, healthy cells is strongly dependent on changes in the expression of oncogenes and tumor suppressor genes. DNA copy number alteration (CNA) is an important mechanism through which tumor cells can modulate expression of cancer genes. CNAs are a result of genomic instability, which has been described as an enabling mechanism of tumorigenesis (1). Aberrations range in size from relatively small regions (<0.5 Mb) to entire chromosomes and the various mechanisms by which they arise are unclear (2). Comparative genomic hybridization (CGH) is a technique that is capable of measuring CNAs. The term CGH was first used for competitive hybridization of differentially labeled DNA on metaphase chromosomes (3) to measure CNAs.

Various microarray platforms have enabled high-resolution genome-wide analysis of CNAs by array-based CGH (aCGH) (4). Many different platforms are currently available for aCGH analysis, such as bacterial artificial chromosome (BAC) clone (4), cDNA clone (5), single nucleotide polymorphism (SNP) (6) and oligonucleotide based platforms (7,8). Most of the platforms sample the genome at specific positions with a certain distance between the measurement points (probes). This distance is referred to as the resolution of the platform. Array CGH data generally consist of the log-ratios of normalized hybridization intensities of fluorescently labeled DNA from disease versus 2*n* control samples measured by the probes spotted on these microarrays.

Array-CGH has been used extensively in cancer research (9) and careful analysis of CNAs can facilitate cancer gene discovery (10,11). Various automated methods have been described to analyze the results obtained from aCGH measurements. They typically either smooth the data and/or try to estimate the location of the aberration by defining 'break-points' at which the CNA is defined to start or end (12,13). This is always done at the single tumor level and procedures which define aberrations often rely on several parameter choices to find

*To whom correspondence should be addressed. Tel: +31-205122000; Fax: +31-206691383; Email: j.jonkers@nki.nl
Correspondence may also be addressed to Lodewyk Wessels. Tel: +31-205127987; Fax: +31-206691383; Email: l.wessels@nki.nl

putative CNAs. It is not always clear how these parameters relate to biological reality. Recently, methods to discretize aCGH data have been extended to include states other than 1, 0 and −1 (13–15). However, these methods do not provide a solid statistical framework to identify CNAs that occur in a significant fraction of the tumors. Especially these recurrent regions may harbor genes relevant for tumor development.

Three methods have been developed to perform multiexperiment analysis to identify recurrent CNAs within a group of tumor samples: CMAR (16), STAC (17) and H-HMM (18), a hidden Markov Model (HMM)-based algorithm. Of these, STAC and H-HMM provide probabilistic outputs. Both STAC and CMAR suffer from the fact that aCGH data have to be discretized into three states (1, 0 and −1) before the method can be applied. H-HMM uses continuous data as input, but still employs three discrete states in its hidden model. CNAs are discrete in nature on a single cell level, as gains and losses of a region of DNA can occur only on a per-copy basis. Although a method for single cell aCGH has recently been published (18), most aCGH analyses are not performed on single cells, but rather on a population of tumor cells mixed with stromal components. This stromal compartment will influence the signal measured by quenching the true CNA level. Moreover, as solid tumors are known to be heterogenic in nature and contain sub-populations of clonally diverse cells (19), it is also safe to assume that not all tumors cells will carry the same CNAs. This implies that an aCGH measurement will measure a population of CNAs present in the tumor sample. Apart from the fact that there is no consensus regarding the best discretization method to use, reducing a signal from such a heterogeneous population of CNAs to values of 1, 0 and −1 leads to loss of potentially valuable biological information.

We have developed a statistical method for multi-experiment aCGH analysis of nondiscretized aCGH data capable of detecting statistically significant aberrations of varying size. To this end, we developed KC-SMART, Kernel Convolution: a Statistical Method for Aberrant Region deTection. This method employs kernel convolution (20) to perform locally weighted regression (21), which produces a smoothed estimate of the CNAs. This approach takes into account (i) the nondiscretized strength of a clone; (ii) the strength of neighboring clones in the same tumor, as well as (iii) the occurrence frequency of this clone across all tumors in the dataset. Kernel regression automatically corrects for the unequal distances between the probes. KC-SMART allows analysis of both small and large aberrations using different values for the width of the kernel function. This is important, as DNA copy number aberrations have a large variation in size (2). Using a published dataset of 89 sporadic breast tumors (22) we find that KC-SMART identifies several aberrations that correlate with breast cancer subtype and clinical parameters. We also identify regions that are enriched for known cancer genes.

Furthermore, KC-SMART performs better than STAC (17), the only other statistically-based method currently available. Using aCGH and expression data from a novel dataset of 19 mammary tumors from a conditional mouse model of sporadic breast cancer, we identify several cancer genes, and we show that complex aberrations can be analyzed by varying the width of the kernel function. Finally, we propose several new cancer gene candidates based on KC-SMART results.

METHODS

Human breast cancer dataset

aCGH data were used as acquired from the Supplementary Data of Chin *et al.* (22). Preprocessing and normalization was performed and described by Chin *et al.* (22). Because no exact mapping information was available for all clones in the Chin dataset, we gave the clones a 3-bp length centered on the mapping position as supplied by Chin *et al.* We removed all clones with more than 50% missing values. We imputated the remaining missing values using the averaged values of their two positional neighbors. Probes mapped to the same area were averaged and represented as a single clone. This resulted in 2149 unique clones. Gene expression data were also acquired from Chin *et al.* (Arrayexpress accession number: E-TABM-158). Probes not mapping to a single ENSEMBL ID were removed; probes mapping to Y chromosome genes were removed. This resulted in 21 339 unique Affymetrix probe measurements.

Mouse p53 dataset

All BAC information was based on NCBI assembly 36 of the mouse genome. The platform used to analyze the mouse DNA was a 3k mouse BAC platform (23). Extraction of DNA, labeling and hybridization to the array was performed as described by Chung *et al.* (23). Dye-swaps were performed for all samples. The data were normalized using median normalization. BAC clones containing more than 10% missing values were deleted from the dataset. The remaining missing values were imputed using the values of their two positional neighbors. Should the imputed BAC clone be the first or the last on a chromosome, it is imputed using the value of its remaining neighbor. The preprocessing left 2895 BACs for analysis. Gene expression analysis was performed on in-house 32k mouse oligo arrays. Detailed methods can be found in the Supplementary Data. The data were normalized using the Rosetta error model. The oligos were BLASTed against the mouse transcriptome as of May 2006. All oligos targeting multiple genes were discarded. All positional information of the target genes was acquired from the NCBI assembly 36 of mouse. Genes for which no positional information was available were discarded. Genes containing more than 10% missing values were discarded. Remaining missing values were imputed using the gene average over the nonmissing values. Preprocessing left 25 809 oligos targeting 19 104 unique genes. p53 conditional knockout mouse aCGH data are under submission at the GEO database, accession number GSE7794. Gene expression data are deposited at the Arrayexpress database accession number E-NCMF-6, details are in the Supplementary Data.

KC-SMART analysis

Both datasets were analyzed equally by permutation analysis. The KC-SMART permutation analysis was performed to create a null distribution, and peaks were detected at $P < 0.05$ (Bonferroni corrected for multiple testing), for kernel widths: $6\sigma = 2$, 4, 6, 8, 10, 12, 16, 20, 24, 30 and 40 Mb. One thousand permutations were used. The interpolation, as discussed below, was performed for an interval [2, 40] Mb, with steps of 0.2 Mb and was only applied to the mouse dataset. MATLAB scripts used for implementation of the method are available from the authors upon request.

Scale space interpolation

To be able to lower the computational load which is required if the significance threshold needs to be determined for a large number of kernel widths, an interpolation step was introduced for more detailed analysis of the mouse dataset. The significance level at a certain *P*-value is determined for several different kernel widths employing the permutation approach outlined earlier. Then, a model is fitted to these data which describe the significance threshold as a function of the kernel width. The function found to fit the relationship best was a modified power function:

$$\tau = a \cdot b^{w_{kernel}} + c \quad \mathbf{1}$$

where τ is the significance threshold for the given w_{kernel}. The coefficients a, b and c are estimated using an iterative procedure that minimizes the sum of squared errors using randomized starting values. This function is then used to determine the significance threshold for all kernel widths within the range of the original widths tested. This allows a detailed examination of the 'scale space', which is defined as the significant areas in the genome given different kernel widths.

Integration of gene expression data

We correlated gene expression profiles of the genes in these significant regions with the CNA data. First we tested the pair-wise correlation of BAC clones in a significant region. If the average pair-wise correlation coefficient exceeded 0.5, we averaged the BAC-signals across all clones, resulting in a region BAC profile, capturing the average behavior of a BAC clone in the aberration. This profile was correlated with the gene expression profile of each of the genes in the same region. We calculated a *P*-value of the correlation of each gene's expression profile with the region BAC profile using the Students *t*-test. All genes that correlated at $P < 0.05$ (Bonferroni corrected) were called significantly correlated.

Cancer gene enrichment

The cancer gene census (CGC) list was retrieved from the Internet (www.sanger.ac.uk/genetics/CGP/Census/). Mapping information was retrieved from the Ensembl database (www.ensembl.org) using the Biomart data mining tool. Only genes with valid mapping information were used. The genes described in the CGC were divided according to their label 'recessive' or 'dominant'. This resulted in 275 unique dominant genes and 65 recessive genes. Enrichment for genes labeled as dominant was tested in gains and enrichment for genes labeled as recessive genes was tested in losses. For enrichment analysis of the genes in the Chin dataset, only those genes with measured gene expression profiles were used and only those genes with an ENSEMBL-gene identifier. This left 11 409 unique genes measured.

DAVID analysis

Genes found significant by correlation were compiled into gene lists according to method (KC-SMART, STAC) and according to gain/loss status. Only genes with valid Entrez and Ensembl IDs that were present on the gene expression microarray platform were considered. As background, gene list of all genes with Entrez and Ensembl Ids on the array platform was used. Association with cancer was calculated using the DAVID web interface (david.abcc.ncifcrf.gov) and as described in ref. 24.

Discrete data

The raw data used for frequency analysis are described in Chin *et al.* (22) and were segmented using the parameter-free segmentation algorithm CBS (25) and assigned copy number change using MergeLevels (13), courtesy of J. Fridlyand. All positive copy number changes were counted as gains (1); all negative copy number changes were counted as losses (−1) and no copy-number change was set to 0.

Frequency analysis. For each BAC clone, the overall frequencies of 1 (gains) and −1 (losses) were counted and plotted as a percentage on its genomic location. For comparison, frequencies of 30% and higher were chosen to be significant.

STAC analysis. STAC inputs data discretized to three levels: 1, 0 and −1. The raw data used for frequency analysis are described in Chin *et al.* (22) and were segmented using the parameter-free segmentation algorithm CBS (25) and assigned copy number change using MergeLevels (13), courtesy of J. Fridlyand. All positive copy number changes were counted as gains (1); all negative copy number changes were counted as losses (−1) and no copy-number change was set to 0. We followed the advice on the STAC website (cbil.upenn.edu/STAC/) and divided each chromosome arm into stretches of 1 Mb. A stretch was called aberrant whenever a BAC clone present in that stretch was called aberrant. Stretches containing no BACs were set to 1/−1 when both neighbors were 1/−1, 0 when one of them was 0. The preprocessed data were loaded into the STAC 1.2 java applet and analyzed using 1000 permutations and a 1-Mb span constant. The OR regions were defined by either the frequency statistic or the footprint statistic being $P < 0.05$. The 'AND' regions were defined by both the frequency statistic and the footprint statistic being $P < 0.05$.

RESULTS

Array CGH data consist of the log-transformed ratios of normalized intensities from case versus control samples measured on certain positions on the genome. In heterogeneous tumors, gains or losses do not assume discrete amplitude values along the genome. To model the fact that CNAs are manifested in a continuous manner in an aCGH profile of a tumor, we start off by producing a smoothed estimate of the cross-tumor averaged genome-wide CNAs profile using kernel convolution (20). Array CGH probes are not distributed equally across the genome. This unequal probe spacing influences the smoothed estimate and needs to be corrected for. To this end, we employed a kernel convolution-based method to perform locally weighted regression (21).

A schematic overview of the KC-SMART method is depicted in Figure 1. The general principle of KC-SMART is to place a kernel function at the position of each probe. At an arbitrary position x on the genome, the kernel smoothed estimate (KSE) of the log2 ratio is given by

$$KSE(x) = \frac{\sum_{M_i} a_i \cdot g_i(x)}{\sum_{M_i} g_i(x)} \qquad 2$$

Where a_i is the sum of all positive or negative log2 values across all tumors for probe i, $g_i(x)$ is the kernel function

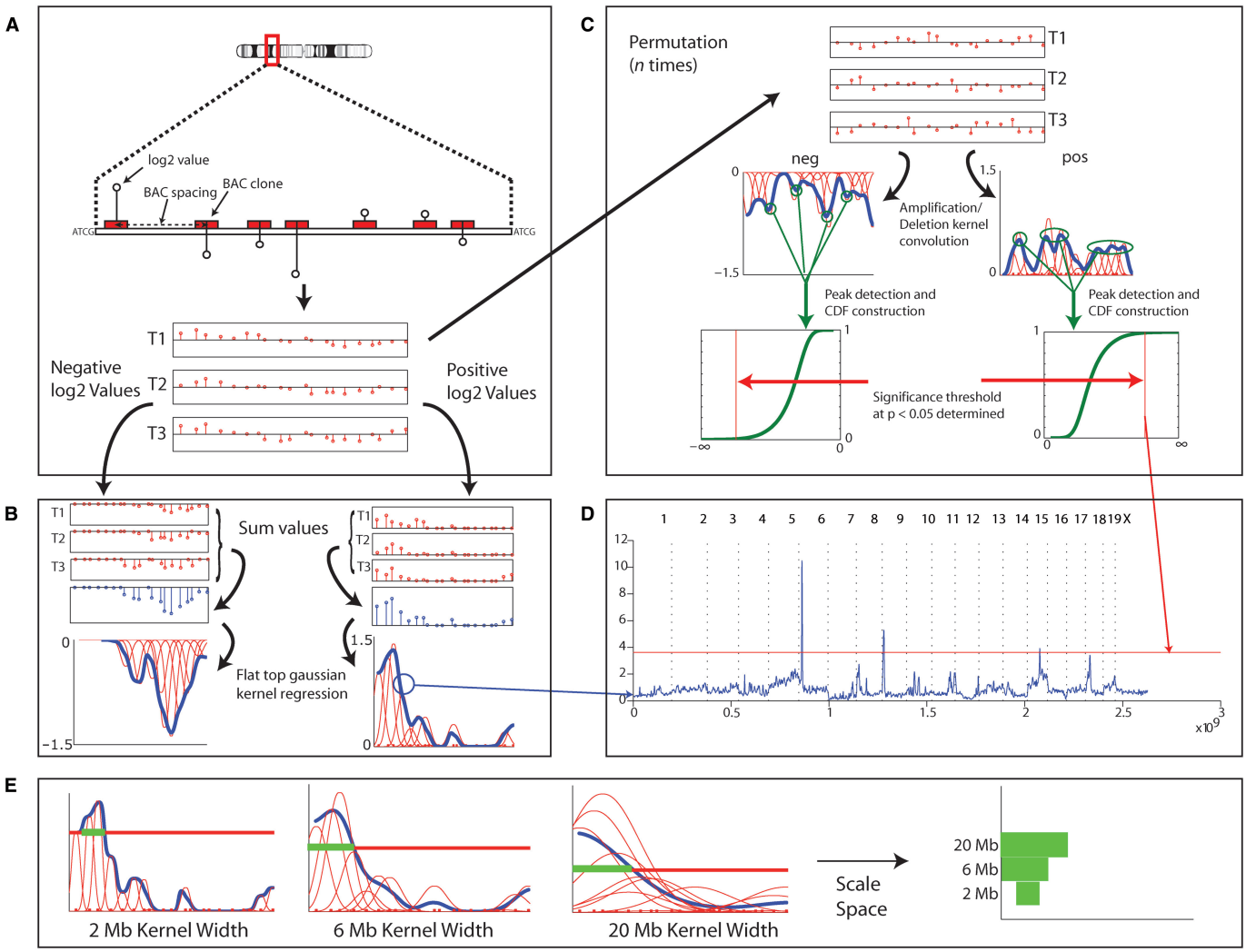

Figure 1. A schematic overview of KC-SMART. T1, T2 and T3 represent three arbitrary tumor samples. (**a**) Illustration of the nature of the data measured and how it is represented on the genome. The BAC clones are spaced along the genome where the sizes of the gaps depend on the platform used. Per BAC clone a log2 value is measured that is a representation of the CNA at that point on the genome. (**b**) The positive and negative log2 values in the data are separated and summed across tumors and per BAC clone. After summation the kernel convolution is applied and the Kernel Smoothed Estimate (KSE: blue line) is determined. (**c**) An overview for the method of determining statistical significance is shown here. First, the original log2 values are shuffled randomly within each tumor. After summation across tumors the KSE is computed. For both the gains and the losses a cumulative density function (CDF) of the detected peaks is calculated. By testing the significance level against this CDF a value is obtained, above which peaks are found to be significant. (**d**) Here the result of a genome-wide analysis is shown. The blue line is the KSE obtained from the data and the red line is the significance threshold at $P = 0.05$, which was determined in (c). (**e**) The scale space is constructed by arranging the significantly aberrant areas (visualized as blocks) in order of scale on the genomic position.

[Equation (3)] and M_i the set of probes contributing to the KSE. The denominator implements the locally weighted regression to account for unequally distributed clones.

Gains and losses are analyzed separately since gains and losses are fundamentally different (only a few copies of a region can be lost, depending on the ploidy of the cell, but many copies can be gained). The separation of gains and losses also prevents positive and negative values from summing to zero. Several kernel functions were considered for constructing the KSE. A flattop Gaussian kernel was chosen since it remains constant across the length of the BAC clone, and then drops off in a smooth, monotonically decreasing fashion. This choice is based on the assumption that probe signals of immediate neighbors are more predictive than more distant probes. A graphical representation of the flattop Gaussian function is shown in Figure 2. The function $g_i(x)$ is defined as:

$$g_i(x) = I_{\{x \le \mu_{i1}\}} \cdot e^{\frac{-(x-\mu_{i1})^2}{2\sigma^2}} + I_{\{x \ge \mu_{i2}\}} \cdot e^{\frac{-(x-\mu_{i2})^2}{2\sigma^2}} + I_{\{x \in [\mu_{i1}, \mu_{i2}]\}} \quad \mathbf{3}$$

The variables μ_{i1} and μ_{i2} represent the mapped genomic start and end position of the probes and σ determines the width of the kernel function. I is the indicator function defined as:

$$I_{\{b\}} = \begin{cases} 1 & \text{if } b \text{ is true} \\ 0 & \text{otherwise} \end{cases} \quad \mathbf{4}$$

Set M_i theoretically contains all probes on the same chromosome as x, as they all contribute to sample point x. To reduce computational load we only consider probes that lie 4σ upstream and downstream from sample point x. In this range, 99.8% of the kernel surface is contained and the error for not including kernels further up- or downstream is negligible. To correct for boundary problems at the chromosome ends and the centromeres, the probes up to half the kernel width from the boundary are mirrored.

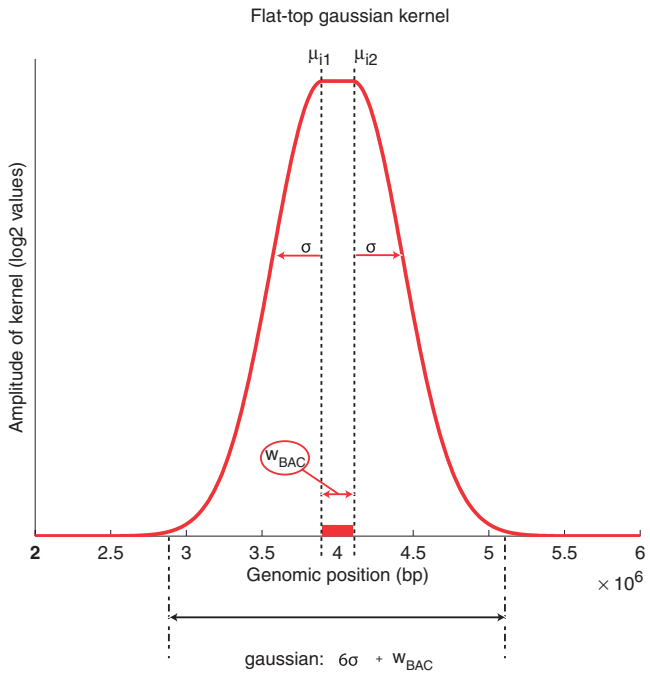

Figure 2. Main properties of the flat top Gaussian function The BAC clone is depicted as a red rectangle. The amplitude of the kernel is determined by the summed log2 value of that BAC clone across all tumors. The red line is a representation of g_i.

It should be noted that the KSE of the aCGH ratio at a clone position is the result of information gathered along two dimensions: (i) information from all tumors is incorporated through cross-tumor averaging and (ii) information from neighboring clones is incorporated through local regression.

The statistical significance of the peaks found in the KSE was determined by testing against a null distribution of peak heights acquired from the KSEs obtained from several random within-tumor permutations of the original data (Figure 1c). The P-value is corrected for multiple testing using a Bonferroni approach by multiplying the resulting P-value by the number of peaks tested, thus controlling the family-wise error (26). Combining the threshold found by permutation with the KSE produces the result visualized in Figure 1d. All regions where the KSE exceeds the significance threshold (red line) are significantly aberrant across the tumors, for a specific kernel width.

Biologically, CNAs are found in a large variety of shapes and sizes, from small amplifications to whole chromosome loss due to uneven mitotic chromatid separation. It is desirable to detect both large and small aberrations. Furthermore, the strength of gain or loss is an important factor, as high-level gains may reflect strong oncogene activation and high-level losses may reflect homozygous deletion of tumor suppressor genes. Different kernel widths must be used to find both very localized, high-level and low-level CNAs that span a large region. The width of the kernel function will determine how broad the region will be across which a probe amplitude will influence the KSE, i.e. how much the kernels will overlap. A narrow kernel can be expected to detect small local aberrations, which will be smoothed away by a wide kernel. Conversely, a wide kernel will identify long aberrations of low to medium strength such as a single copy loss or gain of a chromosome in a fraction of the tumor cells, while missing localized aberrations. By applying KC-SMART to a dataset using different kernel widths, we can construct a scale space of the dataset which gives a comprehensive view of the aberrations in the dataset. Another advantage of the scale-space analysis is that within a single aberration multiple levels of CNA may be present. Analyzing these aberrations using multiple scales can identify the region within an aberration that is significant across scales, and therefore more interesting. A visualization of how the scale space is constructed is depicted in Figure 1e.

A typical recurrent aberrant region is several megabases long and may contain many genes. To filter the genes in a recurrent region identified by KC-SMART we use gene expression measurements of the genes in those regions. When the gene expression of a gene within a recurrent CNA correlates well ($P < 0.05$, Bonferroni corrected) with

the average aCGH profile of all probes within the region, we consider it as an interesting candidate gene.

Validation of KC-SMART

We applied KC-SMART to a publicly available dataset of 89 human breast cancer samples for which both aCGH data and gene expression data were available (22). The results of the KC-SMART analysis are shown in Figure 3. As is immediately apparent, these tumors are characterized by a larger number of recurrent losses as opposed to recurrent gains. Whole chromosome arm aberrations, such as the well-known 1q and 8q gains and 17p loss are found to be significantly recurrent (27,28). Also smaller, local aberrations are identified such as the 17q amplification of the *ERBB2* gene and the gain of the 11q region containing cyclin D1.

To test the biological relevance of the results, we assessed the enrichment for known cancer genes among the significantly correlating genes located in the significantly recurrent regions. (See the Methods section for the determination of significantly correlating genes.) We used the hypergeometric test to assess significant enrichment. As 'known' cancer genes, we used the Cancer Gene Census (CGC) (29). More specifically, we tested genes in gains against the CGC dominant genes and the genes in losses against the CGC recessive genes. We used the dominant/recessive terminology from the Futreal *et al.* (29) paper which is not synonymous with the classical genetic definition of dominant and recessive alleles, but rather is a crude separation between oncogenes and tumor suppressor genes. As shown in Table 1, both KC-SMART analyses show significant enrichment for CGC genes ($P < 0.05$).

As an alternate test we explored the association of the genes in the various lists to cancer using the Genetic Association Database (geneticassociationdb.nih.gov/). Their association with the 'cancer' category was determined using the DAVID Bioinformatics Resources (24). As shown in Table 1, the KC-SMART results also show a significant association with cancer ($P < 0.05$) for this database.

In order to extend our biological validation of the data, we determined whether regions found significant by KC-SMART were associated with molecular breast cancer subtypes (30). For each tumor in the 89 tumor

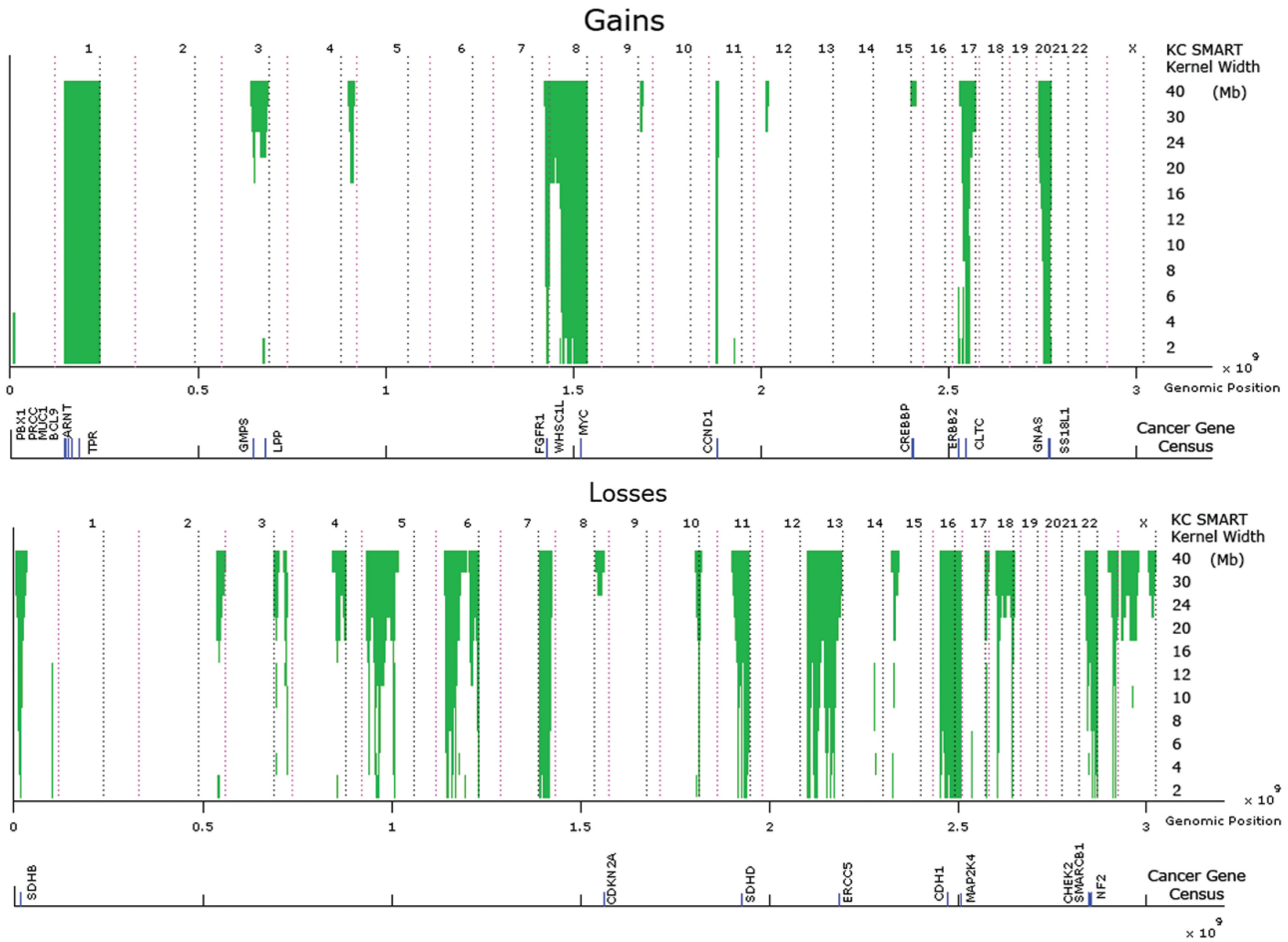

Figure 3. Results of KC-SMART analysis of 89 human breast tumor samples. The human tumor set was acquired from Chin *et al.* (22). Significant recurrent regions found by KC-SMART are shown in green. Significantly correlating genes from the Cancer Gene Census (CGC) list are shown below for each result. The cancer gene census list was split in CGC dominant genes (for gains) and CGC recessive genes (for losses). Black dotted lines represent the end of chromosomes; magenta dotted lines represent the centromere location.

Table 1. Enrichment of correlating genes found by KC-SMART for known cancer genes

Cancer Gene Census Enrichment	Number of genes in gene set[a]	Number of CGC[b] genes in gene set	*P*-value enrichment for CGC[b] genes[c]
KC-SMART gains*	521	17	0.0213
KC-SMART losses*	590	9	0.0006
DAVID analysis	Number of correlating genes[d]	Number of cancer genes	*P*-value association with cancer category (EASE score)
KC-SMART gains*	529	13	0.021
KC-SMART losses*	609	11	0.046

*Gene sets marked with an asterisk showed significant enrichment for cancer genes at $P < 0.05$.
[a]Numbers of genes were based on genes that had an Ensembl ID.
[b]Cancer Gene Census.
[c]Calculated using the cumulative density function of the hypergeometric test. Total number of genes in the genome used for this analysis: 11409, total number of dominant CGC genes: 275, total number of recessive CGC genes: 65.
[d]Numbers of genes were based on genes that had an Entrez ID.

Table 2. Association of KC-SMART significant regions to the five breast cancer subtypes (30)

Molecular Subtypes			ERBB2 subtype		Basal subtype		Luminal A subtype		Luminal B subtype		Normallike subtype	
Chromosome Arm	Start (Mb)	End (Mb)	no tum	pvalue	no tum	pvalue	no tum	pvalue	no tum	pvalue	no tum	pvalue
17 q	36.1	38.4	6	0.0000	1	0.6651	0	0.9853	0	0.6680	3	0.2629
4 q	155.2	191.7	2	0.6378	13	0.0004	7	0.9310	3	0.4408	2	0.6378
5 q	53.3	138.9	0	0.9702	15	0.0000	5	0.9836	3	0.3709	2	0.5737
10 p	0.0	9.3	2	0.3615	10	0.0008	2	0.9972	4	0.0521	1	0.6790
14 q	82.6	86.9	2	0.6679	12	0.0036	6	0.9840	7	0.0034	1	0.8868
16 p	0.0	10.3	1	0.7911	1	0.9961	16	0.0001	0	0.9706	5	0.0167
16 q	48.5	89.5	0	0.9852	5	0.8488	19	0.0001	3	0.5104	2	0.6966
22 q	15.5	49.2	2	0.4699	1	0.9945	14	0.0017	2	0.5450	3	0.2066
8 q	74.3	146.3	3	0.4514	10	0.0811	6	0.9935	9	0.0001	2	0.7238
11 p	0.0	8.6	0	0.9428	8	0.0429	5	0.9227	6	0.0029	2	0.4340
14 q	82.6	86.9	2	0.6679	12	0.0036	6	0.9840	7	0.0034	1	0.8868
9 p	0.2	25.7	2	0.3978	2	0.9452	6	0.7590	3	0.2080	7	0.0001

Gained Regions
Lost Regions
Significant result at 5% FDR

For each subtype the significantly enriched aberrations are listed. Significance was determined using the hypergeometric distribution and corrected using Benjamini–Hochberg multiple testing correction. Locations mentioned in the start and end column are accurate up to maximally 1 Mb, as this is the resolution of the array.

dataset Chin *et al.* (22) determined the molecular subtype based on gene expression profiles. To determine which tumors contribute to a certain significant aberration, we ranked the tumors based on their average log2 value across all BAC clones in the aberrant region. Next, we calculated the cumulative contribution of the top *n* tumors to the total average log2 value in the aberrant region. The top *n* tumors, for which the cumulative contribution reached 75%, were marked as significantly contributing to that aberration. The threshold of 75% is chosen to be a conservative estimate, to ensure that all tumors truly contain this aberration. Among the tumors that contribute to a certain aberration we calculated the enrichment for each molecular subtype using the hypergeometric distribution. As can be seen in Tables 2 and 3, this analysis reveals significant association of several well-known aberrant regions with distinct molecular subtypes (22,31) and clinicopathological features.

Three regions with interesting associations were selected and summarized in Table 4. One of these regions is the 5q loss, which is associated with the basal-like breast cancer subtype and *ER*, *PR* and *ERBB2* negative status. Basal-like breast tumors are known to have a low frequency of *ER* and *PR* expression, and rarely overexpress ERBB2 (32). Similarly, a gain on chromosome 16p can be considered a marker for ER and PR positive tumors and the Luminal A subtype (28). The loss on 9p is most common in normal-like tumors, which are associated with a genomically stable diploid genotype (33). This region contains the well-known tumor suppressor *CDKN2A* ($p14^{ARF}$). *CDKN2A* gene expression was also found to significantly correlate with the 9p CNA profile, and can therefore be considered a putative driver gene for this location. These results show that association studies with regions detected by KC-SMART not only recapitulate known associations (5q, 16p) but also identifies potential new associations (9p).

The classical way of analyzing aCGH data is to produce frequency plots. This approach counts the frequency of occurrence of probes called as gained, lost or unchanged. The method of calling varies from setting a rough threshold on the raw data to applying dedicated segmentation and calling algorithms. To compare KC-SMART results to frequency-based analyses we employed a

Table 3. Association of KC-SMART significant regions to the five breast cancer subtypes (30)

ER				positive		negative	
Chromosome	Arm	Start (Mb)	End (Mb)	pval	no. tum	pval	no. tum
16	p	0.0	10.3	0.0004	20	0.9970	3
22	q	15.5	49.2	0.0008	19	0.9951	3
4	p	0.8	14.8	0.9827	10	0.0047	14
4	p	24.3	37.5	0.9928	11	0.0018	16
5	q	53.3	138.9	0.9979	9	0.0004	16
10	p	0.0	9.3	0.9873	7	0.0029	12
14	q	82.6	86.9	0.9966	11	0.0007	17

PR				positive		negative	
Chromosome	Arm	Start (Mb)	End (Mb)	pval	no. tum	pval	no. tum
1	q	143.2	239.9	0.0001	31	0.9986	11
16	p	0.0	10.3	0.0003	19	0.9975	4
16	q	48.5	89.5	0.0012	22	0.9981	6
5	q*	53.3	138.9	0.9938	8	0.0042	16

p53				positive		negative		not scored
Chromosome	Arm	Start (Mb)	End (Mb)	pval	no. tum	pval	no. tum	no. tum
5	q	53.3	138.9	0.0022	10	0.9891	12	3
14	q	82.6	86.9	0.0030	11	0.9865	15	2
15	q	21.4	42.1	0.0011	11	0.9941	13	0
17	p	0.7	19.5	0.0023	12	0.9894	17	1
16	p	0.0	10.3	0.9900	1	0.0009	20	2

ERBB2				positive		negative		not scored
Chromosome	Arm	Start (Mb)	End (Mb)	pval	no. tum	pval	no. tum	no. tum
17	q	36.1	38.4	0.0000	7	1.0000	0	0
5	q	53.3	138.9	0.9456	0	0.0000	23	3
6	q	78.1	124.4	0.9019	0	0.0000	19	1
11	p	0.0	8.6	0.8704	0	0.0000	17	4
11	q	67.2	73.5	0.8065	0	0.0000	14	2
11	q	67.8	71.1	0.7798	0	0.0000	13	1
15	p	21.4	42.1	0.9456	0	0.0000	23	0
15	p	24.8	28.6	0.9601	0	0.0000	25	1
16	p	0.0	10.3	0.9265	0	0.0000	21	2
16	q	48.5	89.5	0.9712	0	0.0000	27	2

Gained Regions
Lost Regions
Significant result at 5% FDR

For each clinical parameter the significantly associated aberrations are listed. Significance was determined using the hypergeometric distribution and corrected using Benjamini–Hochberg multiple testing correction8. Locations mentioned in the start and end column are accurate up to maximally 1 Mb, as this is the resolution of the array.
*The 5q loss associated with PR-negative staining was included as it was the most significant association with PR-negativity and only borderline not-significant.

Table 4. Summary of association for three recurrent regions identified from 89 human breast tumors

Aberration	Subtype	ER	PR	ERBB2	P53 mutated[b]
5q loss (53 Mb–138 Mb)	Basal	neg	neg[a]	neg	pos
16p gain (0 Mb–10 Mb)	Luminal A	pos	pos	neg	neg
9p loss (0 Mb–25 Mb)	Normal-like	N/A	N/A	N/A	N/A

pos: significant association with positive immunohistochemistry (IHC), neg: significant association with negative IHC, N/A is used when no significant association was used.
[a]This association was identified as borderline significant, but was the most significant association found for PR-negative status.
[b]As determined by immunohistochemistry.

discretized version of the Chin *et al.* dataset, kindly provided by J. Fridyland. These data have been discretized with discretization algorithms, such as CBS (25) and MergeLevel (13). The results of our comparison are shown in Figure 4.

Figure 4a shows a genome-wide view of the frequency of alterations compared to the KC-SMART results for a single kernel width. Since most researchers use a frequency cutoff between 25 and 35% to define interesting regions (31,34–36), we chose a cutoff of 30% to define significant regions in the frequency analysis. The right panel shows the result for chromosome 17. It can be seen that by employing a smaller kernel width, KC-SMART detects the ERBB2 amplification, which is not picked up by the frequency analysis. The overlap between the number of probes in the significant regions for KC-SMART ($P < 0.05$) and the number of probes in regions aberrated in $\geq$30% of the tumors is shown in Figure 4b. The Venn diagrams show that there is a large overlap in the copy number losses identified by both methods, but that the frequency method flags a larger number of probes as gained. To determine whether KC-SMART underestimates the number of significant regions or whether the frequency method overestimates the number of aberrated regions, we used the aCGH data to assess the 'quality' of the identified regions. More specifically, we computed, for each identified consecutive region, all the pair-wise correlation coefficients of clones within that region. Our hypothesis is that a higher degree of correlation will be indicative of a better delineation of the aberration. The smoothed histograms for both KC-SMART and the frequency approach are depicted in Figure 4c, As can be seen, the within-region correlation coefficients are, on average, much higher in the gained regions detected by KC-SMART than in the regions detected by the frequency approach. The frequency approach performs similar to KC-SMART when looking at the lost regions.

To our best knowledge, the only other readily available method that incorporates a statistical framework for multiexperiment aCGH analysis is STAC (17). This method employs two statistics, the frequency statistic and the footprint statistic, to identify aberrant regions. STAC requires aCGH data to be discretized to contain only 1 (gain), −1 (loss) and 0 (no change). We analyzed the same 89-sample breast cancer dataset using STAC. Results are shown in the Supplementary Data. When comparing the regions found significant by both methods it is striking that STAC does not identify the 1q arm gain or the 17q *ERBB2* amplification as significant. Significant regions found by STAC seem to aggregate at boundaries

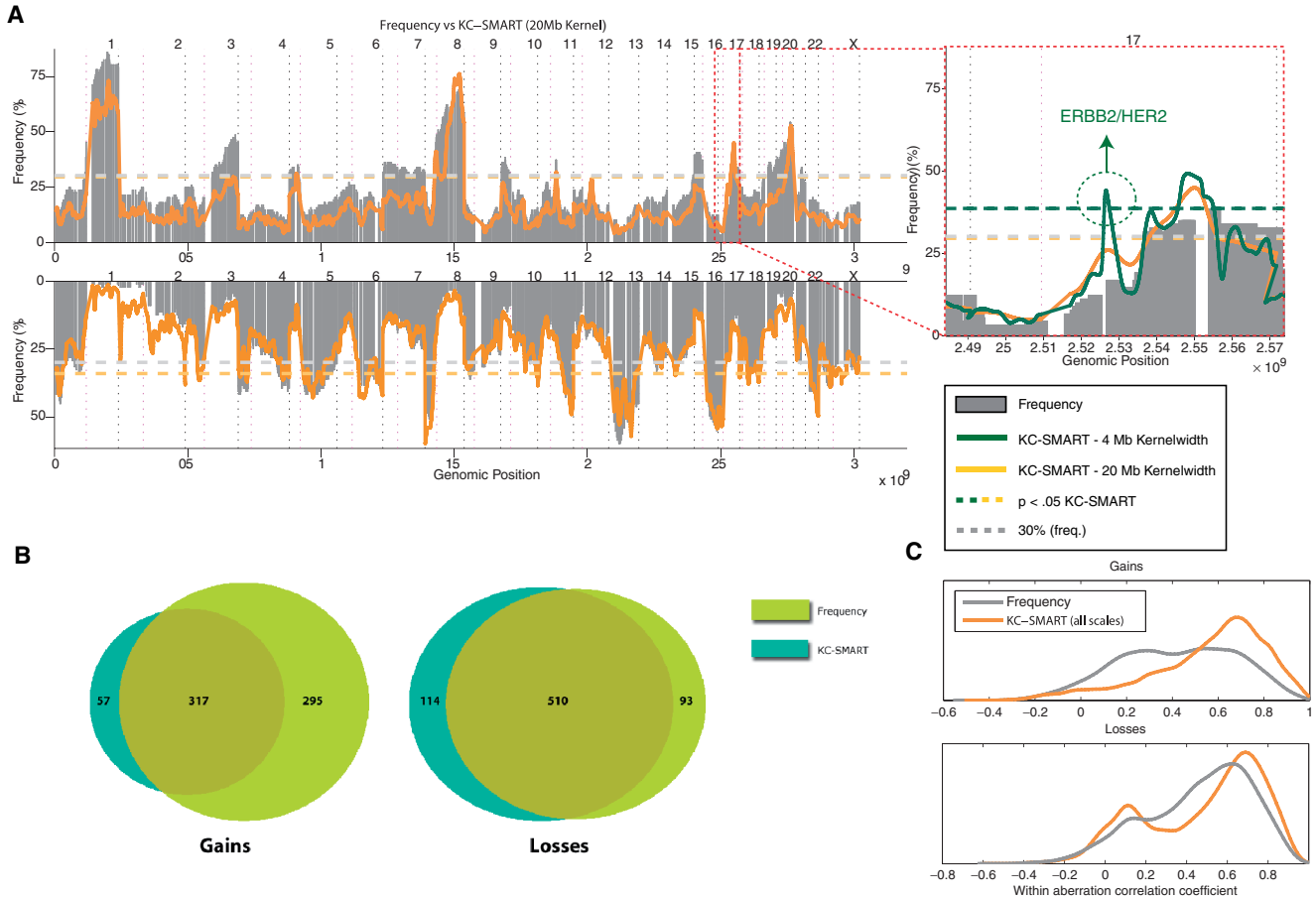

Figure 4. Comparison of KC-SMART to frequency-based analysis. (**a**) A genome-wide frequency analysis of copy-number changes is show (gray bars). On top of the frequency analysis the KC-SMART result for 20-Mb kernelwidth is plotted (orange line). The dataset used for analysis is the 89 breast tumor data published by Chin *et al*. The significance threshold for KC-SMART is shown as an orange dotted line. The 30% frequency level has been shown as a gray dotted line. The zoom-panel shows a magnification of chromosome 17. Here the result for KC-SMART at 4 Mb is shown in green. The green dotted line shows the significant threshold for 4-Mb kernelwidth. (**b**) Proportional Venn diagrams showing the overlap between results from both KC-SMART and the frequency analysis. Overlap is determined on the basis of probes in significant regions (KC-SMART) or in regions over 30% frequency. (**c**) Smoothed histograms of within-region BAC pair correlation coefficients.

of chromosome arms. This could point to a bias in the algorithm for boundary regions. To investigate the nature of the significant regions detected by STAC but not by KC-SMART, we analyzed the raw log2 values and compared them to the discretized values produced by Chin *et al*. (22). Two examples of this analysis are shown in the Supplementary Data. Generally, the areas exclusively detected by STAC are scored as single copy alterations by the MergeLevels analysis done by Chin *et al*. (22). On a log2 scale these gains and losses were very small indeed (~0.1), but nevertheless large enough to be assigned a value of 1 or −1, and are therefore treated equally to large gains and losses. This discrete approach, which neglects the amplitude of an aberration can therefore be considered to be less conservative than KC-SMART and can/may therefore yield more false-positives.

Cancer gene discovery using KC-SMART

To identify genes that are associated with CNA and potentially implicated in tumor formation we applied KC-SMART to a dataset containing aCGH measurements of 19 mammary carcinomas acquired from conditional p53 knockout mice. Sporadic mammary tumor development in these mice is induced by cell type-specific stochastic inactivation of the tumor suppressor gene *p53* (37). Mouse mammary tumors present with relatively few passenger mutations as compared to human sporadic tumors. This is because tumors from genetically engineered mouse models have a more uniform cell type of origin and they develop relatively rapid giving less time for passenger CNA development. This will aid in identification of causal aberrations and will reduce the number of tumors required for cancer gene discovery. Mouse aCGH data were preprocessed as described in the Methods section. The result of the analysis is shown in Figure 5. In total, seven amplified and 11 deleted regions were identified by KC-SMART.

The significant CNAs identified by KC-SMART are large and contain many genes. Oncogenic aberrations typically occur because it is the mechanism through which an oncogene is amplified or a tumor suppressor

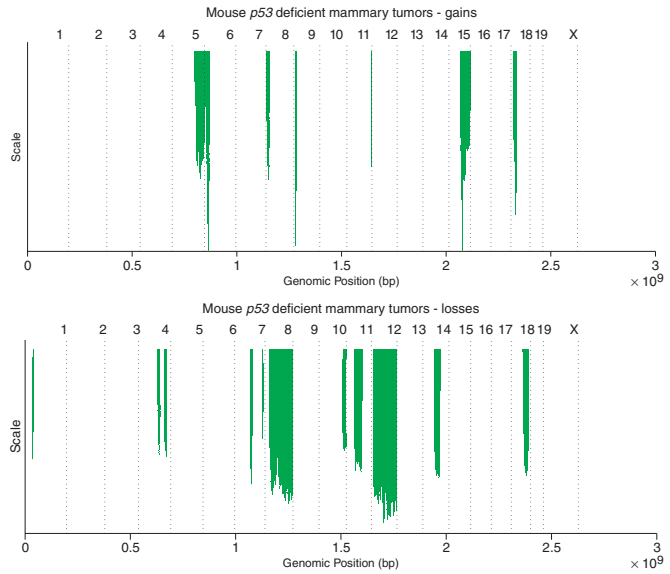

Figure 5. KC-SMART results on the p53-deficient mouse model mammary tumors. The *y*-axis represents the interpolated scale space running from 2 to 40 Mb. Black dotted lines represent the end of chromosomes.

is inactivated. We examined the gene expression profiles of all tumors to find the genes which are driving the aberrations. We used an in-house 32-k oligonucleotide microarray platform with 31 769 longmer probes targeting 19 104 unique genes, which is 68% of the known mouse protein-coding genes. By analyzing only the genes that are measured on the array, we accept the possibility of missing the driver gene because there is no corresponding probe on the array. Furthermore, the gene expression data were obtained using a pool of 12 mouse mammary tumor samples as a common reference. While these measurements will reveal the within-group variability of the tumor dataset, genes that are over-expressed or downregulated in all tumors will not be picked up, because their RNA levels in the reference pool are the same. The complete results of these analyses can be found in the Supplementary Data. As can be seen in Table 5 and Figure 6 the correlation analysis resulted in association of several known and novel cancer genes with significant recurrent CNAs. Our correlation analysis resulted in the association of three known cancer genes with gained regions (*Ptpn11*–Chr. 5, *c-Met*–Chr. 6 and *Birc2*–Chr. 9) and one known cancer gene within a deleted region (*Cyld*–Chr. 8). The mouse homologs from the CGC were used as 'known' cancer genes. The identification of known cancer genes validates this approach for cancer gene discovery.

Novel candidate cancer gene discovery

Correlation analysis can also be used for the identification of new cancer gene candidates. In addition to the purely data-driven approach, the candidate genes were further analyzed based on known functionality and possible links to carcinogenesis. The possibility exists that a gene with an unknown function could be a driver gene, and it will be missed by this approach. Keeping this in mind, one should reconsider these genes when no obvious candidate is found, or when validation experiments do not show any causal role for the identified candidate genes.

Table 5. Overview of the regions found by KC-SMART analysis of the mouse conditional p53 knockout dataset

Chromosome	Position (Mb)	Number of correlating genes	Correlating cancer genes[1]	Candidate cancer genes
Gains				
5	100–151	48	PTPN11	–
6	3–26	9	C-MET	–
8	3–15	22	–	CDC16, TPDF1, CUL4A, FBOXO25
9	6–15	7	BIRC2	–
11	114–118	9	–	BIRC5
15	52–102	0	–	–
18	10–26	7	–	RNF138
Losses				
1	37–40	2	–	GNT-IVA
4	91–103	0	–	–
4	122–133	0	–	–
7	72–85	4	–	–
7	131–134	3	–	BUB3, BCCIP
8	23–131	37	CYLD	–
10	112–129	10	–	USP15
11	37–75	29	–	HINT1, RAD50, IRF-1
12	4–117	10	–	NUMB
14	54–83	1	–	–
18	55–83	1	–	–

The positions mentioned are based on the maximal width of the aberration found across all scales. Genes that either have a start or end position in the region were taken into account. Only genes against which an oligonucleotide probe was present on the gene expression micro-array platform were considered. The last column shows new likely cancer gene candidates, based on the result of the correlation analysis (Pearson correlation, $P < 0.05$, Bonferroni–corrected) and based on functional analysis. When no likely candidates were found in the correlating genes based on functional analysis, none were listed.

Chromosome 8 gain. None of the 22 genes identified by correlation analysis are listed in the CGC. Based on previous studies we identified four possible candidates: *Cdc16*, *Tfdp1*, *Cul4a* and *Fboxo25* (38). The human homolog of *Cdc16*, a subunit of the anaphase promoting complex which governs degradation of G1 checkpoint proteins, has also been linked to cancer (39). *Tfdp1* is a transcription factor that operates in concert with the E2F family to upregulate genes associated with cell cycle progression (40). *Cul4a* is a ubiquitin ligase that targets the cyclin-dependent kinase inhibitor *p27* for degradation and thereby *Cul4a* activity promotes proliferation and cell cycle progression (41). *Fboxo25* is a member of the F-box only protein family (42), which is well known for their ubiquitin-ligase function.

Chromosome 18 gain. None of the seven genes identified by correlation analysis are listed in the CGC. Only one gene was found to have interesting functional annotation: *Rfn138*. The human homolog of *Rnf138*, dubbed *NARF*, has recently been identified as a ubiquitin ligase targeting

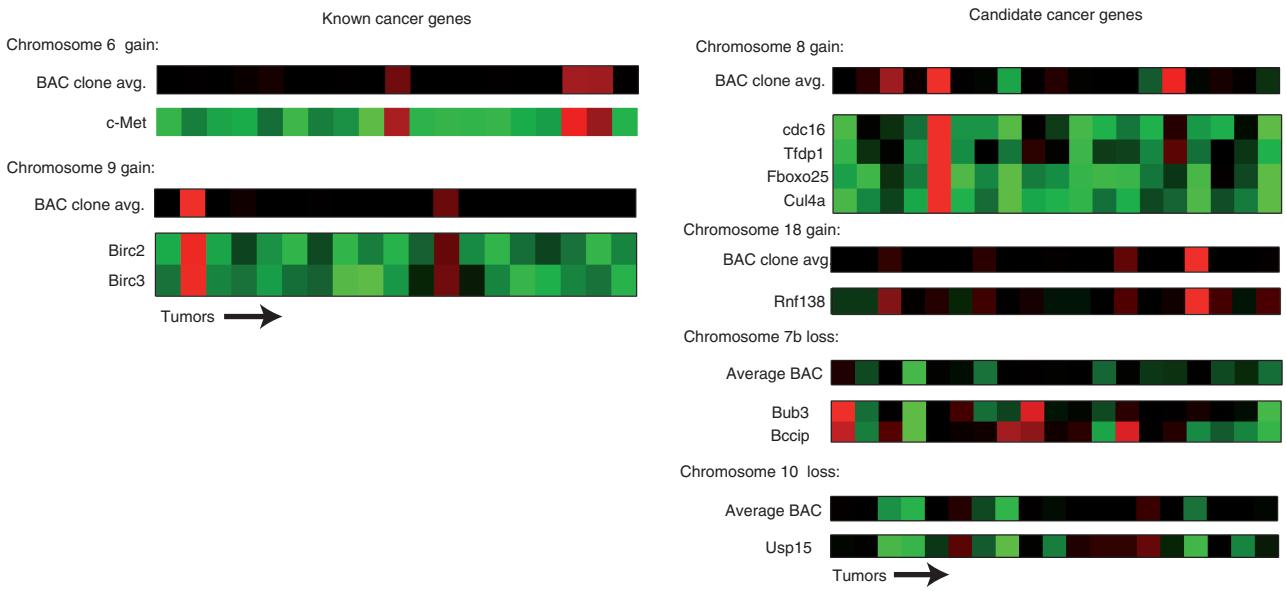

Figure 6. Expression and genomic profiles of the known cancer genes and the cancer gene candidates discovered in p53-deficient mouse mammary tumors. For each aberration the chromosome number is given as well as the average aberration BAC profile of the BAC clones in the aberrant region. The gene expression profiles of the genes that were selected based on their correlation with the BAC profile are depicted below each BAC profile. Green indicates downregulation/loss, red indicates overexpression/gain.

the *TCF/LEF* complex, thereby downregulating the Wnt-signaling target genes (43). Derepression of *Wnt* target genes due to loss of *TCF/LEF* might promote tumorigenesis.

Chromosome 7 loss. Of the three genes found by correlation analysis, two were interesting candidate cancer genes. *Bub3* is a mitotic spindle checkpoint gene that can act as a repressor of cell cycle progression (44). Deletion of this gene could enhance cell cycle progression and thus proliferation. *Bccip* is a *BRCA2* interacting protein and has been found to be important in double-strand break repair (45). Loss of *Bccip* might lead to genomic instability and thus enable tumorigenic aberrations.

Chromosome 10 loss. Of the 10 genes found by correlation analysis, only one gene, *Usp15*, had functional annotation that could be related to tumor suppressor activity. *Usp15* has recently been found to be present in the *COP9* signalosome, which is an important regulator of ubiquitin ligase activity and has been implicated in cancer (46,47). It should be noted that the chromosome 10 region includes several cancer genes, including the known tumor suppressor *Ddit3*, also known as *CHOP*. *CHOP* expression correlated just below the significance threshold set for correlation analysis.

Scale space aided aberration analysis

Our correlation analysis did not yield promising cancer genes for each region. For example, the chromosome 15 copy number gain is a large aberration found by KC-SMART and spans around 50 Mb over all scales analyzed. Yet, no cancer gene candidates were identified using the correlation approach. Since the chromosome 15 gain is analogous to the human chromosome 8q, we expect the oncogene *Myc* to be the target of the amplification. The *Myc* oncogene is a well-known target of DNA amplification in human cancer (48,49). We therefore set out to analyze the chromosome 15 gain with the aid of the scale space. The scale space of mouse chromosome 15 identifies a large region as significant in the large scales, yet only a small region remains significant in the smaller scales (Figure 7a). If this region is analyzed in more detail (Figure 7b) one can see several genes situated next to the narrow significant region. The two genes that directly map onto the significant region are both predicted transcripts with unknown function. The *Myc* oncogene is situated slightly upstream from the significant region. If the heatmaps of the BAC clones are examined (Figure 7d) it is apparent that BAC Clone 7 shows a gain in multiple tumors. This could indicate that amplification of the region downstream of Myc has a tumorigenic effect, either by affecting *Myc* in some way or acting on a previously unknown gene. It could also mean that the chromosome 15 gain is not directed at the *Myc* oncogene, but rather at the *Ddef1* gene. *Ddef1* has been identified as a pathologically relevant target of amplification in human uveal melanoma (50). This result shows that scale-space analysis can be effectively used to gain more insight in complex aberrations and to identify potential target genes.

Besides giving more insight in the nature of aberrations, the scale space is also used to identify both small focal aberrations and large low level CNAs such as chromosomal gains or losses. The smaller scales allow for identification of high-level amplifications, such as the ERBB2/HER2 amplification in human breast

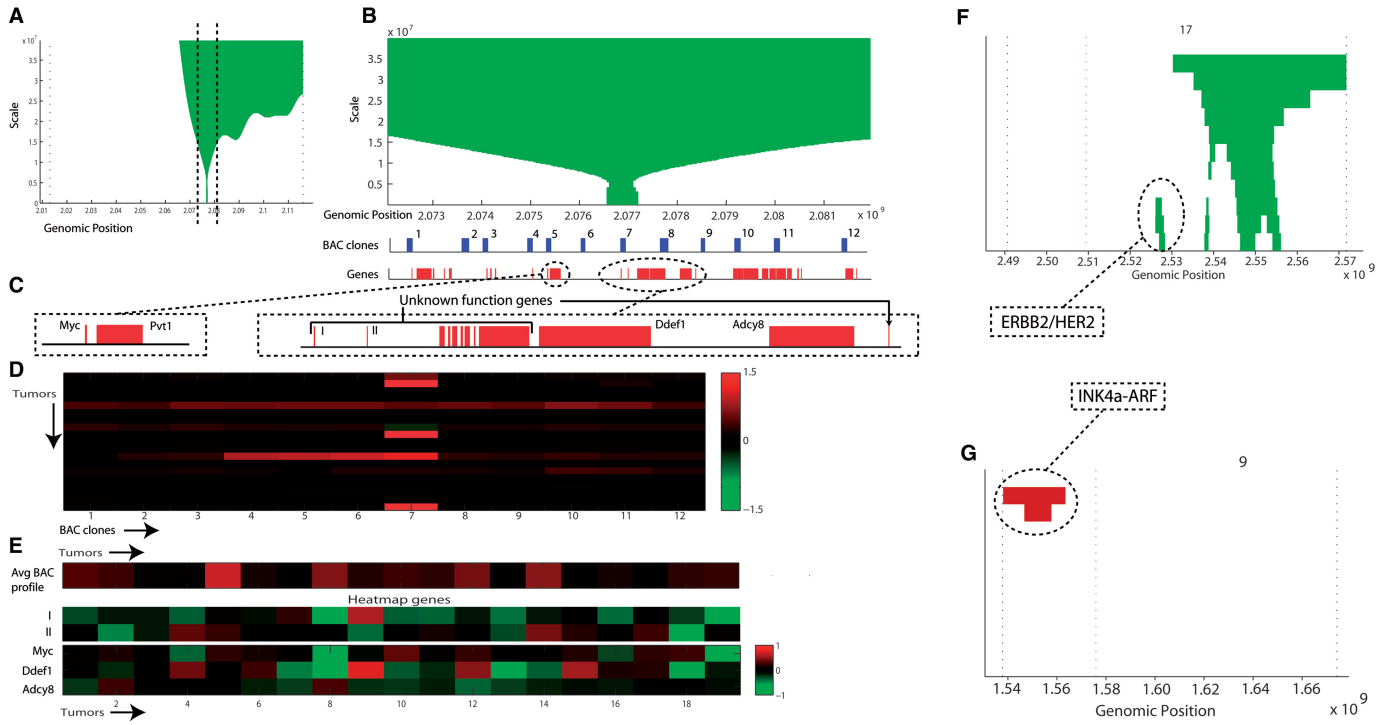

Figure 7. Scale-space analysis of the chromosome 15 aberration. (**a**) Scale-space analysis of the complete chromosome. The outer black dotted lines denote the end of the chromosome. The inner black dotted lines denote the area shown in (**b**). (b) Zoom-in view of the scale-space analysis of chromosome 15. The BAC clones that are mapped to this region are shown as blue blocks. The genes that are situated in this region are depicted as red blocks. (**c**) Genes close to the region that is significant across all scales are shown in more detail. (**d**) This figure shows the heatmap of the BAC-clones shown in (b). Numbers along the horizontal axis correspond to the BAC clone numbers in (c). Positive log2 values as shown as red, negative log2 values as green. (**e**) Heatmap of the gene expression of genes *Myc*, *Ddef1* and *Adcy8*. Note: The tumors are now depicted along the horizontal direction, as opposed to (d), where the tumors are depicted in the vertical direction. Positive log2 values as shown as red, negative log2 values as green. No probe against *Pvt1* was present on the gene expression array. The two unknown-function transcripts overlapping with BAC clone 7 show equally uncorrelated expression profiles, and are denoted by I and II. (**f**) This figure shows a scale-space analysis of significant gains on chromosome 17. The analysis is from a set of 89 human breast tumors. (**g**) A scale-space analysis of chromosome 9 losses. The analysis is from a set of 89 human breast tumors.

cancer (Figure 7f). This small, but strong amplification is found to be significant in the smaller scales, but it is smoothed away in the larger scales. Conversely, a large systemic loss of chromosome 9p containing the *INK4a/ARF* tumor suppressor locus is only detected as significant in the highest two scales analyzed (Figure 7g). By traversing the scale space we can find biologically relevant aberrations of all sizes.

DISCUSSION

Genomic instability is a powerful, yet chaotic way of acquiring aberrations that confer proliferative or survival advantages to a tumor cell. Correctly separating driver mutations from passenger mutations in aneuploid tumors is very difficult. The classical tumor model assumes clonal expansion of a tumor from a single progenitor cell; however, recent views on tumor development preferentially model the tumor as an ecological model, with different populations of tumor cell clones competing for survival (19). In such a diverse population of tumor cells, CNA should also be diverse. As an aCGH measurement is an averaged view of all CNAs within a large heterogenic population of tumor cells, discretization of these values to integer levels of copy number change may result in loss of information, which is exemplified by the fact that the level of ERBB2 amplification is a determinant of breast cancer progression (51). KC-SMART is the first method to use nondiscretized data to find statistically significant recurrent CNAs. KC-SMART uses the log2 values straight from the aCGH platform and is therefore not dependent on preprocessing steps, except for data-normalization. We show that relevant biological data are contained in the continuous log2 aCGH measurements. So, instead of noise reduction, discretization may well cause information loss. We think that approaches that discretize the aCGH data prior to detecting CNAs are likely to perform very well in clonal systems, such as cell lines and hematological tumors. KC-SMART is preferable for analysis of aCGH data from more heterogeneous solid tumor samples.

Tumor tissues are challenging biological samples and their copy number analysis is confounded by several factors. Stromal contamination will quench the signal by introducing normal *2n* DNA in the sample. Very pure (clonal) tumors will have a larger effect on the analysis as their overall measured copy number levels are higher. Both discretization approaches and KC-SMART will be

affected by sample inhomogeneity. While KC-SMART will employ the measured aberration level in the analysis, discretization approaches will, depending on the chosen discretization threshold, either fully count an aberration (regardless of the degree to which it exceeds the threshold) or completely discard a probe when the CNA signal does not exceed the threshold defining an aberration. Very large focal amplifications will be maintained in the KC-SMART analysis, but if a sufficiently large set of tumors is analyzed, these aberrations will not be flagged as significant, unless they appear with sufficient frequency.

Random signal noise is always a confounding factor in aCGH data analysis. Especially formalin fixed paraffin embedded (FFPE) samples are known to produce noisy measurements, making aCGH data from these samples hard to analyze. Since KC-SMART analyzes gains and losses separately, the possibility exists that noise which randomly fluctuates around zero will be detected as aberrations due to the separation of data in positive and negative values. However, this is countered in the random permutation scheme where all values including random fluctuations are incorporated in the construction of the background distribution. By determining a null-distribution for each dataset analyzed, noise specific to that dataset is modeled and therefore also controlled. To gain the most power from an analysis we encourage researchers to keep comparable noise levels across the entire dataset.

We showed association of recurrent aberrations found by KC-SMART with molecular subtype (30) or clinico-pathological features such as *p53*, *ERBB2*, estrogen receptor (*ER*) and progesterone receptor (*PR*) expression status. Identification of the previously reported 16p and 5q regions show that known relations are recapitulated by KC-SMART, and that these regions might be markers for the Luminal A and basal-like breast cancer subtypes, respectively. The 5q loss is an important feature employed in a classifier to separate *BRCA1*-mutated breast cancers from control tumors (52). *BRCA1*-mutated breast tumors are known to be generally basal-like (53), and the fact that aberrations on chromosome 5q can distinguish the *BRCA1* tumors from control tumors is possibly due to their basal-like character. A *BRCA1*-modifier locus for hereditary breast cancer has been mapped to 5q (54) and multiple DNA-damage repair and cell cycle associated genes are present in the 5q region (e.g. *RAD50*, *RAD17*, *APC*). This finding might point to a more general DNA damage-related phenotype in basal breast tumors. Other relations with regions identified by KC-SMART are new, such as the association of 9p loss with normal-like tumors, which could target the well-known *INK4a/ARF* tumor suppressor locus located on 9p21. In a large clinical study, normal-like tumors were found to be mainly diploid, genomically stable tumors (33). In line with this, mouse mesotheliomas induced by *Nf2* and *Ink4a/Arf* loss show significantly less genomic instability than mouse mesotheliomas induced by *Nf2* and *p53* loss (van Montfort,E. and Berns,A., unpublished data). Normal-like tumors may thus be driven by *INK4a-ARF* loss-of-function.

Array CGH data analysis can result in detection of large regions containing many genes. The application of KC-SMART in combination with gene expression micro-array data facilitates prioritization of genes for biological validation. While there are many (epi)genetic mechanisms through which a tumor can upregulate oncogenes or downregulate tumor suppressor genes, it has been shown that CNAs can predict gene expression (55). Nevertheless, one should be aware of the risk of missing important genes that do not seem to correlate very well with the copy number data.

Our results obtained from the KC-SMART analysis combined with gene expression data of the *p53* conditional knockout mice did reveal some well-known cancer genes that have been associated with CNAs before. Two of the seven genes that are identified by the correlation analysis on chromosome 9 were previously implicated in tumorigenesis: *Birc2* and *Birc3* (56). *Birc2* is also listed in the CGC list. Both are members of the *cIAP* family of antiapoptotic caspase-binding proteins. Zender *et al*. (56) demonstrated the tumorigenic effect of *Birc2* in *Myc* overexpressing cells. Their comparative oncogenomics approach uses CNA information to find potential oncogenes and tumor suppressors. As human data are inherently complex, cross-species analysis can provide additional power to identify truly causal genes. CNA detection is an increasingly valued tool in cross-species oncogenomics (57). KC-SMART would be ideally suited to provide the first analysis of CNAs in a comparative oncogenomics study.

The examples of *Birc2/3* and other known cancer genes identified by correlation analysis show that biologically relevant results can be obtained using this approach. However, not every region delivered a promising cancer gene candidate after correlation analysis. Especially, the lost regions did not produce many obvious candidates, because the correlation analysis either delivered very many or very few genes. This could point to the fact that a tumor cell might be more inclined to use epigenetic factors to downregulate tumor suppressor genes so that the correlation between expression and copy number values would be disturbed. It could also mean that a deletion is more suited to incur haploinsuffiency of large numbers of genes, which may collectively contribute to tumor development (58). In contrast, amplification is more likely to result in overexpression of one or a few genes which are therefore more likely to be detected by correlation analysis.

The scale space provided by KC-SMART facilitates analysis of complex aberrations. As our scale-space analysis of the chromosome 15 aberration in the mouse *p53*-deficient mammary carcinoma dataset shows, true targets of gains and losses can be determined. Of course, this analysis is spatially restricted to the resolution of the aCGH platform and quantitatively restricted to the degree with which individual probes on the microarray report CNAs. We foresee that the scale-space analysis will allow a very insightful and accurate determination of aberrant regions across multiple tumors.

We sought to compare our results to those obtained by existing methods. First we compared our approach to

a frequency-based analysis, which is still often used to analyze aCGH data. We used the discretized version of the Chin *et al*. data, as discretized by the authors, using two different algorithms to obtain segmented and called data. Our method compares favorably to this more complicated approach. One intrinsic problem associated with frequency analysis is that it is mainly descriptive. There is no statistical basis for calling a region significantly recurrent aberrated. KC-SMART provides the researcher with a solid statistical analysis, while frequency-based thresholds will be arbitrarily chosen. Furthermore, KC-SMART has the added advantage of incorporating a scale-space analysis. The advantages of this feature have already been clearly demonstrated in the example where KC-SMART can readily identify high-level amplicons, while frequency-based methods need additional adaptations to detect these amplifications.

To the best of our knowledge, the only method that encompasses a statistical framework and has readily available software to perform multiexperiment aCGH data analysis is STAC (17). Several remarks can be made on the comparison between STAC and KC-SMART. STAC results are dependent on the manner in which the two STAC statistics are combined. Results for both the statistics (footprint and frequency) can be very different. This disparity complicated the interpretation and might lead to results-oriented use of statistics. Furthermore, STAC bases the background-distribution of their statistical tests on local chromosome-arm permutation. This means that recurrent loss or gain of whole-chromosome arms cannot be detected. A prime example of this is the well-known gain of human 1q which was not identified as statistically significant by STAC. Also, basing the background distribution just on the chromosome arm is not a correct representation of the biological reality, where the exact position of a gain or amplification is not known and could be located on a completely different chromosome arm. To get a better average background distribution, permutation over the entire genome is preferable.

In summary, we have developed the first parameter-free statistical method to find significantly recurrent CNAs from continuous aCGH measurements. In contrast to other approaches, KC-SMART inputs nondiscretized data, which minimizes loss of information and enables better modeling of the continuous-valued amplitude of CNAs. KC-SMART employs a permutation approach to detect statistically significant CNAs while keeping control of the error rate. A scale space is constructed which facilitates detailed inspection of CNAs at a range of genomic resolutions. KC-SMART compares favorably to existing aCGH analysis methods and produces biologically relevant results. Furthermore, using KC-SMART we identify several new candidate cancer genes. DNA is the static foundation of the dynamic cellular environment. Tumor cells often resort to modification of the DNA content of the cell to further their proliferative and survival strength. The ability to detect these changes will help us understand the genetic basis of cancer, and it allows us to prioritize our subsequent research into causal components of tumor cells.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank the following persons for critically reading the manuscript: E. van Beers, P. Derksen, B. Evers and K. de Visser. We would like to thank J. Fridlyand for kindly providing us with the CBS and MergelLevel data of the 89 human breast tumor samples. This work was supported by grants from the Netherlands Organization for Scientific Research (ZonMw 917.036.347), the Dutch Cancer Society (NKI 2002-2635) and the Susan G. Komen Breast Cancer Foundation (BCTR 0403230). Funding to pay the Open Access publication charges for this article were provided by the Dutch Cancer Society (NKI 2002-2635).

Conflict of interest statement. None declared.

REFERENCES

1. Hanahan,D. and Weinberg,R. (2000) The hallmarks of Cancer. *Cell*, **100**, 57–70.
2. Myllykangas,S. (2006) Manifestation, mechanisms and mysteries of gene amplifications. *Cancer Lett*., **232**, 79–89.
3. Kallioniemi,A., Kallioniemi,O.P., Sudar,D., Rutovitz,D., Gray,J.W., Waldman,F. and Pinkel,D. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.
4. Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet*., **20**, 207–211.
5. Pollack,J.R., Perou,C.M., Alizadeh,A.A., Eisen,M.B., Pergamenschikov,A., Williams,C.F., Jeffrey,S.S., Botstein,D. and Brown,P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet*., **23**, 41–46.
6. Lindblad-Toh,K., Tanenbaum,D.M., Daly,M.J., Winchester,E., Lui,W.O., Villapakkam,A., Stanton,S.E., Larsson,C., Hudson,T.J. *et al*. (2000) Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotech*., **18**, 1001–1005.
7. Lucito,R., Healy,J., Alexander,J., Reiner,A., Esposito,D., Chi,M., Rodgers,L., Brady,A., Sebat,J. *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res*., **13**, 2291–2305.
8. Brennan,C., Zhang,Y., Leo,C., Feng,B., Cauwels,C., Aguirre,A.J., Kim,M., Protopopov,A. and Chin,L. (2004) High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res*., **64**, 4744–4748.
9. Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet*., **37**, 11–17.
10. Shayesteh,L., Lu,Y., Kuo,W.L., Baldocchi,R., Godfrey,T., Collins,C., Pinkel,D., Powell,B., Mills,G.B. *et al*. (1999) PIK3CA is implicated as an oncogene in ovarian cancer. *Nat. Genet*., **21**, 99–102.
11. Albertson,D.G., Ylstra,B., Segraves,R., Collins,C., Dairkee,S.H., Kowbel,D., Kuo,W.L., Gray,J.W. and Pinkel,D. (2000) Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat. Genet*., **25**, 144–146.
12. Lai,W.R., Johnson,M.D., Kucherlapati,R. and Park,P. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.
13. Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.

14. van de Wiel,M.A., Kim,K.I., Vosse,S.J., van Wieringen,W.N., Wilting,S.M. and Ylstra,B. (2007) CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, **23**, 892–894.
15. Huang,J., Gusnanto,A., O'Sullivan,K., Staaf,J., Borg,A. and Pawitan,Y. (2007) Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics*, **23**, 2463–2469.
16. Rouveirol,C., Stransky,N., Hupe,P., Rosa,P., Viara,E. and Barillot,E. (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849–856.
17. Diskin,S.J., Eck,T., Greshock,J., Mosse,Y.P., Naylor,T., Stoeckert,C.J. Jr, Weber,B.L. and Maris,J.M. (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res.*, **16**, 1149–1158.
18. Fiegler,H., Geigl,J.B., Langer,S., Rigler,D., Porter,K., Unger,K., Carter,N.P. and Speicher,M.R. (2007) High resolution array-CGH analysis of single cells. *Nucleic Acids Res.*, **35**, e15.
19. Merlo,L.M.F., Pepper,J.W., Reid,B.J. and Maley,C.C. (2006) Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*, **6**, 924–935.
20. Parzen,E. (1962) On estimation of a probability density function and mode. *Ann. Math. Stat.*, **33**, 1065–1076.
21. Atkeson,C.G., Andrew W Moore, and Stefan Schaal, (1997) Locally weighted learning. *Artif. Intell. Rev.*, **V11**, 11–73.
22. Chin,K., DeVries,S., Fridlyand,J., Spellman,P.T., Roydasgupta,R., Kuo,W.L., Lapuk,A., Neve,R.M., Qian,Z. *et al.* (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**, 529–541.
23. Chung,Y.J., Jonkers,J., Kitson,H., Fiegler,H., Humphray,S., Scott,C., Hunt,S., Yu,Y., Nishijima,I. *et al.* (2004) A whole-genome mouse BAC microarray with 1-Mb resolution for analysis of DNA copy number changes by array comparative genomic hybridization. *Genome Res.*, **14**, 188–196.
24. Dennis,G., Sherman,B., Hosack,D., Yang,J., Gao,W., Lane,H. and Lempicki,R. (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, R60.
25. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
26. de Ridder,J., Uren,A., Kool,J., Reinders,M.J.T. and Wessels,L.F.A. (2006) Detecting Statistically Significant Common Insertion Sites in Retroviral Insertional Mutagenesis Screens. *PLoS Comput. Biol.*, **2**, 1530–1542.
27. Kallioniemi,A., Kallioniemi,O., Piper,J., Tanner,M., Stokke,T., Chen,L., Smith,H.S., Pinkel,D., Gray,J.W. and Waldman,F.M. (1994) Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc. Natl Acad. Sci. USA*, **91**, 2156–2160.
28. Loo,L.W.M., Grove,D.I., Williams,E.M., Neal,C.L., Cousens,L.A., Schubert,E.L., Holcomb,I.N., Massa,H.F., Glogovac,J. *et al.* (2004) Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. *Cancer Res.*, **64**, 8541–8549.
29. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
30. Perou,C.M., Sorlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.
31. Bergamaschi,A., Kim,Y.H., Wang,P., Sørlie,T., Hernandez-Boussard,T., Lonning,P.E., Tibshirani,R., Børresen-Dale,A.-L. and Pollack,J. (2006) Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer*, **45**, 1033–1040.
32. Nielsen,T.O., Hsu,F.D., Jensen,K., Cheang,M., Karaca,G., Hu,Z., Hernandez-Boussard,T., Livasy,C., Cowan,D. *et al.* (2004) Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin. Cancer Res.*, **10**, 5367–5374.
33. Calza,S., Hall,P., Auer,G., Bjohle,J., Klaar,S., Kronenwett,U., Liu,E., Miller,L., Ploner,A. *et al.* (2006) Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Res.*, **8**, R34.
34. Naylor,T., Greshock,J., Wang,Y., Colligon,T., Yu,Q.C., Clemmer,V., Zaks,T. and Weber,B. (2005) High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res.*, **7**, R1186–R1198.
35. van Beers,E. and Nederlof,P. (2006) Array-CGH and breast cancer. *Breast Cancer Res.*, **8**, 210.
36. Nakao,K., Mehta,K.R., Fridlyand,J., Moore,D.H., Jain,A.N., Lafuente,A., Wiencke,J.W., Terdiman,J.P. and Waldman,F.M. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, **25**, 1345–1357.
37. Jonkers,J., Meuwissen,R., van der Gulden,H., Peterse,H., van der Valk,M. and Berns,A. (2001) Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat. Genet.*, **29**, 418–425.
38. Yasui,K., Arii,S., Zhao,C., Imoto,I., Ueda,M., Nagai,H., Emi,M. and Inazawa,J. (2002) TFDP1, CUL4A, and CDC16 identified as targets for amplification at 13q34 in hepatocellular carcinomas. *Hepatology*, **35**, 1476–1484.
39. Wang,Q., Moyret-Lalle,C., Couzon,F., Surbiguet-Clippe,C., Saurin,J.C., Lorca,T. and Navarro,C. (2003) Alterations of anaphase-promoting complex genes in human colon cancer cells. *Oncogene*, **22**, 1486–1490.
40. Bandara,L.R. (1991) Adenovirus E1a prevents the retinoblastoma gene product from complexing with a cellular transcription factor. *Nature*, **351**, 494–497.
41. Li,B., Jia,N., Kapur,R. and Chun,K. (2006) Cul4A targets p27 for degradation and regulates proliferation, cell cycle exit, and differentiation during erythropoiesis. *Blood*, **107**, 4291–4299.
42. Hagens,O., Minina,E., Schweiger,S., Ropers,H. and Kalscheuer,V. (2005) Characterization of FBX25, encoding a novel brain-expressed F-box protein. *Biochim. Biophys. Acta*, **1760**, 110–118.
43. Yamada,M., Ohnishi,J., Ohkawara,B., Iemura,S., Satoh,K., Hyodo-Miura,J., Kawachi,K., Natsume,T. and Shibuya,H. (2006) NARF, an Nemo-like kinase (NLK)-associated ring finger protein regulates the ubiquitylation and degradation of T Cell Factor/Lymphoid Enhancer Factor (TCF/LEF). *J. Biol. Chem.*, **281**, 20749–20760.
44. Yoon,Y.M., Baek,K.H., Jeong,S.J., Shin,H.J., Ha,G.H., Jeon,A.H., Hwang,S.G., Chun,J.-S. and Lee,C.-W. (2004) WD repeat-containing mitotic checkpoint proteins act as transcriptional repressors during interphase. *FEBS Lett.*, **575**, 23–29.
45. Lu,H., Guo,X., Meng,X., Liu,J., Allen,C., Wray,J. and Nickoloff,J.A. (2005) The BRCA2-interacting protein BCCIP functions in RAD51 and BRCA2 focus formation and homologous recombinational repair. *Mol. Cell. Biol.*, **25**, 1949–1957.
46. Richardson,K.S. (2005) The emerging role of the COP9 signalosome in cancer. *Mol. Cancer Res.*, **3**, 645–653.
47. Hetfeld,B.K.J., Helfrich,A., Kapelari,B., Scheel,H., Hofmann,K., Guterman,A., Glickman,M., Schade,R., Kloetzel,P. *et al.* (2005) The zinc finger of the CSN-associated deubiquitinating enzyme USP15 Is essential to rescue the E3 ligase Rbx1. *Curr. Biol.*, **15**, 1217–1221.
48. Collins,S. and Groudine,M. (1982) Amplification of endogenous myc-related DNA sequences in a human myeloid leukaemia cell line. *Nature*, **298**, 679–681.
49. Favera,R.D., Wong-Staal,F. and Gallo,R.C. (1982) onc gene amplification in promyelocytic leukaemia cell line HL-60 and primary leukaemic cells of the same patient. *Nature*, **299**, 61–63.
50. Ehlers,J.P., Worley,L. and Onken,M.D. (2005) DDEF1 is located in an amplified region of chromosome 8q and is overexpressed in uveal melanoma. *Clin. Cancer Res.*, **11**, 3609–3613.
51. Slamon,D.J., Clark,G.M., Wong,S.G., Levin,W.J., Ullrich,A. and McGuire,W.L. (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, **235**, 177–182.
52. Wessels,L.F.A., van Welsem,T., Hart,A.A.M., van't Veer,L.J., Reinders,M.J.T. and Nederlof,P.M. (2002) Molecular classification of breast carcinomas by comparative genomic hybridization: a specific somatic genetic profile for BRCA1 tumors. *Cancer Res.*, **62**, 7110–7117.

53. Sorlie,T., Tibshirani,R., Parker,J., Hastie,T., Marron,J.S., Nobel,A., Deng,S., Johnsen,H., Pesich,R. *et al*. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.
54. Nathanson,K.L., Shugart,Y.Y., Omaruddin,R., Szabo,C., Goldgar,D., Rebbeck,T.R. and Weber,B.L. (2002) CGH-targeted linkage analysis reveals a possible BRCA1 modifier locus on chromosome 5q. *Hum. Mol. Genet*., **11**, 1327–1332.
55. Pollack,J.R., Sorlie,T., Perou,C.M., Rees,C.A., Jeffrey,S.S., Lonning,P.E., Tibshirani,R., Botstein,D., Borresen-Dale,A.-L. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *PNAS*, **99**, 12963–12968.
56. Zender,L., Spector,M.S., Xue,W., Flemming,P., Cordon-Cardo,C., Silke,J., Fan,S.T., Luk,J.M., Wigler,M. *et al*. (2006) Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. *Cell*, **125**, 1253–1267.
57. Peeper,D. and Berns,A. (2006) Cross-species oncogenomics in cancer gene identification. *Cell*, **125**, 1230–1233.
58. Santarosa,M. and Ashworth,A. (2004) Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way. *Biochim. Biophys. Acta*, **1654**, 105–122.