GENES, CHROMOSOMES & CANCER 46:75–86 (2007)

Prediction of Chromosomal Aneuploidy from Gene Expression Data

Libi Hertzberg,[1,2,3] David R. Betts,[4] Susana C. Raimondi,[5] Beat W. Schäfer,[4] Daniel A. Notterman,[6,7] Eytan Domany,[3] and Shai Izraeli[1,2*]

[1]Department of Pediatric Hemato-Oncology, The Sheba Cancer Research Center, Sheba Medical Center, Tel Hashomer, Israel
[2]Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel
[3]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel
[4]Department of Oncology, University Children's Hospital, CH-8032 Zürich, Switzerland
[5]Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN
[6]Department of Pediatrics, Robert Wood Johnson Medical School, New Brunswick, NJ
[7]Department of Molecular Genetics, Robert Wood Johnson Medical School, New Brunswick, NJ

Chromosomal aneuploidy is commonly observed in neoplastic diseases and is an important prognostic marker. Here we examine how gene expression profiles reflect aneuploidy and whether these profiles can be used to detect changes in chromosome copy number. We developed two methods for detecting such changes in the gene expression profile of a single sample. The first method, fold-change analysis, relies on the availability of gene expression data from a large cohort of patients with the same disease. The expression profile of the sample is compared with that of the dataset. The second method, chromosomal relative expression analysis, is more general and requires the expression data from the tested sample only. We found that the relative expression values are stable among different chromosomes and exhibit little variation between different normal tissues. We exploited this novel finding to establish the set of reference values needed to detect changes in the copy number of chromosomes in a single sample on the basis of gene expression levels. We measured the accuracy of the performance of each method by applying them to two independent leukemia datasets. The second method was also applied to two solid tumor datasets. We conclude that chromosomal aneuploidy can be detected and predicted by analysis of gene expression profiles. This article contains Supplementary Material available at http://www.interscience.wiley.com/jpages/1045-2257/suppmat.

INTRODUCTION

Numerical and structural abnormalities of chromosomes are frequently detected in cancer and constitute a key mechanism in cancer progression (Albertson and Pinkel, 2003; Vogelstein and Kinzler, 2004). The level of cellular proto-oncogenes may be enhanced by genomic amplification (e.g., *MYCN* in neuroblastoma), and tumor suppressor genes may be inactivated by deletion (e.g., the retinoblastoma gene *RB1* on chromosome band 13q14). However, whether whole-chromosome aneuploidy (i.e., the gain or loss of a chromosome) has a central role in tumor initiation (Nowak et al., 2002; Duesberg and Li, 2003) or is a consequence of the genomic instability of cancer cells (Zimonjic et al., 2001; Wang et al., 2004) remains unknown. Genomic amplification of a region of DNA on a specific gene such as *MYCN* could be as much as a 300-fold increase (Brodeur et al., 1984), which would dramatically affect the expression level of that gene. The biological mechanism that causes whole-chromosomal aneuploidy is different from that which causes DNA amplification, which typically duplicates (or deletes) one copy of either the whole chromosome or a large segment of it. The effect of whole-chromosome aneuploidy on gene expression also differs from that of DNA amplification in that it potentially affects the expression of all of the genes on the chromosome but in a moderate manner (Schoch et al., 2005; Tsafrir et al., 2006).

Recent studies support a crucial role for whole-chromosomal aneuploidy in cancer pathogenesis (Hanks et al., 2004; Hernando et al., 2004; Izraeli, 2005; Teixeira and Heim, 2005). Leukemias may be defined as a group of malignant diseases in which genetic abnormalities in a hematopoietic cell give rise to clonal proliferation. Acute lymphoblastic leukemia (ALL) is the most common childhood neoplasm. Among the most common subtypes of child-

Supported by: Ministry of Science, Israel; Wolfson Family Charitable Trust, London; NCI/NIH; Grant number: P01 CA 65930-06; Mr. Curtis Katz.

*Correspondence to: Shai Izraeli, Department of Pediatric Hemato-Oncology, Sheba Medical Center, Tel-Hashomer 52621, Israel. E-mail: sizraeli@sheba.health.gov.il

Received 1 July 2006; Accepted 18 September 2006

DOI 10.1002/gcc.20391

Published online 16 October 2006 in
Wiley InterScience (www.interscience.wiley.com).

hood, ALL is high-hyperdiploid (HD) ALL (so called because the modal number of the leukemic cells is >50 chromosomes), which is associated with a good prognosis (Pui et al., 2001). Trisomies of specific chromosomes are also of prognostic value (Heerema et al., 2000; Moorman et al., 2003). Therefore, it is important to identify HD ALL, which can be detected by cytogenetic analysis and determination of the DNA index of the leukemic cells. However, cytogenetic analysis may be unsuccessful, especially in multicenter settings, and is not routinely performed on most solid tumors.

Gene expression profiling is being evaluated for implementation in the routine diagnostic workup of leukemias and solid tumors. Recent studies reported an association between numerical chromosomal aberrations and gene expression levels (Hughes et al., 2000; Virtaneva et al., 2001; Pollack et al., 2002; Nigro et al., 2005). They suggest that when a chromosome is missing or extra copies are present, the expression of its genes may be enhanced or decreased. However, it is not clear what proportion of genes is expressed in concordance with gene dosage (Phillips et al., 2001; Platzer et al., 2002). If this proportion is large enough, then numerical chromosomal aberrations could be identified from the gene expression profile (GEP). Schoch et al. (2005) explored the effect of chromosomal aneuploidy on gene expression in acute myeloid leukemia (AML). For the aberrant chromosomes, they calculated the proportion of genes with fold changes in expression, either greater than (duplication) or less than (deletion) one. The fold change in expression was calculated by dividing the average expression level of each gene in the sample with aberrational chromosomes by its average expression level in the normal karyotype samples. Most of the genes (67–87%) showed fold changes greater than one. Note that this does not mean that 67–87% of the genes were affected by the aneuploidy; we would expect 50% of the genes to show this much change in expression by chance. However, this finding suggests that GEP analysis detects chromosomal aneuploidy.

Recently, Crawley and Furge (2002) presented a method for identifying cytogenetic changes in hepatocellular carcinoma GEPs. They used a "sign test" to detect expression biases associated with whole chromosomes (or whole chromosomal arms). In their study, the regions with copy number changes that were found in >35% of the samples were compared with modifications in regions that were detected by comparative genomic hybridization (CGH) in previous studies. However, the accuracy of detection of such regions in a single sample was not measured. Tsafrir et al. (2006) investigated the relationship between DNA copy number (as measured by array CGH and using single nucleotide polymorphism [SNP] chip arrays) and gene expression in colon cancer. They showed that although the relation between gene copy number and expression level is complex and noisy when viewed on the single-gene level, when examined on a larger scale (by smoothing both types of data), a clear correlation between these factors is observed. Yi et al. (2005) performed two statistical tests to find genes that were differentially expressed in samples of tumor and normal tissues. The chromosomes were scanned using sliding-window analysis to detect regions with significant, differential gene expression. In each window, the scores of the statistical tests for the different genes were summed, and the region was detected if the scores exceeded a certain empirically determined threshold. This method was used to detect potential regions of genomic alterations in prostate cancer gene expression data. One of the regions they found is well known for high-frequency deletions; however, this detection method is based on analysis of a group of samples and cannot be used to detect chromosomal alterations in a single sample.

To what extent can changes in whole-chromosome (or chromosome arm) copy number be identified by GEP analysis of a single sample? Here we present a systematic analysis of this unresolved question. We explore two approaches for detecting changes in chromosomal copy number from GEP analysis at the single-sample level. One approach relies on the availability of a large gene expression dataset of patients with the same disease, and the other, more general method relies on analysis of expression data from the tested sample alone. We measured the accuracy of both methods by applying them to two independent leukemia datasets that included corresponding cytogenetic information. The second method was also applied to two solid tumor datasets. To the best of our knowledge, this is the first report of a reliable method that detects changes in chromosome copy number via GEP analysis of a single sample.

MATERIALS AND METHODS

Gene Expression Datasets

The following six gene expression datasets were measured on the Affymetrix HG-U133A array and are analyzed in this paper.

1. St. Jude dataset comprising 132 ALL samples, 18 of which were HD >50 (Ross et al.,

2003). Cytogenetic information on copy number(s) of chromosome 21 was available for all samples in this dataset (see Supplementary Table 1).

2. Johns Hopkins Hematopoietic Stem Cells (Johns Hopkins SC) dataset included 6 normal hematopoietic stem cell samples (Georgantas et al., 2004).
3. Colon cancer dataset containing samples of normal lung ($n = 5$), liver ($n = 10$), and colon ($n = 24$) and samples of colon polyps ($n = 30$), colon tumors ($n = 114$), and metastasized tumors ($n = 19$) (Tsafrir et al., 2006).
4. Scripps dataset of 158 normal tissue samples obtained from 79 tissues (Su et al., 2004).
5. Zurich dataset included 20 ALL samples and cytogenetic data for all chromosomes (see Supplementary Table 2) (Betts and Schaefer, unpublished).
6. Human Stem Cells (Human SC) dataset included 17 samples of embryonic stem cells, keratinocyte stem cells, and hematopoietic stem cells (Golan-Mashiach et al., 2005).

The following two gene expression datasets were measured on the Affymetrix HG-U95Av2 array:

1. Uveal melanoma dataset included 20 samples of uveal melanoma; 10 of the samples had one copy of chromosome 3 (monosomy 3); the other 10 samples had 2 copies of it (disomy 3) (Tschentscher et al., 2003).
2. Scripps-U95 dataset contains 58 samples from 29 normal tissues (Su et al., 2002).

Preprocessing the datasets

The datasets that were measured on the Affymetrix HG-U133A array were normalized by Affymetrix Microarray Suite 5.0 (MAS 5.0) software. The datasets that were measured on the Affymetrix HG-U95A array were normalized by MAS 4.0 software. Each dataset was filtered by omitting probe sets that got a "present" call in two or fewer of the samples. A threshold was applied by setting expression levels at 30 for any probe set with expression level less than or equal to 30 in a particular sample.

When there were two or more probe sets for the same gene (defined as those with the same gene symbol), their expression levels were combined as follows. The Pearson correlation coefficient was calculated for each pair of probe sets in the group; then the largest subgroup in which each pair of probe sets had a correlation coefficient higher than 0.5 was found by simple scanning. If the size of the chosen subgroup was larger than 1, then the average expression level of that group of probe sets represented that of the gene in each sample. Otherwise, one probe set was chosen to represent the gene: first, all probe sets without '_s_' or '_x_' in their identity were found ('_s_' or '_x_' indicates probe sets that are less specific). If there were no such probe sets, one probe set was chosen randomly. Otherwise, the probe set without '_s_' or '_x_' and with the largest number of present calls throughout the samples was chosen. After this process, each gene that passed the filter had one expression value in each sample.

Identifying Differentially Expressed Genes

To find genes that differentiate two groups of samples in a given dataset, we used the Wilcoxon rank-sum test for each gene. To deal with the problem of multiple comparisons, we applied the false discovery rate (FDR) method (Benjamini and Hochberg, 1995) to the resulting P-values (parameter $Q = 0.05$).

To detect overrepresented chromosomes in the group of differentiating genes, we calculated a P-value for each chromosome. The P-value reflects the extent to which that group of genes is enriched on a particular chromosome. Specifically, for each chromosome i (i = 1–22, X, and Y), we calculated the number of genes, n_i, and the number of differentiating genes, dn_i. Then we calculated a hypergeometric P-value for getting dn_i or more genes from this chromosome in the differentiating group of genes, assuming it is a random group of genes.

Fold-Change Analysis

To identify extra chromosomes in a single sample from a large dataset, we assume that when a gene is located on a chromosome that has more than two copies, its expression level will be increased. Therefore, the basis of this analysis is calculating for each gene its fold-change value, f_g, which is defined as the expression level of gene g in the given sample divided by its median expression level in all samples included in the dataset. Two analyses were done. First, the median fold change for each chromosome i = 1–22, X and Y, in the sample was calculated as follows: $\overline{f_i} = \text{median}(\{f_g | g = 1, \ldots, n_i\})$. The value $\overline{f_i}$ is indicative of the likelihood that chromosome i has extra copies in the given sample. The more $\overline{f_i}$ exceeds 1, the greater the likelihood that an additional copy (or copies) of chromosome i is present.

Second, the fold-change P-value for each chromosome $i = 1$–22, X and Y, was calculated. We determined the P-value for the number of upregulated genes. The null assumption is that for each gene g from chromosome i ($g = 1, \dots, n_i$), $P(f_i > 1) = 0.5$. Let A_i be the group of genes whose fold-change value is larger than 1. Then $A_i = \{gf_g > 1, g = 1, \dots, n_i\}$. Let N_i be the size of this group. Then a P-value, p_i, is assigned to each chromosome by calculating the probability that this group is of size N_i or more, by using the binomial distribution, under the null assumption. Like $\bar{f}_i$, p_i reflects how likely it is that chromosome i has extra copies in the given sample but in a slightly different way. The smaller the value of p_i, the more likely it is that chromosome i has extra copies in the sample.

Fold-change prediction of extra chromosomes

We predict that chromosome i has extra copies in a given sample only if $\bar{f}_i$ and p_i both satisfy their respective conditions. Specifically, if p_i is lower than a cutoff value and $\bar{f}_i$ is higher than another cutoff value, then we predict that the chromosome has extra copies in the sample. We used the following sets of cutoff values: fold change at 1.01, 1.02, …, 1.49, 1.50 and P-value at 0, 0.01, …, 0.39, 0.40. For each pair of fold-change and P-value cutoff values, we made a prediction for each chromosome. Then we measured the quality of these predictions.

Fold-change detection of missing chromosomes

The analysis described above can be also applied to detect missing chromosomes. The only change is that the fold-change value (or median fold-change value) should be <1 for missing chromosomes.

Chromosomal Relative Expression Analysis

The chromosomal relative expression (CRE) is calculated for each chromosome i as follows: For each sample α, we calculate the median expression level of the genes in the chromosome. We then divide each of those individual median values by the median expression level of all genes in the sample. The log-relative expression level of chromosome i in sample α is determined by performing the logarithm (base 2) to get $l_{i,\alpha}$. We then calculate the mean, a_i, and SD, s_i, of the $l_{i,\alpha}$ in all samples.

Because the number of genes on chromosome Y is small (i.e. 49 probe sets), we excluded it from this analysis. The output is two vectors of size 23, $\bar{a} = (a_1, \dots, a_{22}, a_X)$ of the mean values and $\bar{s} = (s_1, \dots, s_{22}, s_X)$ of the SD values. For each chromosome, the distribution of the log-relative expression level is close to Gaussian (Supplementary Fig. 1). For a given sample, α, the CRE vector contains the 23 values $l_{i,\alpha}$. This analysis yields for each dataset a pair of CRE vectors, $\bar{a}$ and $\bar{s}$. The datasets were combined, and all samples were analyzed together, yielding a "representative" CRE.

CRE prediction of chromosomal gains and losses

Using the representative CRE mean and SD vectors, we can predict whether a chromosome in a given sample has extra copies or is missing. We calculate the CRE vector ($l_{1,\alpha}$, $l_{2,\alpha}, \dots, l_{X,\alpha}$) of the given sample α, as described above. Then a P-value is assigned to each chromosome by calculating the probability of its log-relative expression value being higher (or lower) than an effective Gaussian distribution, with mean and SD values of the representative CRE in this chromosome. The smaller the P-value, the more confident we are of our prediction. After calculating a P-value for each chromosome, we predict that a chromosome is extra (or missing) in the sample if its P-value is lower than a predetermined cutoff value that is applied to the analysis of all of the chromosomes. We used the following set of P-value cutoffs: 0.001, 0.002, …, 0.149, 0.150.

Combining the Fold-Change and Chromosomal Relative Expression Analyses

For each sample, the predictions by both methods are combined by applying an AND gate: for each chromosome, if both methods predict extra copies, then that chromosome is indeed predicted to have extra copies; if neither or only one method predicts extra copies, then no extra copies of that chromosome are predicted. As described above, the fold-change prediction is made according to two cutoff values, and the CRE prediction is made according to one cutoff value. The combined prediction is therefore made according to all three of those cutoff values.

Quality of the Prediction

To measure the quality of the prediction, we assume that there is a large group of samples. Each prediction is made for all of the samples together (and for all of the chromosomes). Now we look at the set of all "points" (i.e. "the number of samples multiplied by the number of chromosomes analyzed"). For example, the St. Jude dataset contains 132 points (only chromosome 21 was analyzed). According to conventional cytogenetics, we define positives (P) in the sample as points in which extra copies of the chromosome are present, and negatives (N) as points in which they are not. We consider the pair a true positive (TP) if a chromosome is predicted to be added in a specific sample and the cytogenetic data agree. The pair is considered a false positive (FP) if a chromosome is predicted to be added in a specific sample and the cytogenetic data do not agree. In the same way, we define true negatives (TN) and false negatives (FN). TP/(TP + FN) is called the TP rate (TPR), and FP/(FP + TN) is called the FP rate (FPR).

For each prediction made, we calculated its TPR and FPR. The tradeoff between these rates is presented in a "receiver operating characteristic" (ROC) curve. The TPR and FPR values are obtained by varying the parameters of the predictor (i.e. the cutoff values). Therefore, we can choose the cutoff values that result in a prediction of a specific accuracy (i.e. at specific values of TPR and FPR).

RESULTS

Extra Chromosome Copies are Reflected in the Gene Expression Profile

Our analysis of the St. Jude dataset to examine the effect of extra chromosome(s) on gene expression revealed that the genes that differentiate the HD > 50 samples from the rest of the samples come from the extra chromosomes. We calculated a *P*-value for each chromosome to determine its enrichment in genes that differentiated these groups (Fig. 1). Genes residing on chromosomes 4, 6, 10, 14, 18, 21, and X were significantly enriched (the standard FDR method was applied to the 24 *P*-values; they pass it when parameter $Q = 0.08$). These seven chromosomes are the ones most commonly involved in trisomies of HD > 50 ALLs (i.e. trisomies that appear in >50% of patients with HD > 50 ALL) (Teixeira and Heim, 2005). Chromosome 17 also frequently undergoes trisomy in HD > 50 ALL, but it did not achieve a significant *P*-value in our analysis. However, the same analysis of differentiating genes whose average level of expression is higher in HD > 50 samples than in others did identify chromosome 17. These findings indicate that the group of differentiating genes is enriched with genes from the extra chromosomes. In addition, as might be expected, the expression levels of multiple genes located on extra chromosomes are usually higher than that seen when no extra chromosomes are present (Supplementary Fig. 2).

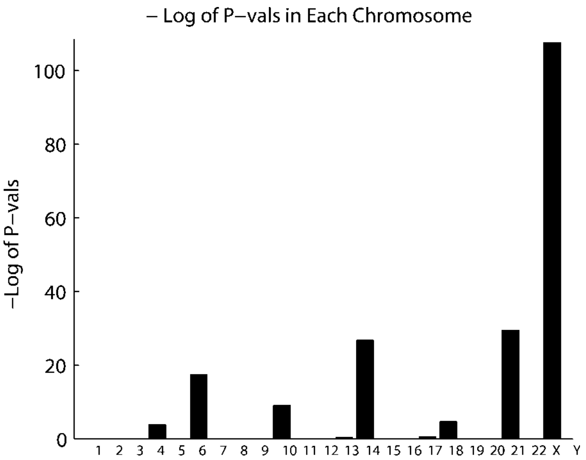

Figure 1. *P*-values for enrichment of chromosomes in the differentiating group. The *x*-axis represents the chromosomes, and the *y*-axis represents the negative logarithm of the chromosomes' enrichment *P*-values.

Gene Expression Profile Identifies Changes in the Number of Chromosomes in a Single Sample

Chromosomes with frequent trisomies can be detected on the basis of the GEP from several samples; however, does a single sample contain enough information to allow us to identify extra chromosomes? This question is clinically relevant, because clinical laboratories routinely need to analyze individual samples that are not part of a large study group. In our first approach to addressing this question, we assumed that the sample was part of a large experiment and compared its expression data with that obtained in the large experiment. In our second approach, we assumed that there were no similar expression data with which to compare the sample. In both of these approaches, we measured the accuracy of the detection.

Comparison with a large gene expression dataset: Chromosomal expression fold-change analysis

First we compared the expression level of each gene in the sample with that of the same gene in the dataset by calculating its fold change. This comparison may indicate whether a gene is located on an extra chromosome. Because the dataset could have included samples that contain additional chromosomes, we assumed that there was no information available that identified the samples that contained numerical chromosome changes. This approach made the analysis generally applicable. To deal with this problem, the dataset must be large; thus, we expected that for all chromosomes, most of the samples did not contain extra copies or deletions. By using the median gene expression

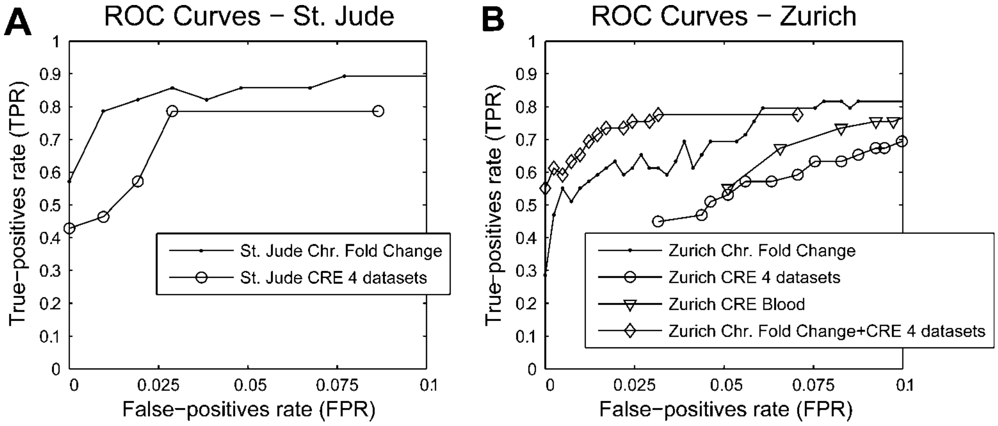

Figure 2. ROC curves for the St. Jude and Zurich datasets. The *x*-axis represents the false-positives rate, and the *y*-axis represents the true-positives rate. A: Fold-change and CRE analyses of chromosome 21 aneuploidy in the St. Jude dataset. B: Fold-change analysis, CRE analysis, CRE analysis based only on normal blood samples, and the combined analyses (i.e. fold change and CRE) of chromosomes 1–22 and X in the Zurich dataset.

TABLE 1. Fold-Change Analysis of Chromosome 21 Aneuploidy in 132 ALL Samples from the St. Jude Dataset

Group (*n*)	No. of samples with chr 21 aneuploidy[a]	TPR (%)[b]	No. of samples without chr 21 aneuploidy[a]	FPR (%)[b]
HD > 50 (18)	18	17 (94)	0	0 (0)
Others (114)	10	7 (70)	104	3 (2.9)

Abbreviations: chr, chromosome; FPR, false-positives rate; HD, high-hyperdiploid; TPR, true-positives rate.

[a] Aneuploidy of chromosome 21 was determined by cytogenetic analysis.

[b] The true-positives rate and false-positives rate of chr 21 aneuploidy was predicted by fold-change analysis.

level in the fold-change calculation, we minimized the effect of additional chromosomes in a minority of the samples.

We analyzed each of the 132 St. Jude samples as individual samples; the whole dataset served as the comparative dataset. We obtained the cytogenetic copy number information for chromosome 21 in each of the 132 samples; thus, the analysis was confined to chromosome 21, and the two factors, median fold change and fold change *P*-value, were calculated for each sample to predict whether any contained extra copies of chromosome 21. The quality of the prediction is represented by an ROC curve (Fig. 2).

For an FPR value of 2.9%, the TPR value was 86% (Table 1). Thus, of 28 samples that had extra copies of chromosome 21, 24 (86%) were identified by this method. The rate of identification was higher for HD > 50 samples: 17 (94%) of the 18 were identified. The actual rate is even higher, because the two *BCR*-ABL$^+$ samples with an extra chromosome 21 were also HD > 50, and both were captured. Aneuploidy is therefore more easily predicted in HD > 50 samples. This finding may be explained, at least in part, by the presence of at least two extra copies of chromosome 21 in about 70% of the HD > 50 samples. In contrast, <30% of the non-HD > 50 samples containing additional copies of chromosome 21 had more than one extra copy.

We have shown that, given a gene expression dataset that can be used as a reference and a single test sample, we can accurately predict whether there are extra copies of a specific chromosome in a given sample.

Analysis of a single sample: CRE analysis

If a sample contains an extra copy (or copies) of a chromosome, we expect that chromosome to show a median expression level higher than that of the other chromosomes. However, even in cytogenetically normal samples, the expression levels of some chromosomes may not be similar. For example, the median expression level of all genes on chromosome 1 is not equal to the that of all genes on chromosome 2 (Caron et al., 2001). Because chromosomes generally have different median expression values, we evaluated their relative expression values (i.e. the ratios of median chromosomal expression levels). If the relative expression level of chromosomes is constant in normal sam-

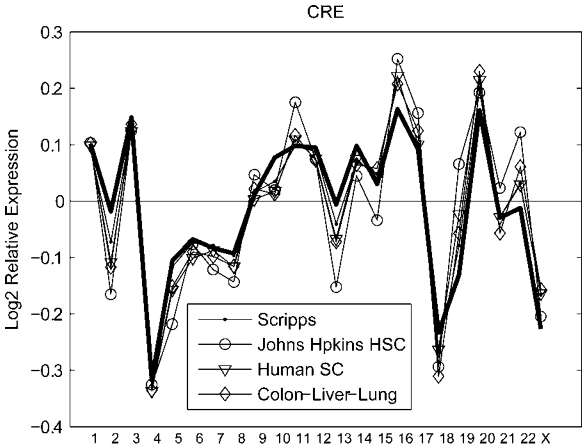

Figure 3. Mean CRE values of four datasets of samples with normal cytogenetics. The *x*-axis represents the chromosomes 1–22 and X, and the *y*-axis represents the logarithm (base 2) of the mean CRE value. The bold line shows the representative normal CRE calculated on the basis of the datasets indicated.

ples, this information could be used to identify extra chromosomes in a given sample.

We calculated the CRE for normal samples in each of four datasets. The mean CRE values were similar (although not identical) throughout different, cytogenetically normal tissue groups from different experiments (Fig. 3). The pairs of datasets were strongly correlated (average Pearson correlation coefficient, $r = 0.96$) (Supplementary Table 3). When two values deviated by 0.2, their ratio was still <1.15 (because $\log_2$ values are plotted on the *y* axis).

We then calculated a representative "normal" CRE vector (Fig. 3 heavy line). Chromosomal copy number in a single sample was determined by comparing the sample with the normal CRE vectors for each chromosome. If the sample's CRE for a specific chromosome deviated from its representative CRE, we predicted that extra copies of that chromosome were present in the sample. This analysis was performed only on chromosome 21 from each sample in the St. Jude dataset (Fig. 2). The copy number prediction was less accurate than that of the fold-change analysis. For example, for an FPR of 2.9%, the TPR was 79% (the *P*-value cutoff was 0.06). At this FPR, 22 of 28 (79%) samples with extra copies of chromosome 21 were identified. As in expression fold-change analysis, the rate of identification was higher for HD > 50 samples: 17 of the 18 (94%) HD > 50 samples, all of which contained an extra copy (or copies) of chromosome 21, were identified. Therefore, the presence of extra chromosome(s) can be accurately predicted from the GEP of a single sample without the use of a large comparative dataset.

Testing the Methods on all Chromosomes in an Independent Leukemia Dataset

To validate our novel methods for predicting the copy number of any chromosome, we analyzed the copy numbers of all chromosomes in a separate gene expression dataset of 20 ALL samples from Zurich.

Chromosomal expression fold-change analysis

Each of the 20 samples was analyzed as if it was a single, isolated sample. The St. Jude dataset served as the comparable control. The resulting ROC curve (Fig. 2) showed that although the accuracy of the prediction was good, it was lower than that obtained by comparing a sample from the St. Jude dataset with the whole St. Jude dataset. This lower accuracy may result from comparing data from different experiments. Additionally, in the St. Jude dataset, we looked only at changes in chromosome 21, which often has four extra copies in HD > 50 samples, whereas here we looked at changes in all chromosomes, which usually have only one extra copy (Supplementary Table 2).

As shown in Figure 2, when the FPR was 6.1%, the TPR was 80%; this performance is realized for median fold-change cutoff value of 1.08 and *P*-value cutoff of 0.08. The prediction of additional chromosomes by using these FPR and TPR values and the cytogenetics information is presented in Supplementary Figure 3.

CRE analysis

We performed the CRE analysis on each sample from the Zurich dataset, and again, the accuracy of the prediction was lower than that obtained by using the St. Jude dataset (Fig. 2). As in the fold-change analysis, we explain this loss of accuracy by the fact that we were looking at all chromosomes rather than just chromosome 21. Figure 3 shows that there were chromosomes for which the CRE was very stable (e.g., chromosomes 1, 8, 12), but there were also chromosomes for which the CRE varied across tissues (chromosomes 2, 13, 22). Therefore, this method performs better when applied only to chromosome 21.

In Figure 2, we also see that the accuracy of the prediction obtained with the CRE analysis is lower than that obtained with the fold-change analysis. One possible explanation for this observation is the fact that for some chromosomes, the CRE is variable, as mentioned above. Alternatively, the loss of accuracy may be attributed to differences in the comparison data sets. In the fold-change analysis, we compared ALL samples from the Zurich dataset

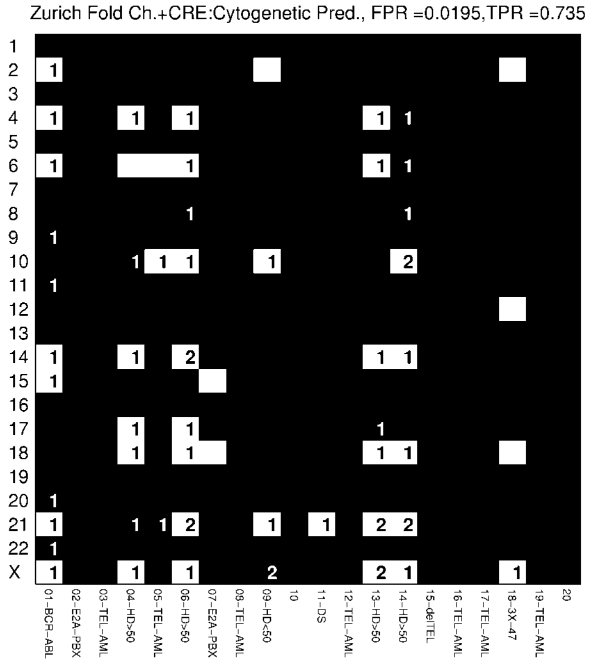

Figure 4. Predictions of aneuploidy made using cytogenetic information compared with those made using the combined fold-change and CRE methods. The prediction was based on the analysis of the Zurich dataset by combining the fold-change and CRE methods; the fold-change analysis used the St Jude dataset as a reference. The *x*-axis lists the samples, and the *y*-axis lists the chromosomes. Predicted additional chromosomes are white. Numbers show the cytogenetic information (i.e. the number of extra copies of the chromosome in the sample). FPR = 2% and TPR = 74%.

with ALL samples from St. Jude dataset, whereas in the CRE analysis, we compared the Zurich dataset with samples from different normal tissues.

To examine this possibility, we performed the CRE analysis by using only normal blood samples from the Johns Hopkins HSC, the Human SC, and the Scripps datasets. The results of this analysis were only slightly more accurate than those of the general CRE analysis (Fig. 2). Therefore, we conclude that when investigating all chromosomes, composing a representative CRE from only relevant normal tissue samples will not necessarily improve the results.

Combining the CRE Analysis with the Fold-Change Analysis

Because the CRE and fold-change analysis methods are based on different approaches, it may be beneficial to combine them when possible. When we combined the predictions of these two methods for the Zurich dataset, the accuracy of the prediction substantially improved (Fig. 2). For example, with an FPR of 2%, the TPR was 74%. The predicted chromosomal copy number is compared with the cytogenetic information in Figure 4.

Discrepancies Between Gene Expression Profile and Cytogenetic Analyses

To understand potential causes of discrepancy between the GEP-based calculation of chromosomal copy number and the cytogenetic data, we re-examined the cytogenetic information on the discrepant predictions shown in Table 1 and Figure 4.

In the St. Jude dataset, there were four samples in which the bioinformatics approach failed to identify extra copies of chromosome 21 that were detected by cytogenetic analysis. For one sample (T-ALL-C7, Supplementary Table 2), a repeated fluorescence in situ hybridization (FISH) analysis demonstrated two copies of chromosome 21, in agreement with the GEP analysis. Thus, the TPR of the prediction described in Table 1 increased to 96% (FPR = 2.8%) or to 92% for the CRE analysis at the same FPR. In two of the other three samples, the abnormal clone composed only a small proportion (60% or less) of the sample and possibly contributed less to the overall gene expression level. This small contribution could explain why it was missed by the bioinformatics approach. However, our method successfully identified four other samples containing only a small amount of the abnormal clone (Supplementary Table 1).

Nine samples in the Zurich dataset showed discrepancy between the cytogenetics and bioinformatics findings (Supplementary Table 2). Of those, the discrepancies in three samples (1, 4, and 9) can be explained. In sample 1, cytogenetic analysis showed extra material of chromosomes 9, 20, and 22 that was not reflected in the gene expression analysis. However, re-examination of the cytogenetics findings showed that the gains of der(9) and 20 were present only in a subclone of 33% of the cells, and the gain of the der(22) represented a gain only of 22pter→22q11.2 in a subclone of 67% of the cells. In sample 4, cytogenetics analysis showed an extra copy of chromosome 10 that was not reflected in the gene expression analysis. Re-examination of the cytogenetics findings showed that the extra material resulted from a partial gain of 10pter→10q22, so a substantial portion of chromosome 10 was not aneuploid. In sample 9, cytogenetic analysis showed two extra copies of chromosomes X, and the prediction missed it. For this sample, FISH analysis demonstrated an extra copy of chromosome X in 60% of the cells. The cytogenetics findings were confirmed by FISH analysis in an additional four samples (8, 13, 15, and 18). Taking into account the revised cytogenetics findings, the TPR of the bioinformatics prediction shown in Figure 4 is adjusted to 84% for an FPR of 1.9%.

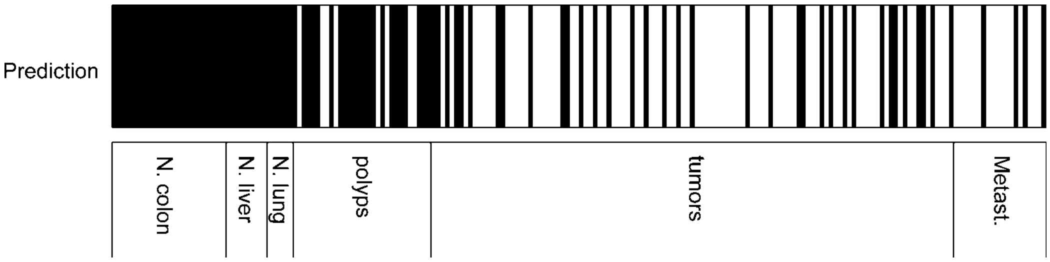

Figure 5. Prediction of chromosome 20 aneuploidy in the colon cancer dataset by CRE analysis. The prediction used a *P*-value cutoff of 0.06. The *x*-axis lists the sample groups of the colon cancer dataset. Samples in which chromosome 20 is predicted to be at least duplicated are white; the others are black. The CRE analysis was composed of normal samples of colon, liver, and lung taken from the colon cancer and Scripps datasets.

CRE Analysis of Two Solid Tumor Datasets

To further evaluate the applicability of the CRE method, we applied it to two solid tumor datasets: the first is a colon cancer dataset that contains samples with whole chromosomal arm amplifications; the second is a uveal melanoma dataset that contains samples with whole-chromosome losses.

Colon cancer dataset

Amplification of the long arm of chromosome 20 (20q) is a known marker of progression of colon cancer, and overexpression of E2F1 in lung and liver metastases of human colon cancer is associated with gene amplification (Iwamoto et al., 2004). Most of the genes on chromosome 20 reside on the q arm (i.e. 401 of the 597 chromosome 20 probe sets on the HG-U133A array recognize genes on the q arm). Thus, aneuploidy in 20q should result in a detectable change in gene expression. We performed CRE analyses of chromosome 20 in the colon cancer dataset and in samples from normal colon, liver, and lung tissues and composed the representative CRE from samples of normal epithelial, colon, liver, and lung tissues. We used the same *P*-value cutoff as that used for the St. Jude data, with an FPR of 2.9% and TPR of 79% ($P = 0.06$). The percentage of samples with predicted duplication of chromosome arm 20q correlated with progression of the malignant phenotype (Fig. 5). No duplication was predicted in the normal samples; however, duplications of 20q were predicted in 26% of polyp samples, 69% of tumor samples, and 79% of metastases samples. These results confirm previous findings from the same samples (Tsafrir et al., 2006), that is, the percentage of 20q duplications appeared to increase with disease progression.

Unlike our approach, the analysis done (Tsafrir et al., 2006) was not on the single-sample level but on groups of samples of the same disease progression state. For validation of their results, FISH analysis was done on six tumor samples (Tsafrir et al., 2006); 5 of 6 samples gave clear results. Per our CRE prediction, 2 of the 5 samples showed no amplification of 20q, and 3 showed clear amplification of the region. Thus, the CRE prediction was confirmed by FISH analysis.

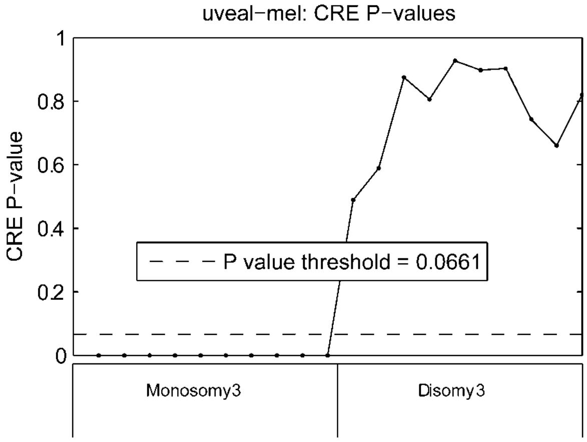

Figure 6. Prediction of monosomy 3 in the uveal melanoma dataset by CRE analysis. The *x*-axis lists the sample groups of the uveal melanoma dataset. The *y*-axis lists the CRE *P*-values for each sample. The prediction used a *P*-value cutoff of 0.06. The *P*-values of the monosomy 3 samples are consistently <0.06, and those of the disomy 3 samples are >0.06. The CRE analysis was composed of normal samples of 29 normal tissues taken from the Scripps dataset.

Uveal melanoma dataset

Uveal melanoma is the most common intraocular malignancy. About 50% of patients die of metastasis, which almost exclusively originates from primary tumors that have lost one copy of chromosome 3 (Tschentscher et al., 2003). Therefore, the detection of monosomy 3 is clinically important. To examine the performance of our CRE method in detection of whole-chromosomal loss, we applied it to chromosome 3 in a uveal melanoma dataset of 20 samples; 10 contained monosomy 3. The composition of probe sets on the gene expression array is important for the CRE calculation; therefore, we measured the 58 samples of normal tissues from the Scripps-U95 dataset on the same array as the uveal melanoma dataset (HG-U95Av2). We used the same *P*-value cutoff as that used for the St. Jude data, with an FPR of 2.9% and TPR of 79% ($P = 0.06$). The CRE prediction was 100% accurate for the 20 uveal melanoma samples (Fig. 6). Therefore, we conclude that CRE analysis of GEP is applicable to solid tumors and to detection of both chromosomal gains and losses.

DISCUSSION

We have developed two methods that predict prognostically significant chromosomal aneuploidies that are discernible in GEPs: fold-change analysis and the CRE method. Fold-change analysis is more accurate than the CRE method, but it requires the use of a large dataset of the same disease, or at least the same tissue type, to which the test sample can be compared. In the fold-change analysis, each gene's expression level in the test sample is compared with the same gene's median expression level in all samples from a closely related dataset. This comparison minimizes the variability in the expression of a single gene across datasets and in different tissues, regardless of numerical chromosomal changes.

Although the CRE method is less accurate, it is generally more applicable, because it does not require a comparable dataset. CRE analysis examines the average expression levels for all genes on the chromosome and relates that value to the overall expression level in the sample; thus, the analysis is meaningful, even if different tissue types were used to construct the reference CRE.

As was demonstrated in our comparison of the fold-change analysis with the CRE analysis, there is a tradeoff between accuracy and generality. Combining the two methods improves the accuracy of prediction; this observation implies that the two methods generate different types of errors. In the fold-change analysis, errors are not specific to a chromosome, but they might be specific to a sample. The errors generated by the CRE method are caused by the variability of the relative expression of some chromosomes in different samples and tissues. Although the CRE method can always be applied to an isolated sample, if a comparable dataset of the same disease exists, we recommend combining these two analyses. Although the level of expression of specific genes greatly varies among different tissues, we found that the relative chromosomal expression is stable across various normal tissues. This novel finding is the basis of the CRE method, and it suggests that the regulation of the expression of the genes on a specific chromosome is somehow averaged to maintain the same relative expression of those genes throughout normal tissues.

We also addressed the detection of whole-chromosome (or whole-arm) aneuploidy in a single sample. This type of aberration is common in many cancers (Teixeira and Heim, 2005), including hyperdiploid ALL and myeloma, monosomy 7 in AML, gain of chromosome 7 in brain tumor, and monosomy 3 in melanoma. The detection of specific chromosomal gains and losses is, therefore, clinically important. Our methods can also be applied to smaller genomic regions to obtain results of a higher resolution. However, the expression of a limited number of genes is affected by genomic amplification or deletion (Schoch et al., 2005; Tsafrir et al., 2006). Therefore, the resolution of an expression-based aneuploidy-detection method is also limited. In addition, the genes affected might differ across chromosomes (Schoch et al., 2005), and their location along a single chromosome may not be uniform. Thus, detection of aneuploidy in smaller chromosomal regions is a more complex task. Furthermore, to measure the accuracy of the detection, one must compare their results with genomic (CGH and SNP) array analysis of the same samples rather than with results from cytogenetic analysis.

A comparison of gene expression data from with that from genomic arrays to detect small regions of chromosomal amplifications or deletions is not straightforward, because the lengths and locations of the regions must be defined in both types of data. By examining whole chromosomes or whole arms, we simplify this task and facilitate the measurement of the accuracy of detection. In the case of smaller regions, accuracy of detection is less properly defined. In summary, determining the resolution limits of the methods proposed here will require a more complex analysis.

We have clearly shown that GEP analysis captures the prognostically significant aneuploidies in a tumor sample. Cytogenetic analysis has the advantage of identifying subclones with specific aneuploidies that may not be reflected in the GEP; the prognostic significance of such findings, however, is unknown. Nevertheless, most cases of HD ALL can be identified by either of the methods we have developed. Results from our tests of the CRE method applied to the colon cancer and uveal melanoma datasets indicated that this approach is also clinically relevant for detecting chromosomal aneuploidies in solid tumors, which are usually not amenable to cytogenetic analysis. The relevance of the CRE method is strengthened by the fact that gene expression profiling is being evaluated for implementation in the routine diagnostic workup on leukemias and solid tumors. We note that genomic (CGH and SNP) arrays are more suitable for detection of chromosomal aberrations than gene expression arrays. However, it appears that it will take more time for these analyses to be included in the routine diagnostic workup.

Finally, the potential importance of this work as a research tool is greater than its potential as a clinical tool. For example, many large gene expression datasets of multiple types of cancers already exist; these have good clinical annotation but no genomic data (i.e. no cytogenetics or CGH). Our methods could be used to draw a virtual map of chromosomal gains and losses for each patient. A large amount of data could also be reanalyzed using our methods to explore, for example, whether certain trisomies are associated with a particular outcome (bad or good). When the cytogenetics data do not agree with the gene expression data, another interesting question arises—What is more relevant for outcome, cytogenetic properties or gene expression? Is there a clinical difference between cases in which a specific aneuploidy is reflected in the gene expression profile and those in which it is not? The converse question is even more interesting—Can one elucidate a gene expression signature of chromosomal aneuploidy in the absence of actual excess or loss of an individual chromosome? Could an epigenetic process create a situation of "pseudomonosomy" or "pseudotrisomy"? These research questions can now be directly approached by using our novel methods that are capable of analyzing single samples. The relevance of our systematic approach to analyzing how chromosomal aneuploidy is reflected in gene expression will be further strengthened when more precise genomic tools become widely available for direct measurement of aneuploidy in the routine diagnostic workup.

ACKNOWLEDGMENTS

We thank Drs. Sheila Shurtleff, Francis Barany, Janet Davies, and Angela McArthur for their contributions. This study was done as a partial fullfilment of the requirement for the Ph.D. Degree of Libi Hertzberg, Sackler Faculty of Medicine, Tel-Aviv University.

REFERENCES

Albertson DG, Pinkel D. 2003. Genomic microarrays in human genetic disease and cancer. Hum Mol Genet 12:R145–R152.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc 57:289–300.

Brodeur GM, Seeger RC, Schwab M, Varmus HE, Bishop JM. 1984. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. Science 224:1121–1124.

Caron H, Schaik BV, Mee MV, Baas F, Riggins G, Sluis PV, Hermus MC, Asperen RV, Boon K, Vote PA, Heisterkamp S, van Kampen A, Versteeg R. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science 291: 1289–1292.

Crawley JJ, Furge KA. 2002. Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. Genome Biol 3:0075.1–0075.8.

Duesberg P, Li R. 2003. Multistep carcinogenesis: A chain reaction of aneuploidizations. Cell Cycle 2:202–210.

Georgantas RW III, Tanadve V, Malehorn M, Heimfeld S, Chen C, Carr L, Martinez-Murillo F, Riggins G, Kowalski J, Civin CI. 2004. Microarray and serial analysis of gene expression analyses identify known and novel transcripts overexpressed in hematopoietic stem cells. Cancer Res 64:4434–4441.

Golan-Mashiach M, Dazard JE, Gerecht-Nir S, Amariglio N, Fisher T, Jacob-Hirsch J, Bielorai B, Osenberg S, Barad O, Getz G, Toren A, Rechavi G, Itskovitz-Eldor J, Domany E, Givol D. 2005. Design principle of gene expression used by human stem cells: Implication for pluripotency. FASEB J 19:147–149.

Hanks S, Coleman K, Reid S, Plaja A, Firth H, Fitzpatrick D, Kidd A, Mehes K, Nash R, Robin N, Shannon N, Tolmie J, Swansbury J, Irrthum A, Douglas J, Rahman N. 2004. Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. Nat Genet 36:1159–1161.

Heerema NA, Sather HN, Sensel MG, Zhang T, Hutchinson RJ, Nachman JB, Lange BJ, Steinherz PG, Bostrom BC, Reaman GH, Gaynon PS, Uckun FM. 2000. Prognostic impact of trisomies of chromosomes 10, 17, and 5 among children with acute lymphoblastic leukemia and high hyperdiploidy (>50 chromosomes). J Clin Oncol 18:1876–1887.

Hernando E, Nahle Z, Juan G, Diaz-Rodriguez E, Alaminos M, Hemann M, Michel L, Mittal V, Gerald W, Benezra R, Lowe SW, Cordon-Cardo C. 2004. Rb inactivation promotes genomic instability by uncoupling cell cycle progression from mitotic control. Nature 430:797–802.

Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, Friend SH, Marton MJ. 2000. Widespread aneuploidy revealed by DNA microarray expression profiling. Nat Genet 25:333–337.

Iwamoto M, Banerjee D, Menon LG, Jurkiewicz A, Rao PH, Kemeny NE, Fong Y, Jhanwar SC, Gorlick R, Bertino JR. 2004. Overexpression of E2F-1 in lung and liver metastases of human colon cancer is associated with gene amplification. Cancer Biol Ther 3:395–399.

Izraeli S 2005. Perspective: Chromosomal aneuploidy in leukemia-lessons from Down syndrome. Hematol Oncol 24:3–6.

Moorman AV, Richards SM, Martineau M, Cheung KL, Robinson HM, Jalali GR, Broadfield ZJ, Harris RL, Taylor KE, Gibson BE, Hann IM, Hill FG, Kinsey SE, Eden TO, Mitchell CD, Harrison CJ. 2003. Outcome heterogeneity in childhood high-hyperdiploid acute lymphoblastic leukemia. Blood 102:2756–2762.

Nigro JM, Misra A, Zhang L, Smirnov I, Colman H, Griffin C, Ozburn N, Chen M, Pan E, Koul D, Yung WK, Feuerstein BG, Aldape KD. 2005. Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. Cancer Res 65: 1678–1686.

Nowak MA, Komarova NL, Sengupta A, Jallepalli PV, Shih Ie M, Vogelstein B, Lengauer C. 2002. The role of chromosomal instability in tumor initiation. Proc Natl Acad Sci USA 99:16226–16231.

Phillips JL, Hayward SW, Wang Y, Vasselli J, Pavlovich C, Padilla-Nash H, Pezullo JR, Ghadimi BM, Grossfeld GD, Rivera A, Linehan WM, Cunha GR, Ried T. 2001. The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. Cancer Res 61:8143–8149.

Platzer P, Upender MB, Wilson K, Willis J, Lutterbaugh J, Nosrati A, Willson JK, Mack D, Ried T, Markowitz S. 2002. Silence of chromosomal amplifications in colon cancer. Cancer Res 62:1134–1138.

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO. 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci USA 99:12963–12968.

Pui CH, Campana D, Evans WE. 2001. Childhood acute lymphoblastic leukaemia—Current status and future perspectives. Lancet Oncol 2:597–607.

Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK, Liu HC, Mahfouz R, Raimondi SC, Lenny N, Patel A, Downing JR. 2003. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood 102:2951–2959.

Schoch C, Kohlmann A, Dugas M, Kern W, Hiddemann W, Schnittger S, Haferlach T. 2005. Genomic gains and losses influence expression levels of genes located within the affected regions: A study on acute myeloid leukemias with trisomy 8, 11, or 13, monosomy 7, or deletion 5q. Leukemia 19:1224–1228.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB. 2002. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci USA 99: 4465–4470.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067.

Teixeira MR, Heim S. 2005. Multiple numerical chromosome aberrations in cancer: What are their causes and what are their consequences? Semin Cancer Biol 15:3–12.

Tsafrir D, Bacolod M, Selvanayagam Z, Tsafrir I, Shia J, Zeng Z, Liu H, Krier C, Stengel RF, Barany F, Gerald WL, Paty PB, Domany E, Notterman DA. 2006. Relationship of gene expression and chromosomal abnormalities in colorectal cancer. Cancer Res 66:2129–2137.

Tschentscher F, Husing J, Holter T, Kruse E, Dresen IG, Jockel KH, Anastassiou G, Schilling H, Bornfeld N, Horsthemke B, Lohmann DR, Zeschnigk M. 2003. Tumor classification based on gene expression profiling shows that uveal melanomas with and without monosomy 3 represent two distinct entities. Cancer Res 63:2578–2584.

Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle A, Krahe R. 2001. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. Proc Natl Acad Sci USA 98:1124–1129.

Vogelstein B, Kinzler KW. 2004. Cancer genes and the pathways they control. Nat Med 10:789–799.

Wang X, Jin DY, Ng RW, Feng H, Wong WC, Cheung AL, Tsao SW. 2004. Significance of MAD2 expression to mitotic checkpoint control in ovarian cancer cells. Cancer Res 62:1662–1668.

Yi Y, Mirosevich J, Shyr Y, Matusik R, George AL, Jr. 2005. Coupled analysis of gene expression and chromosomal location. Genomics 85:401–412.

Zimonjic D, Brooks MW, Popescu N, Weinberg RA, Hahn WC. 2001. Derivation of human tumor cells in vitro without widespread genomic instability. Cancer Res 61:8838–8844.