Research

Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data
Joseph J Crawley and Kyle A Furge

Address: Bioinformatics Program, Van Andel Research Institute, Grand Rapids, MI 49503, USA.

Correspondence: Kyle A Furge. E-mail: kyle.furge@vai.org

Published: 25 November 2002
Genome Biology 2002, 3(12):research0075.1–0075.8
The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/12/research/0075
© 2002 Crawley and Furge, licensee BioMed Central Ltd
(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: Hepatocellular carcinoma (HCC) is a leading cause of death worldwide. Frequent cytogenetic abnormalities that occur in HCC suggest that tumor-modifying genes (oncogenes or tumor suppressors) may be driving selection for amplification or deletion of these particular genetic regions. In many cases, however, the gene(s) that drive the selection are unknown. Although techniques such as comparative genomic hybridization (CGH) have traditionally been used to identify cytogenetic aberrations, it might also be possible to identify them indirectly from gene-expression studies. A technique we have called comparative genomic microarray analysis (CGMA) predicts regions of cytogenetic change by searching for regional gene-expression biases. CGMA was applied to HCC gene-expression profiles to identify regions of frequent cytogenetic change and to identify genes whose expression is misregulated within these regions.

Results: Using CGMA, 104 HCC gene-expression microarray profiles were analyzed. CGMA identified 13 regions of frequent cytogenetic change in the HCC samples. Ten of these regions have been detected in previous CGH studies (+1q, -4q, +6p, -8p, +8q, -13q, -16q, -17p, +17q, +20q). CGMA identified three additional regions that have not been previously identified by CGH (+5q, +12q, +19p). Genes located in regions of frequent cytogenetic change were examined for changed expression in the HCC samples.

Conclusions: Our results suggest that CGMA predictions using gene-expression microarray datasets are a practical alternative to CGH profiling. In addition, CGMA might be useful for identifying candidate genes within cytogenetically abnormal regions.

Background

Aneuploidy is a common feature of cancer. Genetic alterations such as amplification, deletion, translocation and rearrangement could result in either gain-of-function or loss-of-function mutations in genes that modulate aspects of cell proliferation, differentiation, motility and survival. Whereas cytogenetic profiling techniques, such as comparative genomic hybridization (CGH) [1], have been useful in finding genetic abnormalities, other experimental approaches are frequently used to identify which specific gene(s) drive selection for the genetic aberration and contribute most to tumor progression. Common gene identification techniques include determining if a candidate gene contains a sequence mutation and/or determining if the candidate gene or gene product is abnormally expressed. As mutation analysis and protein expression studies are time-consuming, increasingly...
high-throughput gene-expression profiling is being used to identify abnormally expressed genes within a region of cytogenetic change [2-6].

Recently, several groups have observed that chromosomal changes can lead to regional biases in gene-expression values both in yeast (Saccharomyces cerevisiae) and in human tumors and tumor-derived cell lines [2,3,7,8]. These studies suggest that a fraction of gene-expression values (15-25%) are regulated in concordance with gene dosage. A computational technique termed comparative genomic microarray analysis (CGMA) has previously been used to identify regions of allelic imbalance indirectly from gene-expression profiles of human tumors [8]. CGMA predicts chromosomal amplifications and deletions by organizing gene-expression data by genomic mapping location and scanning for regions that contain a statistically significant number of gene-expression values that change in the same relative direction. In this study, we apply CGMA analysis to a large hepatocellular carcinoma microarray dataset to demonstrate its validity as an alternative to CGH and to identify candidate genes in regions of frequent cytogenetic change.

Primary liver cancer in adults is the sixth most common form of cancer and the fourth leading cause of death from cancer worldwide [9,10]. Through the examination of hepatitis B virus (HBV)- and hepatitis C virus (HCV)-induced tumors, two landmark CGH studies have suggested that a subset of cytogenetic changes frequently occurs in HCC [11,12]. These include frequent gain of chromosomes 1q, 6p, 8q, 17q and 20q and frequent loss of chromosomes 1p, 4q, 6q, 8p, 13, 16 and 17p [11,12]. In particular, gain of chromosomes 1q and 8q has been associated with the early development of HCC [12], whereas loss of chromosome 4q has been linked to increased aggressiveness of established tumors [11]. To determine whether gene-expression data could be used to identify cytogenetic changes accurately, we applied CGMA to a microarray dataset of HCC tumors and compared the CGMA predictions to existing CGH data. For HCC, CGMA was able to predict nearly all chromosomal aberrations identified previously by CGH. In addition, from the gene-expression data we also identified a set of genes whose expression values change most within the regions of cytogenetic change. These genes may represent candidate genes whose expression changes drive selection for chromosomal gains or losses.

Results

CGMA predictions of cytogenetic changes

Normalized, log-transformed gene-expression data from 104 HCC gene-expression arrays [13] were obtained from the Stanford Microarray Database [14]. As CGMA analysis yields more intuitive predictions if the tumor expression data is compared to a normal tissue reference, the original HCC gene-expression data was mathematically transformed such that the pooled cell-line reference was replaced by a normal tissue reference ([3], see Materials and methods). Using this transformation, gene-expression values from the tumor sample are compared to corresponding values from non-cancerous tissue. Genomic regions that contain a disproportion-ate number of genes that change expression in the same relative direction (that is, show a gene-expression bias) may indicate an underlying chromosomal gain or loss (Figure 1a) [2,3,7,8]. Chromosomal regions that contained a gene-expression bias with at least 95% confidence (a sign test z-statistic of at least 1.96, see Materials and methods) were identified for all 104 HCC expression profiles (Figure 1b). In addition, genomic regions that contained significant gene-expression biases in at least 35% of non-replicate samples were identified (Figure 1c). A 35% threshold was chosen because in previous CGMA profiling experiments this threshold yielded the highest CGMA to CGH agreement ([8] and data not shown). CGMA predicted frequent gains for chromosome 1q (gained in 72% of tumor samples), 6p (56%), 8q (49%), 17q (46%), 20q (46%), 5q (42%), 19p (37%) and 12q (35%). Frequent chromosomal losses were predicted for chromosome 4q (lost in 66% of tumor samples), 17p (48%), 13 (39%), 16 (37%), and 8p (35%). To determine if CGMA predictions were consistent with other cytogenetic profiling studies, the CGMA data were compared with data from two of the largest CGH profiling studies (67 and 50 samples, respectively) using HCC tumors [11,12]. Of the 13 regions detected by CGMA, 10 (77%) were also implicated by CGH (Figure 1c). CGMA also detected three gained regions, chromosomes 5q, 12q and 19p, which were not implicated in the CGH analysis. CGMA failed to discover two regions of loss detected by CGH - chromosomes 1p and 6q. It is noteworthy that these particular losses were not identified in both CGH studies. These data suggest that CGMA predictions produce results very similar to traditional CGH profiling studies.

Comparison to previous HCC studies

To date there have been at least 20 reports on the application of CGH to HCC [11,12,15-32]. To determine whether the differences between the CGMA predictions and the two large CGH studies were similar to the experimental variation observed between different HCC CGH studies, predictions produced by CGMA were compared to 13 different HCC CGH profiling studies (Figure 2). CGMA produced 10 of 13 (77%) predictions that matched a consensus chromosomal aberration profile. On average, each CGH study matched the consensus profile in 78±14% of the chromosomal regions analyzed. Therefore, the variation in CGMA results was similar to the variations between independent CGH studies. These results suggest that CGMA profiling is able to predict regions of frequent chromosomal imbalance in HCC as well as CGH profiling. As only 4 of the 104 (4%) samples analyzed scored positive for HCV infection, we could not use this dataset to detect significant cytogenetic differences between HCV-infected versus HBV-infected individuals (data not shown).
CGMA predictions of multifocal tumors

Included in the set of HCC gene-expression profiles were several cases in which multiple tumor nodules were removed from the same patient. In some cases the nodules had related gene-expression profiles (patients HK63, HK64, HK66) whereas in other cases tumors from the same patient had distinct profiles (HK65, HK67, HK85) [33]. In particular, in patient HK67 the gene-expression profile from nodule HK67.1 was distinct from the expression profiles from nodules HK67.2 and HK67.3 [33]. Array CGH was used previously to determine the cytogenetic profiles of the tumor nodules from this patient [33]. Array CGH identified common cytogenetic abnormalities in patient HK67’s tumors, including loss of chromosome 19q, an unusual gain of chromosome 19q, and loss of the centromeric region of chromosome 22. However, additional cytogenetic changes were found in tumors HK67.2 and HK67.3 that were not present in HK67.1. The cytogenetic profiling data coupled with the observation that HK67.2 and HK67.3 were both smaller in size and had an increased mitotic index, suggested that HK67.1 was the primary tumor and HK67.2 and HK67.3 were divergent HK67.1 subclones. To determine whether CGMA predictions agree with this monoclonal origin hypothesis, CGMA profiles of patient HK67’s tumor nodules were isolated and organized by hierarchical clustering (Figure 3). CGMA also detected the common chromosome 19q gain and chromosome 15q loss in the HK67 tumors. CGMA did not identify a common loss on chromosome 22, however, CGMA identified other genetic aberrations (+8q, -16q and -19q) consistently found in the HK67 tumors. CGMA also identified additional aberrations (+2q, +5q, +12q) present in HK67.2 and HK67.3 that were not found in HK67.1. Taken together, the CGMA data supports the hypothesis that the HK67.2 and HK67.3 tumor nodules probably arose from HK67.1, but that additional distinct cytogenetic events had occurred in these nodules during tumor progression. In contrast, tumor nodules from patient HK64 have very similar gene-expression profiles and very similar cytogenetic profiles as predicted by CGMA, suggesting that these tumors have common origins and these nodules have not diverged significantly from the original lineage. In addition, tumor nodules from patient HK85 showed distinctive expression profiles and distinct HBV integration sites [33]. Similarly, the tumors from patient HK85 also show distinct CGMA-predicted cytogenetic profiles, reflecting the independent transforming mechanism (Figure 3).

Identification of candidate genes in regions of frequent cytogenetic change

Frequent cytogenetic abnormalities suggest that tumor-modifying genes (oncogenes or tumor suppressors) may be
driving selection for the amplification or deletion of these particular genetic regions [6,11,12,34-36]. An advantage of using CGMA profiling rather than traditional molecular genetic profiling is that access to gene-expression data is inherent in the analysis. CGMA allows cytogenetic analysis and the candidate gene approach to be performed with the same dataset. For example, the c-myc oncogene has been postulated to drive selection for frequent chromosome 8q amplification. Though c-myc is located on a region that both CGMA and CGH identify as frequently gained, c-myc’s expression is increased more than twofold in less than 6% of the samples. In fact, in 52% of the HCC tissue samples, c-myc’s expression is downregulated (Table 1). This implies that increased c-myc expression is not driving the selection for the amplification of chromosomal region 8q in these samples. In the small region of chromosome 8q presented in Table 1, two other genes (for squalene monooxygenase and pro2000) do show increased expression in a majority of HCC samples. Consistent with previous reports examining gene-expression levels in regions of cytogenetic change, expression levels for a large percentage of genes in this amplified region remain unchanged [3,5,8].

The set of genes that are consistently misregulated in regions of frequent cytogenetic change as predicted by CGMA are shown in Table 2. Platelet-derived growth factor receptor alpha is consistently downregulated in a region of frequent cytogenetic loss and this suggests that loss of a member of this receptor gene family is important in HCC progression. It has previously been reported that a transcript (PRLTS) with sequence similarity to the extracellular domain of platelet-derived growth factor receptor may also be a tumor suppressor for HCC [35]. In addition, consistently increased expression of the pituitary tumor transforming gene 1 oncogene (PTTG) is observed in these samples (Table 2).
PTTG maps to chromosome 5q, a region that was identified as frequently changed by CGMA, but not identified in the majority of CGH profiling studies. PTTP overexpression in NIH 3T3 cells induces these cells to form tumors when injected into nude mice. Overexpression of this gene may result from frequent chromosomal amplification and may participate in HCC tumor progression.

**CGMA prediction software**

To assist in identifying regions of unidirectional gene-expression bias, we have constructed a web-based program that processes two-color gene-expression data and identifies genomic regions that contain gene-expression biases. The input for this program is a simple tab-delimited gene-expression matrix file consisting of columns for the probe
Because different microarray technologies use different identifiers to describe the microarray probe, the program translates probe sequence identifiers (ids) such as GenBank accession numbers and UniGene cluster ids to Ensembl transcript ids using precompiled sequence comparisons. After data analysis, a summary table is displayed showing chromosomal regions that show significant (α ≤ 0.05) unidirectional gene-expression bias highlighted in either red or green, indicating either increased or decreased expression biases, respectively. The program can also send several output files to the user via e-mail. These files include a summary report that contains the z-statistic for each chromosomal region (positive for upregulated regions and negative for downregulated regions) and a list of genes located in regions of frequent cytogenetic change. The program is available at [37].

**Discussion**

In this study we have used gene-expression profiling data to predict cytogenetic changes that frequently occur in HCC.
Two landmark CGH analyses identified 12 different regions of frequent imbalance. However, one study found 8 regions and the other study found 11 [11,12]. Five of these 12 regions were not found in both experiments. CGMA successfully identified 10 of 12 regions previously distinguished by CGH. CGMA also detected three regions that have not been implicated by these CGH studies. On average however, 22% of genomic regions identified in a particular HCC CGH study are not constantly identified in other studies. Therefore, the three inconsistent CGMA predictions (3 of 13; 23%) are comparable to the inconsistencies between independent CGH studies for HCC.

Three additional regions were identified by CGMA that were not identified by CGH. While these CGMA-predicted regions were near the 35% cutoff for detection, they could represent other regions of allelic imbalance yet to be detected by CGH. It is also possible that biological mechanisms other than cytogenetic change could influence expression in large genomic regions and produce regional gene-expression biases. Additional molecular genetic work will be required to resolve these differences.

If CGH data are not available for a particular cancer type, but gene-expression profiling data are, then CGMA could allow rapid prediction of the cytogenetic abnormalities that frequently occur within that cancer type. Moreover, in instances where gene-expression profiling reveals previously unrecognized cancer subtypes, CGMA could determine whether cytogenetic differences are responsible for these different subtypes. In cancer types where traditional cytogenetic profiling has already been carried out, CGMA predictions could serve to confirm existing cytogenetic profiling data and be used further to examine candidate genes whose expression changes most within a region of frequent cytogenetic change. In this way CGMA can be combined with the candidate gene approach to identify genes that are directly involved in tumor progression.

Conclusions
CGMA can be used to indicate chromosomal imbalances by detecting chromosomal regions that contain a disproportionate number of gene-expression values that change in the same relative direction. This analysis provides good evidence that CGMA is a practical alternative to CGH cytogenetic profiling when gene-expression profiling data is available.

Materials and methods
Normalization and filtering
Normalized, log-transformed gene-expression data for 104 HCC samples and 76 corresponding non-cancerous liver gene-expression profiles [13] were obtained from the Stanford Microarray Database [14]. Genes that were present in at least 75% of samples (10,037 genes) were used for further analysis. In this study, both the tumor samples and normal tissue samples were compared to a pooled cell-line reference [3]. To allow comparison of tumor gene-expression values to gene-expression values from surrounding non-cancerous tissue, new gene expression ratios, tumor versus normal (T/N), were estimated. To create the new ratios, log-transformed non-cancerous tissue ratios (N/U) were subtracted from the log-transformed HCC tissue ratios (T/U) for each gene such that \[ \log_2(T/N) = \log_2(T/U) - \log_2(N/U). \] If an HCC sample did not have a corresponding non-cancerous sample, the global mean of the non-cancerous tissue gene-expression ratios was used.

**CGMA analysis**
To identify regional gene-expression biases, gene-expression values that map within a given chromosomal arm were collected and a sign test for a one-sample mean/median was used to determine whether a significant upward or downward bias was present in the expression values. An exception was made for chromosomes 13-16, 21 and 22. These chromosomes are more telocentric and therefore only their q-arms were tested for expression biases. Sequence comparisons were used to map microarray probe sequences (the sequences that are placed on the microarray) to predicted Ensembl transcripts (Ensembl version 6.28) [8]. Included in the Ensembl transcript annotations are chromosomal mapping locations at base-pair resolution. Redundancy introduced by replicate probes on the array and/or multiple probes mapping to the same gene were eliminated by averaging expression values that map to identical transcripts. Of the filtered set of 10,037 genes, 6,274 genes (63%) were unique and had associated genomic mapping information.

A sign test for the one-sample mean/median was used to determine whether a significant number of genes that map to a given chromosomal region change in a unidirectional manner. The algorithm scores a gene as up (+) or down (-) regulated if the magnitude of the expression value change is at least 1.8-fold. The sign test computes the probability, in the form of a z-statistic, of finding x upregulated genes out of n genes that change in a given genomic region. For simplicity, the z-statistic is computed using the normal approximation to the binomial distribution such that \[ z = \frac{2x - n}{\sqrt{n}}. \] Genomic regions that contained less than 15 changed gene-expression values were excluded from further analysis. On average, 160 gene-expression values were located to each genomic region. The sign test z-statistic can be converted to a significance value (\( \alpha \)) based on the two-tailed z-statistic (\( z_{\alpha/2} \)) critical values. For example, if \( z = 1.96 \), then \( \alpha = 0.05 \); if \( z = 2.58 \) then \( \alpha = 0.01 \), and so on.

**Acknowledgements**
Funding was through the generosity of the Van Andel Research Institute and Michigan Center for Biological Information (MCBI). We specially thank Ramsi Haddad for helpful discussion in the preparation of this manuscript.
References

1. Kallioniemi A, Kallioniemi OP, Sudar D, Rutzovit D, Gray JW, Waldman F, Pinkel D: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992, 258:818-821.

2. Virtanen K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri M, Bloomfield CD, de La Chapelle A, Krahe R: Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated 8 and normal cytogenetics. Proc Natl Acad Sci USA 2001, 98:124-1129.

3. Phillips JL, Hayward SW, Wang Y, Vasselli J, Pavlovich C, Padilla-Nash P, Parkin DM, Pisani P, Ferlay J: The incidence of eighteen major cancers in 1985. Int J Cancer 1993, 53:110-116.

4. Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burdach J, Dow S, Ward TR, Kidd MJ, et al.: Insight into hepatocellular carcinogenesis at transcriptional level by comparing expression profiles of hepatocellular carcinoma with those of corresponding non-cancerous liver. Nat Genet 2001, 25:333-337.

5. Platerz P, Upender MB, Wilson K, Willis J, Lutterbaugh J, Nosrati A, Wilson JK, Mack D, Ried T, Markowitz S: Evidence for the presence of two tumour-suppressor genes at 17q23 ameliorating male breast cancer. Proc Natl Acad Sci USA 1998, 95:11-17.

6. Haddad R, Furge KA, Miller J, Schoumans J, Haab B, Teh B, Barr L, Webb C, et al.: Genomic profiling and cDNA microarray analysis of human colon adenocarcinoma and associated peripheral metastasis reveals consistent cytogenetic and transcriptional aberrations associated with progression of multiple metastases. Appl Genomics Proteomics 2002, 1:123-134.

7. Parkin DM, Pisani P, Ferlay J: Estimates of the worldwide mortality from eighteen major cancers in 1985. Implications for prevention and projections of future burden. Int J Cancer 1993, 55:891-903.

8. Novel recurrent genetic imbalances in human hepatocellular carcinoma: correlation with fluorescence in situ hybridization in paraffin-embedded tissue. Mol Diagn 2001, 6:273-37.

9. Takeo S, Araki H, Kasano N, Harada T, Furuya T, Kawauso S, Oga A, Hiranro T, Yoshihida T, Okita K, Sasaki K: Examination of oncogene amplification by genomic microarray in hepatocellular carcinomas: comparison with comparative genomic hybridization analysis. Cancer Genet Cytogenet 2001, 130:127-132.

10. Kusano N, Okita K, Shirahashi H, Harada T, Sasaki K, Oga A, Kawausco S, Furuya T, Sasaki K: Chromosomal imbalances detected by comparative genomic hybridization are associated with outcome of patients with hepatocellular carcinoma. Cancer 2002, 94:746-751.

11. Yasui K, Ari S, Zhao C, Imoto I, Ueda M, Nagai H, Emi M, Inazawa J: TDP31, CUL4A, and CDC616 identified as targets for amplification at 19q13 in hepatocellular carcinomas. Hepatology 2002, 35:1476-1484.

12. Kuroki T, Fujiwara Y, Nakamori S, Imaoka S, Kanematsu T, Nakamura Y: Hypomethylation of chromosome 1 heterochromatin DNA correlates with q-arm copy gain in human hepatocellular carcinomas. Am J Pathol 2001, 159:465-471.

13. Fujwara Y, Osata H, Kuroki T, Koyama K, Tsuichita E, Monden M, Nakamura Y: Isolation of a candidate tumour suppressor gene on chromosome 8q22.2 that is homologous to an extracellular domain of the PDGF receptor beta gene. Oncogene 1995, 10:891-895.

14. Kuroki T, Fujwara Y, Nakamori S, Imaoka S, Kanematsu T, Nakamura Y: Evidence for the presence of two tumour-suppressor genes for hepatocellular carcinoma on chromosome 13q. Br J Cancer 1995, 72:383-385.

15. CGMA - comparative analysis of microarray data [http://www.vai.org/vari/downloads/development/clam-0.2.pl]

16. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998, 95:14863-14868.