Relationship of Gene Expression and Chromosomal Abnormalities in Colorectal Cancer

Dafna Tsafrir,[1] Manny Bacolod,[2] Zachariah Selvanayagam,[5] Ilan Tsafrir,[1] Jinru Shia,[3] Zhaoshi Zeng,[4] Hao Liu,[5] Curtis Krier,[5] Robert F. Stengel,[6] Francis Barany,[2] William L. Gerald,[3] Philip B. Paty,[4] Eytan Domany,[1] and Daniel A. Notterman[5]

[1]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel; [2]Department of Microbiology, Weill Medical College of Cornell University; Departments of [3]Pathology and [4]Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York; [5]Departments of Pediatrics and Molecular Genetics, University of Medicine and Dentistry of New Jersey/Robert Wood Johnson Medical School, Piscataway, New Jersey; and [6]School of Engineering and Applied Science, Princeton University, Princeton, New Jersey

Abstract

Several studies have verified the existence of multiple chromosomal abnormalities in colon cancer. However, the relationships between DNA copy number and gene expression have not been adequately explored nor globally monitored during the progression of the disease. In this work, three types of array-generated data (expression, single nucleotide polymorphism, and comparative genomic hybridization) were collected from a large set of colon cancer patients at various stages of the disease. Probes were annotated to specific chromosomal locations and coordinated alterations in DNA copy number and transcription levels were revealed at specific positions. We show that across many large regions of the genome, changes in expression level are correlated with alterations in DNA content. Often, large chromosomal segments, containing multiple genes, are transcriptionally affected in a coordinated way, and we show that the underlying mechanism is a corresponding change in DNA content. This implies that whereas specific chromosomal abnormalities may arise stochastically, the associated changes in expression of some or all of the affected genes are responsible for selecting cells bearing these abnormalities for clonal expansion. Indeed, particular chromosomal regions are frequently gained and overexpressed (e.g., 7p, 8q, 13q, and 20q) or lost and underexpressed (e.g., 1p, 4, 5q, 8p, 14q, 15q, and 18) in primary colon tumors, making it likely that these changes favor tumorigenicity. Furthermore, we show that these aberrations are absent in normal colon mucosa, appear in benign adenomas (albeit only in a small fraction of the samples), become more frequent as disease advances, and are found in the majority of metastatic samples. (Cancer Res 2006; 66(4): 2129-37)

Introduction

The initiation and progression of human solid tumors is associated with accumulation of alterations in the function of key regulatory genes. Many different factors, including changes in genome copy number and structure, can disrupt proper gene functioning. There is wide agreement that particular recurrent genomic aberrations may encompass genes that are important for tumor development (1, 2). This is particularly clear in cases involving gene dosage changes; tumor suppressor genes may be inactivated by a physical deletion and oncogenes may be enhanced by amplification (e.g., the oncogene *MYC*; ref. 3). However, the functional consequence of recurrent abnormalities is not always apparent, because a change in DNA copy number does not necessarily induce actual alterations in expression (4). The issue is further complicated by the observation that many aberrations span large chromosomal regions that contain multiple genes (1), including many that are not directly related to cancer.

The origins of chromosomal abnormalities and aneuploidy in cancer are the subject of debate (5). Opinions range from viewing aneuploidy as a central cause of tumor initiation (6–8) to regarding it as just a consequence of the derangements in the cell division cycle (9, 10). Large-scale technologies, such as comparative genomic hybridization (CGH), have been used to observe the role of genomic imbalances in solid tumors. In colorectal cancer, genomic aberrations are already present in high-grade dysplasias and adenomas but are significantly more abundant in carcinomas (8). In one study (4), in which CGH was used to determine frequent amplifications in metastatic colorectal cancer, the effects on expression levels were monitored by DNA microarrays. A fold change analysis of the expression data seemed to indicate that only a small minority of amplified genes were also overexpressed, suggesting that increased expression within amplicons in colorectal cancer is rare.

Similar work done on other solid tumors reveals a different picture. A recent study of the relationships between loss and gain of chromosomal material and global expression in head and neck squamous cell carcinoma (11) concluded that large chromosomal regions are transcriptionally affected, although many genes seemed to be unrelated to malignant progression. In work done on 14 breast cancer cell lines, >40% of highly amplified genes were also overexpressed (12), whereas in a similar study of primary breast tumors, 62% of highly amplified genes show moderately or highly elevated expression (13). Such studies suggest that alterations in DNA content can directly influence global expression patterns, and that some genomic aberrations may be selected because they alter the expression of multiple genes that coordinately promote tumor progression (1).

Because apparently in colorectal cancer the relationship between changes in DNA content and gene expression is not yet clear our study was designed to address three goals: *first*, to define how chromosomal abnormalities (changes in DNA content) are reflected in changes of expression of the genes in the affected region. A closely related issue concerns the expression levels of

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Eytan Domany, Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, 76100 Israel. Phone: 972-8-934-3964; Fax: 972-8-934-4109; E-mail: eytan.domany@weizmann.ac.il.

doi:10.1158/0008-5472.CAN-05-2569

single genes: how frequently are the expression and copy number of a gene discordant? Our *second* goal: to compile a comprehensive list of chromosomal regions that exhibit abnormalities in adenomatous polyps, carcinoma, and metastases.

This leads to our *third* goal: to describe whether chromosomal abnormalities and associated changes in expression form a coherent pattern that evolves with the clinical stage of the patient.

Therefore, we have systematically explored both gene expression and chromosomal content data related to samples collected from colorectal cancer patients at various stages. We probed the epithelial component of normal colon mucosa, adenomatous polyps, colon adenocarcinoma from patients with various clinical stages of disease, and metastases, using three approaches: the Affymetrix GeneChip Human Genome U133A Array (expression), the Affymetrix GeneChip Mapping Array (single nucleotide polymorphism, SNP), and a 3000 BACs spotted array (CGH). We thus produced stage-stratified chromosomal maps that present changes at the DNA level and at the transcription level as a function of chromosomal location. When viewed at the level of the chromosome arm, we find strong and significant correlations between perturbations in gene expression and DNA content. However, this relationship is complex; particular genes that lie in amplified regions may nevertheless exhibit reduced expression. We introduce *mutual correlation*, a novel measure of cooperativity, and use it together with the traditionally used fold change analysis to identify coherently amplified chromosomal regions and to study the manner in which chromosomal abnormalities change with the pathologic progression and clinical stage of disease.

Materials and Methods

Expression data analysis. Expression profiles were determined for five types of samples dissected under a microscope, using the Affymetrix U133A GeneChip (14): 24 normal colon epithelium, 30 adenomas, 114 carcinomas, 10 liver metastases, and 9 lung metastases. The location of each probe set was retrieved from Affymetrix (exact locations were available for 18,067 of 22,283 probes on the U133A chip). The data were preprocessed using MAS 5.0, and standard thresholding and filtering operations were used (15); the data were log-transformed, and each gene's expression levels were centered and normalized to yield a matrix whose elements e_{gi} represent the expression level of gene g in sample i.

The metastasis samples were subjected to *electronic microdissection*, a preanalysis step described in detail in ref. 16, using a newly developed algorithm, sorting point into neighborhoods (SPIN), to identify the purest metastasis samples (10 in liver and 9 in lung), least contaminated by surrounding normal tissue. In ref. 16 we also established that our polyps and low-grade adenoma samples are not contaminated by surrounding tissue, whereas the adenocarcinoma samples contain higher and varying levels of contamination.

The fold change ratio f_c was calculated for each gene by dividing its median expression in the relevant group of samples (polyps, primary tumors, liver metastasis, or lung metastasis) by its median expression over the normal colon samples.

To evaluate the cooperativity in expression of physically adjacent genes, we introduced *mutual correlation*, calculated for a group G of genes $g = 1, 2, \ldots, n$ over a set of samples $i = 1, 2, \ldots, N$ as follows: denote the median expression of the genes of G for each sample is $\tilde{e}_i = \text{median}(e_{gi})$. The mutual correlation $C_g(G)$ of gene g with the group G by the Pearson correlation of the two sets of expression levels, $(e_{g1}, e_{g2}, \ldots, e_{gN})$ and $(\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_N)$. Each of the genes annotated to a particular chromosomal location was indexed accordingly; chromosomal interval $[i,j]$ includes the genes of indices i, $i + 1$, $i + 2, \ldots, j$. For each gene in a particular interval $[i,j]$, we calculated $C_g([i,j])$. The score for the interval is defined as the median of the mutual correlations of all the genes in the interval, $C[i,j] = \text{median}\{C_k\}_{k=i}^{j}$.

DNA fold change analysis. A preliminary analysis of the expression data that was collected for our samples allowed us to select the best candidates for CGH or SNP arrays. We focused on a particular chromosomal region (specifically 20q) known to be frequently gained in colorectal cancer. We sampled both tumors that exhibited a clear coordinated increased expression, as well as tumors whose expression in the region was comparable with that of normal samples. Another requirement was a high percentage of tumor cells in the sample, as judged by the pathologic annotation provided with the sample.

Array CGH data was acquired from two different sources:

(*a*) We did array CGH on 12 of our samples (10 primary tumors and 2 liver metastasis samples); 3,100 probes (PCR products) derived from BAC clones obtained from the Welcome Sanger Trust Institute were spotted in duplicates in 48 subgrids. This array provides ~1 Mb resolution (17, 18). The control DNA used was human placenta. Probes were labeled by Cy3 and Cy5, so that data values are log 2 (Cy3/Cy5).

(*b*) Array CGH data of 1Mb resolution for 37 primary colon tumors was taken from Douglas et al. (17).

The Affymetrix GeneChip Human Mapping 50K array Xba 240 array ("SNP array") was used on seven of our primary colon tumors (five of these samples were also measured by CGH, as described). The protocol is detailed on "GeneChip Mapping 100K Assay Manual".[7] Following digestion of 0.25 μg of genomic DNA with *Xba*I, ligation of adapter DNA to the fragments, and PCR amplification, such that the products are in the range of 250 to 2,000 bp, the purified products are fragmented, labeled, and hybridized to the array, which is next scanned to generate the image (DAT) and cell intensity (CEL) files. The CEL file is imported to GeneChip DNA Analysis Software 3.0 (GDAS 3.0, Affymetrix) to generate the SNP calls. The Chromosomal Copy Number Analysis Tool version 2.0 (Affymetrix), then uses the probe intensity data, as well as the SNP calls to generate genomic-smoothed copy number estimate (using the default 0.5 Mb smoothing), log *P*s for the copy number estimate, and loss of heterozygosity calls for each SNP (19). Fold change values per SNP position were calculated by dividing the gene copy values by the median copy number of the entire chip (the median was very close to 2 for all of our samples).

Correlating DNA fold change with expression. To correlate fold change values from expression with either CGH or SNP:

(*a*) Each of the three data sets was ordered according to the chromosomal locations of the measured probes, permitting display of a linear alignment plot of sets of data.

(*b*) Each alignment plot was smoothed using a moving average that highlights regions with a unidirectional change in either expression or DNA dosage of many adjacent probes (20). This straightforward approach reduces variance and extracts trends and patterns from ordered data series. Averaging was done in windows of the following size: 150 probe sets for the expression data, 100 for the SNP data, and 40 for CGH, while taking into account the actual physical distances between adjacent probes (see Supplementary Material for details, including the interpolation procedure used).

Results

Significant correlation is observed between "large-scale" expression and DNA copy number. The following analysis allows identification of chromosomal regions where expression and (or) DNA dosage were altered in a consistent fashion. For each probe location, two types of fold change values were calculated per tumor sample (one for DNA content and the second for RNA expression intensity) relative to normal tissue. A probe with fold change $f_c > 1$ has a higher ($f_c < 1$ indicates lower) measurement in the malignant sample relative to normal tissue. The $\log_2$ fold change is displayed as a function of chromosomal location in Fig. 1. The three techniques (expression, CGH, and SNP) produce three data series, which are presented after smoothing (to dampen noise and

[7] http://www.affymetrix.com.

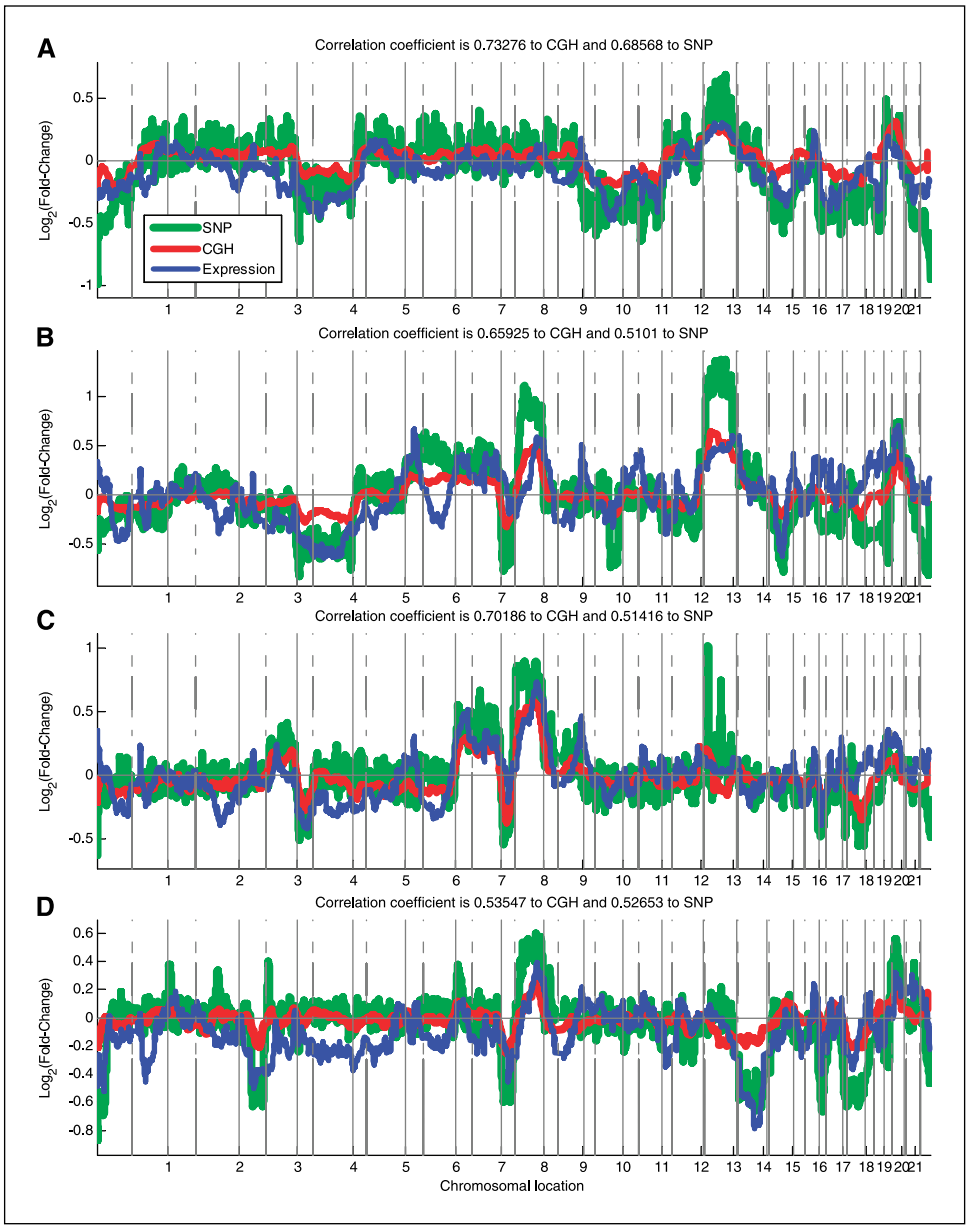

Figure 1. Dosage changes in genomic material and expression levels. A fold-change map is presented per tumor for four different primary tumors. Expression fold-change (*blue*) was calculated per tumor sample by comparing with a normal control (the median of 24 normal colon mucosa samples). DNA fold-change is measured on two different platforms: (*A*) SNP based fold-change was calculated by dividing the copy number per probe by the median copy number for each sample (*green*). (*B*) Values for CGH array are $\log_2$ (Cy3/Cy5) (*red*). To reduce noise and improve visibility, the plots were smoothed in a running average of size 150 for the expression data, 100 for the SNP data, and 40 for CGH, while taking into account the actual physical distances between adjacent genes (the same smoothing technique is also used in Figs. 2–4). *Vertical lines,* transition between chromosomes; *dotted vertical lines,* transition from arm p to q.

enhance trends and patterns) and interpolation (to allow calculation of a correlation coefficient).

There was a strong genome-wide correlation between gene expression and DNA content, for both CGH-derived and SNP-derived data. For the nine primary tumor samples and two liver metastases studied with CGH, the mean correlation was 0.67 $\pm$ 0.1 (SD). For the six primary tumor samples characterized by SNP, the mean correlation between DNA content and expression was 0.58 $\pm$ 0.09. Five of the primary tumor samples were measured on all three platforms. Results from four of these samples are presented in Fig. 1: these are microsatellite-stable (MSS) carcinomas with (*A*) 80%, (*B*) 90%, (*C*) 80%, and (*D*) 85% tumor cells. The fifth sample did not show significant correlation between expression and copy number. Interestingly, that particular sample is the only microsatellite instable (MSI^+) tumor for which we recorded DNA copy number.

The strong correlation between expression and copy number means that when viewed on a chromosomal scale, gain or loss of chromosomal regions is usually accompanied by a corresponding change in transcription of genes. Overall, 63% of the significantly overexpressed genes also display DNA content gains, and 62% of the down-regulated genes show a noticeable loss of DNA content (see Supplementary Material for definitions). In particular, several large chromosomal regions, such as those located in 8q and 20q, repeatedly display increased signal at both the DNA and expression levels, whereas other regions, such as chromosome 4, 8p, and 18, exhibit a coordinated reduction in signal. Several array CGH studies have verified gain and loss of these regions (4, 21–23), supporting the hypothesis that the observed chromosomal-scale changes in expression are linked with underlying alterations in copy number.

The correlation between expression and DNA copy number data is not perfect; in particular, as seen in Fig. 1, chromosomal arms 1q, 2q, 6q, 9q, and 19p show a measure of discrepancy between expression and CGH/SNP in several of the samples shown.

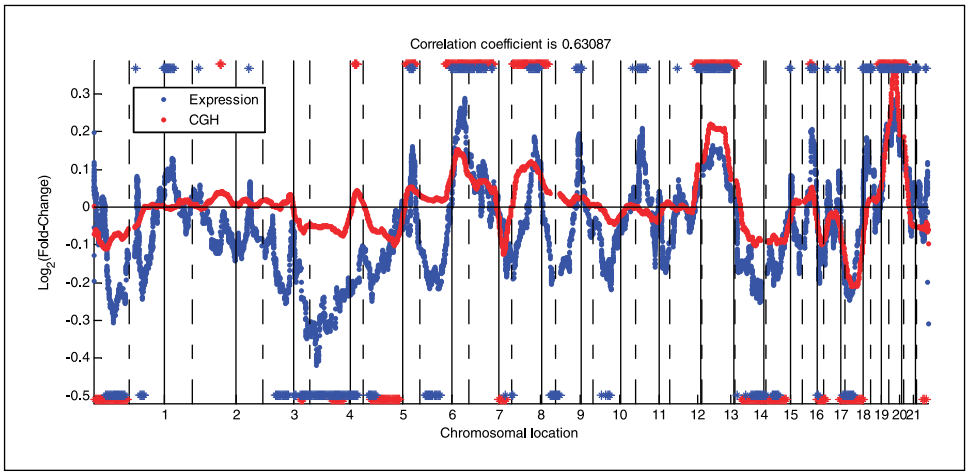

Figure 2. Fold change in CGH and expression averaged over two populations of primary tumors. Expression fold change was calculated per probe by taking the median over 114 primary tumor samples and dividing it by the median of 24 normal colon mucosa samples (*blue*). CGH fold change data for 37 primary tumors (*red*) was taken from Douglas et al. (17). A blue star at the *top* (*bottom*) of the axes marks a location with significant overexpression (underexpression). Correspondingly, a red star has a similar meaning in terms of DNA copy number.

When focusing on individual genes, the situation is even more complex. As can be seen in Supplementary Fig. S1, even within regions with large gains of DNA content, expression of some genes is down-regulated in tumor. Furthermore, as will be discussed below, even within a chromosomal arm that is amplified in its entirety, one may find contiguous regions whose genes are expressed at levels similar to normal tissue.

Characterization of a population of tumors: specific genomic regions are implicated. Next, we proceeded to identify the most prevalent sites of alterations in both genome and transcriptome in the context of colorectal cancer progression. We used expression from our large set of 114 primary colon carcinomas and 24 normal colon mucosa samples, and CGH data from 37 primary tumors, collected by Douglas et al. (17). As in the previous section, alterations in both DNA and transcription are presented together in the same graph (see Fig. 2). Here, the fold change values are produced by averaging over large groups of tumors (rather than viewing one patient at a time, as in Fig. 1).

The average expression fold change displayed in Fig. 2 is calculated by dividing the median expression per probe obtained from 114 carcinomas with that probe's median expression in the 24 normal colon mucosa samples. Hence, only genes that are overexpressed in the majority of the tumor population will receive a significantly high fold change (and likewise for underexpressed genes). A similar argument is true for the average DNA fold change values, so that Fig. 2 indicates that large regions in chromosomes 7, 8q, 13, and 20 are gained and overexpressed in a majority of colon tumors, whereas parts of 1p, 4, 5q, 8p, 14q, 15q, 17p, and 18 are lost and hence become underexpressed. Note the good qualitative agreement between expression and CGH data and the high correlation of 0.63 obtained in spite of the fact that expression and CGH were measured over different sets of samples.

Abnormal expression patterns become more pronounced with more advanced stage. To learn whether there is a relationship between the clinical or pathologic stages of colorectal cancer and gene expression, we present in Fig. 3 a fold change chromosomal map for each of the disease stages. The combined image allows clear identification of chromosomal regions where many genes are progressively perturbed together. The expression patterns of tumors at stages II, III, and IV are very highly correlated with one another (0.95-0.96). Referring to this group as "advanced stage," we observe in Fig. 3 that the expression patterns evolve and

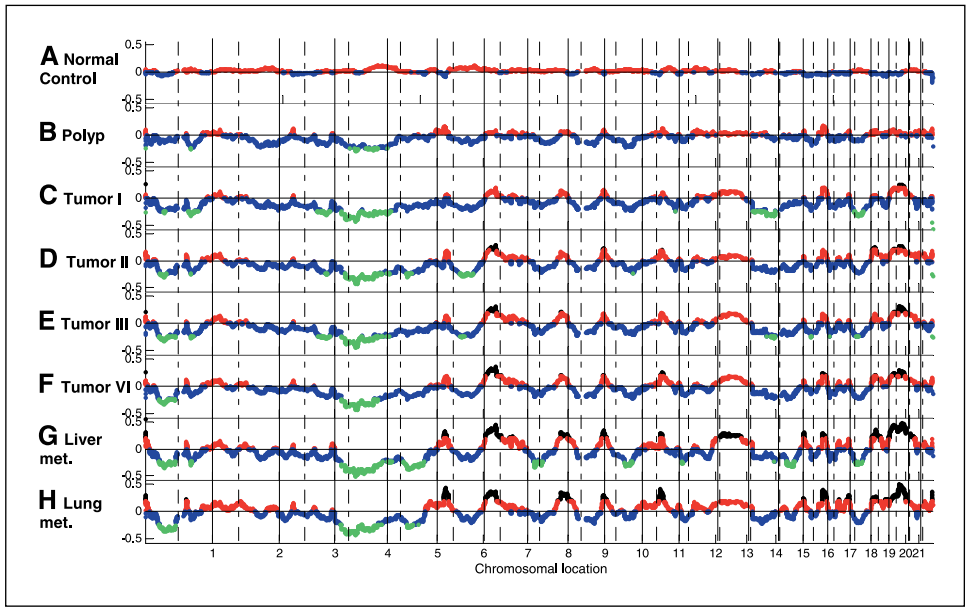

Figure 3. Large-scale expression biases during disease progression. For each annotated gene, the $\log_2$ of the median fold change ratio, $\log_2(f_c)$, is presented in (*A*) normal colon tissue (this is a control value generated by randomly dividing the normal samples into two arbitrary groups), (*B*) adenoma, (*C-F*) primary tumors of stages I to IV, (*G*) liver metastasis, and (*H*) lung metastasis versus normal colon tissue. *Red* and *black points,* increased expression: *black,* top 5% fold-change values (5% of all fold-change values shown in *A-H*). *Blue* and *green points,* reduced expression; *green,* lowest 5% values (again of all values). Note that in the normal control, the values are very close to 0, while in the progressive stages of the disease there is a marked deviation from 0 in several specific regions.

become closer to the metastases with progression from adenoma to stage I tumors and to advanced tumors. Note that the correlation to liver metastases varies as follows: 0.32 (normal colon), 0.74 (adenoma), 0.82 (carcinoma stage I), and 0.9 (advanced stage).

Figure 3 shows that chromosome 4 displays broad underexpression even in adenomas, more so in carcinoma and most profoundly in metastases. A similar link with disease stage is seen for overexpressed regions, such as chromosomes 7p, 13q, and 20, where the increase in transcript level is already noticeable in adenoma but becomes more pronounced at later stages of the disease. Furthermore, it was established (see Fig. 2) that altered expression in those regions is associated with gain in genomic material.

A different and complementary approach for evaluating the relationships between genes that are located within a given chromosomal region is by calculating their mutual correlation. Mutual correlation was defined in Materials and Methods: here, for each gene g, we calculate the $C_g[i,j]$ for a *window* $[g - 10, g + 10]$ of 21 genes centered on g (we have tested a range of window sizes and determined that the results are not dependent on the particular choice of window size). $C_g[g - 10, g + 10]$ is measured separately for each clinical group (e.g., liver metastasis). A high C_g indicates that the fold change profile of gene g exhibits high similarity to that of its 20 closest neighbors on the chromosome. A region in which most genes display relatively high C_g is inclined to be overexpressed (or underexpressed) collectively in the same samples (see Fig. 4). This implies deletion or gain of a contiguous region of genes.

Detailed exploration of chromosome 20. A large chromosomal region that undergoes a change in DNA copy number may contain many different genes. An important question in this context is whether the transcriptional effect on all genes annotated to that region is the same. Here, we concentrate on chromosome 20, which has been repeatedly shown to be subject to amplification in colorectal cancer (ref. 17; in Supplementary Figs. S3 and S4, we present a similar analysis of chromosomes 7 and 8).

Both CGH and SNP arrays documented the expected large-scale gain of chromosome 20 (see Figs. 1 and 2). Furthermore, we have established a gradual increase of transcriptional changes in both fold change (Fig. 3) and mutual correlation (Fig. 4) for genes annotated to that chromosome. Figure 5 provides informative visualization of copy number changes on chromosome 20. Figure 5*A* displays the correlation matrix for the 75 probes annotated to chromosome 20 on the CGH array (correlations were measured for 37 primary carcinomas; ref. 17). The probes are ordered according to their relative positions on the chromosome, so that the first 25 are located on 20p and the following 50 on 20q. Thus, it is possible to recognize groups of physically adjacent probes that share similar gain or loss profiles; they are visually manifested as red (standing for high similarity levels) squares around the matrix's main diagonal. Two major clusters are clearly apparent in the CGH-based correlation matrix of chromosome 20 (Fig. 5*A*); one includes all the probes annotated to 20p, and the second is composed of the 20q probes. The corresponding fold change matrix in Fig. 5*B* helps with the biological interpretation; here, the probes are ordered according to chromosomal location, as in Fig. 5*A*, whereas the samples are ordered by an unsupervised sorting algorithm (i.e., SPIN, which is described in detail in ref. 16). The ordering generated by SPIN achieves a clear visualization of the different chromosomal instability profiles that appear in the tested population. Over all, the 20p arm is positively gained in a small subset of the primary tumors, whereas the entire 20q arm is gained in a larger fraction of tumors. However, the two arms are not necessarily gained together in a given sample. In fact, some of the samples that show the strongest increase for 20q display a moderate loss of 20p. Interestingly, in most samples, each of the arms of chromosome 20 tends to behave as a single unit in the context of copy number changes.

A more complex picture is revealed when the same type of analysis is done on our expression data. The expression-based correlation matrix (Fig. 5*C*) also has a clear distinction between 20p and 20q probes. However, whereas the CGH analysis showed a clear tendency of the entire 20q arm to be gained together as a single unit, only some of the 201 genes that are annotated to that region are overexpressed simultaneously. In particular, there are two main subregions (marked 1 and 2 in Fig. 5*C*), where adjacent genes share highly similar expression profiles. The expression matrix (Fig. 5*D*) reveals that in those two subregions, which contain tens of transcripts, gain of contiguous chromosomal material in the transformed tissue is accompanied by overexpression of most genes contained within those segments. However, even within such contiguous overexpressed regions, there are individual genes that are not overexpressed. This could be attributed to alternate mechanisms of transcriptional control or to differences in gene copy number that are below the limit of resolution of the CGH approach.

The first subregion spans 20q11.2-12, an area that has been previously implicated with DNA amplification in germ cell tumors

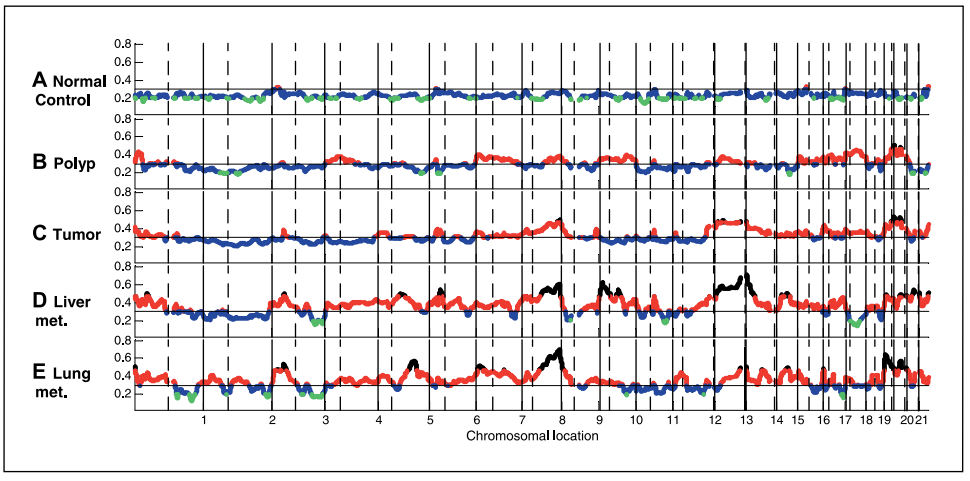

Figure 4. Mutual correlation chromosome scan. Mutual correlation per gene is calculated in the context of the chromosomal region spanning from 10 genes upstream to 10 genes downstream (for a total of 21 genes). The presented plots are for (*A*) normal samples, (*B*) adenoma, (*C*) primary tumors, (*D*) liver metastasis, and (*E*) lung metastasis samples. Coloring of points is dictated by the correlation's strength: *red* and *black points,* values above the median; *black,* top 5% (5% of all correlation values from *A-E*); *blue* and *green points,* below-median correlations; *green,* lowest 5% values (again of all values).

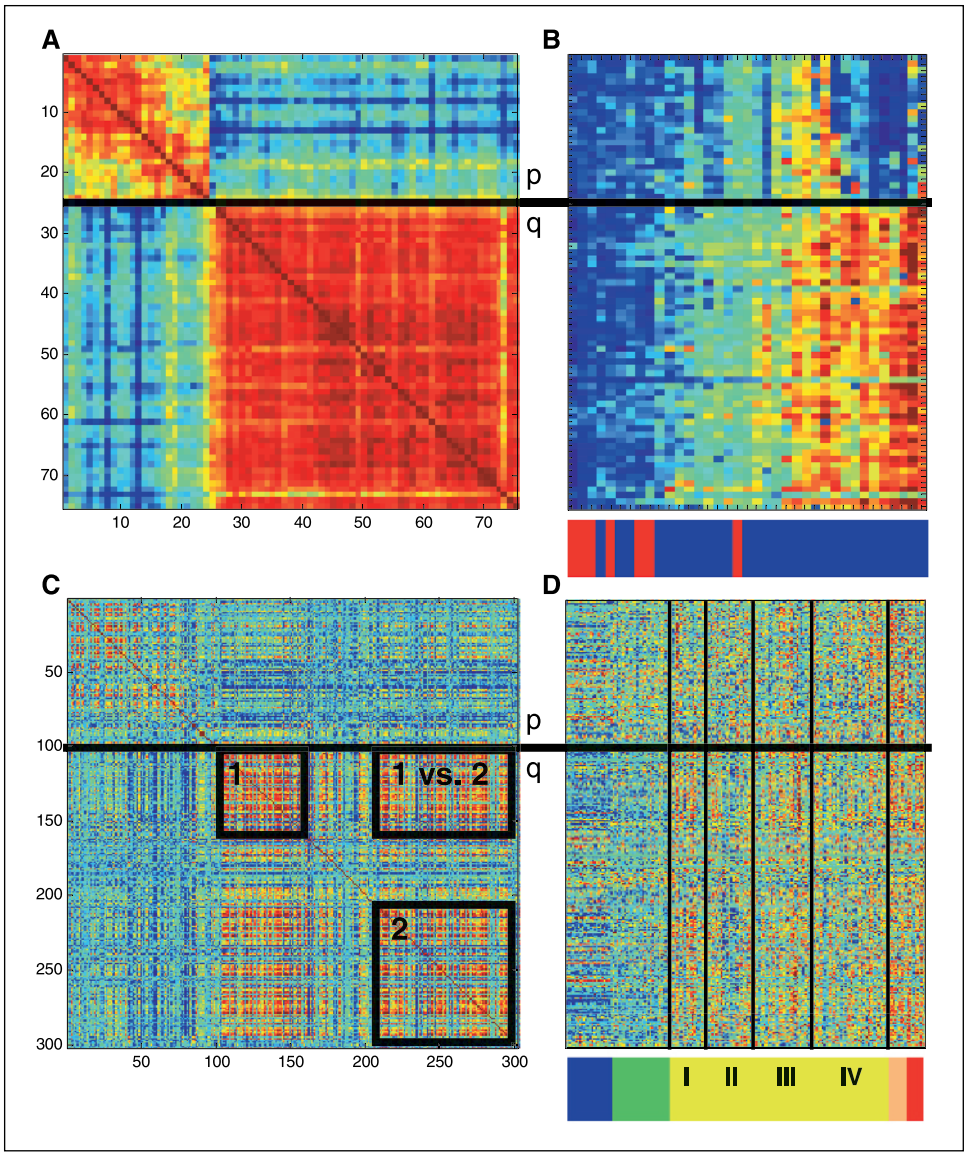

Figure 5. Instability patterns in chromosome 20. *A,* correlation matrix calculated from CGH data for the 75 probes annotated to chromosome 20. The probes are ordered according to their relative chromosomal locations. Each element in the matrix represents the Pearson correlation between the DNA fold change profiles of two probes, as measured across 37 primary colon tumors (17). *Blue,* small correlations; *red,* high correlations. A group of adjacent genes that shares a highly similar copy number profile is visually manifested here as a red square on the main diagonal of the correlation matrix. *B,* the correspondingly ordered DNA fold change matrix. Each row is a probe (in the same order as in *A*), and each column is a sample. Colors stand for relative fold-change. The samples were ordered using SPIN, an unsupervised sorting algorithm (16). The MSI status is given in the colored bar below the matrix: *red,* MSI^+; *blue,* chromosomal instability. *C-D,* corresponding figures for expression data. *C,* correlation matrix for the 302 genes annotated to chromosome 20 on the U133 Affychip, calculated in the sample space spanned by 187 samples (24 normals, 30 adenomas, 114 adenocarcinomas, 10 liver metastases, and 9 lung metastases). Note that there are two main subregions that score highly for similarity and are highlighted by black boxes numbered *1* and *2*. Furthermore, one can also observe the inter-regional similarity in the off-diagonal region of the correlation matrix (black box marked *1* versus *2*). *D,* the corresponding expression matrix. Colors stand for centered and normalized expression values: *blue* (*red*), relatively low (high) expression. Each row is the expression profile of a single gene. The ordering of genes is identical to the one in (*C*). Each column depicts a single sample. The samples are ordered according to tissue identity, indicated for each sample by the colored bar below the matrix: *blue,* normal; *green,* adenomas; *yellow,* carcinomas; *orange,* liver metastases; *red,* lung metastases. The tumor samples were divided into four clinical stages and ordered accordingly. Some samples exhibit high expression across most of the chromosomal arm, visually manifested as reddish vertical stripes and interpreted as indicative of chromosomal amplification. Note that no such normal samples are observed, very few among the adenomas, much more among the stage II to IV carcinoma, and metastasis samples are predominantly amplified.

(24) and breast cancer (25). Candidate genes annotated to this region include *BCL2L1*, a known regulator of programmed cell death, whose overexpression has been reported in >60% of human colorectal cancer (26). *S*-adenosylhomocysteine hydrolase (*AHCY*) has been previously observed to be up-regulated in colorectal cancer (27). *TGIF2* is known to be amplified and overexpressed in ovarian cancer (28). The second subregion covers 20q13, with a higher-resolution scan implicating 20q13.33 as the best-scoring area on chromosome 20 in terms of mutual correlations (see Materials and Methods). This region was observed by prior CGH experiments to be most frequently gained in colon tumors (17). In that region, *LIVIN* (*BIRC7*), an inhibitor of apoptosis, has been associated with the progression of bladder cancer and detected at high levels in a colorectal cancer cell line (17). An independent

validation of our results is shown in Supplementary Fig. S2, where results from fluorescence *in situ* hybridization (FISH) are presented for the 20q13.2 region for six adenocarcinoma, showing amplification of up to 10 to 20 copies in some samples.

An interesting question we address briefly concerns the onset of the abnormalities in different chromosomes: do amplifications of different chromosomal regions occur independently of each other, or are they synchronized? Our data suggest that once cells become susceptible to chromosomal instability, they experience several instances of amplification or loss. Because 20q is so clearly and frequently amplified, it is reasonable to examine whether other specific chromosomal regions are amplified or lost concurrently with 20q. Figure 6 presents the median correlation of (the genes on) each chromosomal arm with the average expression profile of 20q. There is a clear and gradual increase in correlation with disease development, evident from the fact that the median correlation values increase as we move to polyps, to tumors, and to metastases. This observation suggests that in tumors that contain a positive change in copy number for 20q, there is also a high likelihood for a positive change in the 8q arm. On the other hand, the negative correlation with 18p, for example, suggests that a gain of 20q tends to be coupled to loss of 18p. Over all, there is a positive correlation of 20q overexpression with potential amplifications in 7p, 8q, 13q, 16p, 19p, and 20p and negative correlation with losses in 18, 4, and 15q.

In Fig. 5*D*, we present the expression matrix of the 302 genes, ordered according to location on chromosome 20 (101 on 20p and 201 on 20q), measured in our samples, ordered according to their clinical labels, including a breakdown of the 114 carcinoma according to clinical stage. This provides a striking manifestation of the increased chromosomal abnormality with disease progression; samples that exhibit overall high expression levels appear as vertical yellow-red stripes. None of the normal colon samples exhibit such coordinated overexpression of the 20q genes; a few of the polyps do, and the percentage of these amplified samples increases with progression; consistently amplified samples constitute a small minority of the polyps but are the majority of the stage II to IV carcinoma and even more so in the metastasis samples. Similar variation in chromosomal abnormalities with disease stage was observed in other chromosomes as well.

The full richness, heterogeneity, and complexity of transformed malignant tissues versus their normal counterparts is seen very clearly in Supplementary Fig. S5, which presents the expression matrix of the 201 genes of 20q over the samples, with normal liver and normal lung also added. One striking observation is the homogeneity of normal tissue, which is visually manifested in horizontal stripes across each of the three types of normal tissue (also seen in normal colon on Fig. 5*D*). The meaning of these horizontal streaks is that genes that are high (low) in one normal sample tend to be high (low) in all normal samples from the same tissue. On the other hand, the tumors and metastasis samples are very heterogeneous, as exhibited by the vertical, rather than horizontal, stripes discussed above.

Discussion

Our findings suggest that disease progression is associated with coordinated changes in the expression of substantial groups of contiguous genes. These changes in gene expression are associated with and presumably caused by changes in the copy number of contiguous genes along large segments of the chromosome. This explanation is directly supported by performing expression, CGH, and SNP arrays, as well as FISH measurements on the same tumor samples and finding a high degree of correlation between chromosome-scale expression and gene copy number profiles. Thus, >60% of overexpressed (or underexpressed) sites are shown to be associated with gains or loss in the genetic material. Consistent with our results, several studies in different types of solid tumors, breast (12, 13), prostate (29), and head and neck squamous cell carcinoma (17), concluded that alterations in DNA copy number directly influence the expression levels of multiple genes. In a budding yeast study (30), aneuploidy has been linked with chromosome-wide gene expression biases.

The discordance with the results reported by Platzer et al. (4) most probably stems from a difference in the analysis methodology. Platzer et al. concentrated on individual genes and found

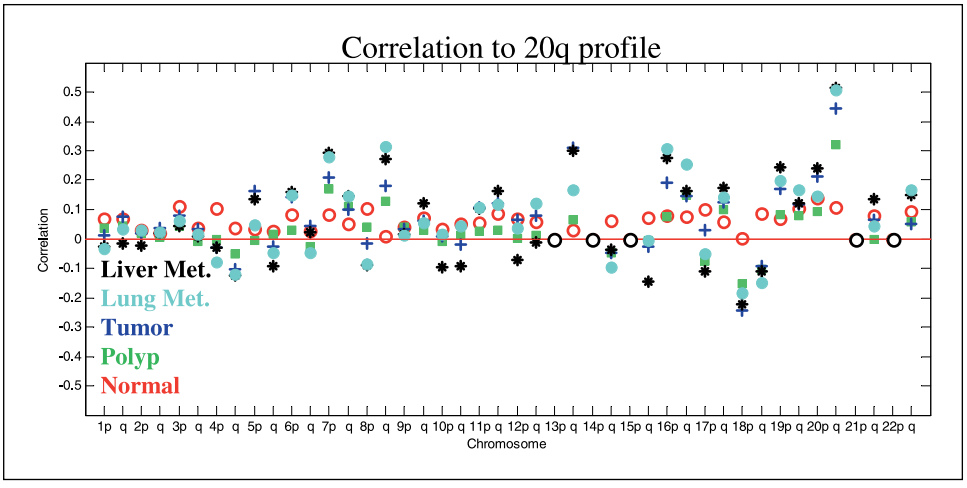

Figure 6. Chromosome scan for correlation to the 20q profile. We created the following five sample groups: (*a*) normals, (*b*) normals and polyps, (*c*) normals and primary adenocarcinomas, (*d*) normals and lung metastasis, (*e*) normal and liver metastasis. Next, we calculated for each group $i = 1, 2, \ldots, 5$, the average expression profile $<E_i(20q)>$ of the genes on 20q. For every chromosomal arm, we now calculate, for every gene on the arm, the correlation of its expression (over each of the five sample groups) with $<E_i(20q)>$. The median of these correlations is denoted c_i^{med}(arm); these are presented in the figure for each chromosomal arm. A high value of c_i^{med}(arm) indicates that most of the genes on that arm are highly correlated with the 20q profile. Note that for 20q, we do *not* get $c_i^{med}(20q) = 1$ because the correlation of each gene with $<E_i(20q)>$ is <1; hence, the median of these correlations is also <1.

that only 3.8% of the genes in the most frequently gained regions are overexpressed at >2-fold over normal colon. Using the same procedure on our samples resulted in similar percentages (only 6% of genes on the amplified chromosome 20 pass the 2-fold threshold). Taking a more global approach, we see that the vast majority (>80%) of genes on chromosome 20 were elevated to some degree (albeit <2-fold) in tumors that gained DNA copy number. Again, Platzer et al. report a similar percentage (almost 90%) of genes that are elevated to some degree.

On the scale of individual genes, the picture becomes more complex; we do find genes, the expression of which is not correlated (may even be negatively correlated) with copy number changes in the region where the gene resides. This is attributed to there being a multiplicity of mechanisms responsible for normal and abnormal control of gene expression, including those related to mutation, promoter methylation, and micro-RNA expression.

We show that it is possible to use gene expression microarray data to identify cytogenetically abnormal regions that are associated with malignant transformation (20, 31). A practical implication is that expression array data can be used to supplement CGH or SNP profiling. Furthermore, combining these data sources allows for a more biologically relevant interpretation of the expression data by highlighting the dependence of gene expression on gene dosage. A more fundamental inference is that recurrent genomic aberrations in colorectal cancer have a significant influence on gene expression and hence may play an important role in the development of colorectal cancer. At the moment, the part that genomic instability plays in tumorigenesis remains unclear (5), with opinions ranging from genomic instability being a cause to being an effect of malignant transformation (6, 7, 9, 10). Two observations suggest that some chromosomal changes have an important pathogenetic role. The first is that specific chromosomal regions are repeatedly observed to undergo gain or loss associated with a regional bias in transcription (in the current work, −1p, −4, −5q, +7, −8p, +8q, +13q, −14q, −15q, −18, and +20q). That these changes are recurrent and not stochastic in different tumors suggests that *as a whole*, they carry a selective advantage for the transformed cell even if not every embedded gene is relevant to transformation. A second observation is that these changes are progressive: we have shown that the expression of continuous segments of genes becomes progressively deranged with disease progression. Hence, we showed that in the 20q arm, for example, many contiguous genes are overexpressed, and that this regionally biased transcription does not occur in normal tissue, is observed in a small fraction of adenoma samples, and is much more prevalent in invasive carcinoma and metastases.

Changes at the chromosomal scale will bias expression over large regions and is likely to affect genes that are unrelated to malignant progression (11). It is likely that bystander genes, located in close physical proximity to cancer pathway genes, are included within these segments but do not confer a selective advantage. This would be analogous to genetic polymorphisms, which once fixed in a population, persist in equilibrium, although they no longer confer a selective advantage. The number and type of samples analyzed by CGH or SNP in the present study is neither sufficiently large nor sufficiently heterogeneous with respect either type (MSI^+ versus MSI^-) or discordant gene expression and copy number changes to permit one to distinguish expression changes that are essential to the neoplastic process from those that are merely bystanders. However, by analyzing and comparing many more samples by both expression and copy number, it may become possible to focus on genes or chromosomal segments that are repeatedly subject to correlated changes in copy number and expression. Extending this analysis into a larger number of MSI^+ samples may provide more information regarding expression changes in the absence of high-level chromosomal instability and may point to genes for which, for example, overexpression is not biologically relevant. In summary, we have shown that in colorectal cancer the mRNA expression of large groups of contiguous genes varies in a coordinated way. These expression changes reflect gain or loss of the associated genes. Furthermore, these changes are progressive and reflect the clinical and pathologic stage of disease. Perturbations in gene copy number are common in colorectal cancer, affect gene expression, and presumably the biology of the malignant cells that comprise the tumor.

Acknowledgments

Received 7/21/2005; revised 11/9/2005; accepted 12/2/2005.

Grant support: R.F. Stengel, F. Barany, W.L. Gerald, P.B. Paty, E. Domany, and D.A. Notterman are Principal Investigators of a National Cancer Institute Program Project Grant (P01-CA65930). Work in the Barany laboratory is sponsored in part by a sponsored research grant from Applied Biosystems, Inc., for which F. Barany also serves as a consultant. Work of the Domany group was partially supported by the Ridgefield Foundation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank the use of Gene Expression Core Facility of Cancer Institute of New Jersey.

References

1. Albertson DG, Collins C, McCormick F, Gray JW. Chromosome aberrations in solid tumors. Nat Genet 2003;34:369–76.
2. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–99.
3. Alitalo K, Schwab M, Lin CC, Varmus HE, Bishop JM. Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-*myc*) in malignant neuroendocrine cells from a human colon carcinoma. Proc Natl Acad Sci U S A 1983;80:1707–11.
4. Platzer P, Upender MB, Wilson K, et al. Silence of chromosomal amplifications in colon cancer. Cancer Res 2002;62:1134–8.
5. Marx J. Debate surges over the origins of genomic defects in cancer. Science 2002;297:544–6.
6. Li R, Sonik A SR, Rasnick D, Duesberg P. Aneuploidy vs. gene mutation hypothesis of cancer: recent study claims mutation but is found to support aneuploidy. Proc Natl Acad Sci U S A 2000;97:3236–41.
7. Duesberg PH. Are cancers dependent on oncogenes or on aneuploidy? Cancer Genet Cytogenet 2003;143:89–91.
8. Shih IM, Zhou W GS, Lengauer C, Kinzler KW, Vogelstein B. Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis. Cancer Res 2001;61:818–22.
9. Zimonjic D, Brooks MW, Popescu N, Weinberg RA, Hahn WC. Derivation of human tumor cells *in vitro* without widespread genomic instability. Cancer Res 2001;61:8838–44.
10. Lamlum H, Papadopoulou A, Ilyas M, et al. APC mutations are sufficient for the growth of early colorectal adenomas. Proc Natl Acad Sci U S A 2000;97:2225–8.
11. Masayesva BG, Ha P, Garrett-Mayer E, et al. Gene expression alterations over large chromosomal regions in cancers include multiple genes unrelated to malignant progression. Proc Natl Acad Sci U S A 2004;101:8715–20.
12. Hyman E, Kauraniemi P, Hautaniemi S, et al. Impact of DNA amplification on gene expression patterns in breast cancer. Cancer Res 2002;62:6240–5.
13. Pollack JR, Sorlie T, Perou CM, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A 2002;99:12963–8.
14. Tsafrir D, Liu W, Yamaguchi Y, et al. A novel mathematical approach to analyzing gene expression data: results from an international colon cancer consortium. Proc Am Assoc Cancer Res 2004;45:4799.
15. Rozovskaia T, Ravid-Amir O, Tillib S, et al. Expression profiles of acute lymphoblastic and myeloblastic

leukemias with ALL-1 rearrangements. Proc Natl Acad Sci U S A 2003;100:7853–8.
16. Tsafrir D, Tsafrir I, Ein-Dor L, et al. Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. Bioinformatics 2005;21:2301–8.
17. Douglas EJ, Fiegler H, Rowan A, et al. Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. Cancer Res 2004;64:4817–25.
18. Fiegler H, Carr P, Douglas EJ, et al. DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. Genes Chromosomes Cancer 2003;36:361–74.
19. Huang J, Wei W, Zhang J, et al. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics 2004;1:287–99.
20. Furge KA, Dykema KJ, Ho C, Chen X. Comparison of array-based comparative genomic hybridization with gene expression-based regional expression biases to identify genetic abnormalities in hepatocellular carcinoma. BMC Genomics 2005;6:67.
21. Ried T, Knutzen R, Steinbeck R, et al. Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. Genes Chromosomes Cancer 1996;15:234–45.
22. He QJ, Zeng WF, Sham JS, et al. Recurrent genetic alterations in 26 colorectal carcinomas and 21 adenomas from Chinese patients. Cancer Genet Cytogenet 2003;144:112–8.
23. Nakao K, Mehta KR, Fridlyand J, et al. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. Carcinogenesis 2004;25:1345–57.
24. Rao PH, Houldsworth J, Palanisamy N, et al. Chromosomal amplification is associated with cisplatin resistance of human male germ cell tumors. Cancer Res 1998;58:4260–3.
25. Hodgson JG, Chin K, Collins C, Gray JW. Genome amplification of chromosome 20 in breast cancer. Breast Cancer Res Treat 2003;78:337–45.
26. Krajewska M, Moss SF, Krajewski S, et al. Elevated expression of Bcl-X and reduced Bak in primary colorectal adenocarcinomas. Cancer Res 1996;56:2422–7.
27. Birkenkamp-Demtroder K, Christensen LL, Olesen SH, et al. Gene expression in colorectal cancer. Cancer Res 2002;62:4352–63.
28. Imoto I, Pimkhaokham A, Watanabe T, et al. Amplification and overexpression of TGIF2, a novel homeobox gene of the TALE superclass, in ovarian cancer cell lines. Biochem Biophys Res Commun 2000;276:264–70.
29. Phillips JL, Hayward SW, Wang Y, et al. The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. Cancer Res 2001;61:8143–9.
30. Hughes TR, Roberts CJ, Dai H, et al. Widespread aneuploidy revealed by DNA microarray expression profiling. Nat Genet 2000;25:333–7.
31. Crawley JJ, Furge KA. Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. Genome Biol 2002;3:RESEARCH0075.