Genome-wide Identification of Chromosomal Regions of Increased Tumor Expression by Transcriptome Analysis

Yan Zhou,* Shiu-Ming Luoh, Yan Zhang, Colin Watanabe, Thomas D. Wu, Michael Ostland, William I. Wood, and Zemin Zhang


ABSTRACT

Genes up-regulated in tumor cells provide attractive anticancer therapeutic targets. Although the general underlying mechanism for the increased expression in tumors is unknown, tumor-specific up-regulation of some genes can be attributed to aberrant DNA amplification, a phenomenon common to many tumors. Using a computational method, we constructed a general transcriptome map with the human genomic sequences and expressed sequence tags in the public database. The transcriptome map revealed nonrandom chromosomal regions (termed region of increased tumor expression) where clusters of genes exhibited increased expression in the 10 tumor tissue types tested. These genomic regions often correspond to experimentally verified tumor amplicons. Our large-scale transcriptome analysis led to identification of many additional chromosomal regions with increased tumor expression, regions that represent potential tumor amplicons.

INTRODUCTION

Allelic imbalance and altered gene expression patterns are two major characteristics of cancer cells. Genes with elevated expression in tumor cells provide potential anticancer therapeutic targets, as is the case for the ERBB2 gene (1–3), whereas aberrant DNA amplification in tumors provides markers for molecular diagnosis. Although the general underlying mechanism for gene up-regulation remains unknown, parallel microarray measurement of mRNA levels and DNA copy number has shown that aberrant DNA amplification contributes to increased gene expression in cancer cells (4). Identification of DNA amplicons in tumors typically requires laborious approaches such as FISH (5), CGH (6), representational difference analysis (7, 8), and restriction landmark genome scanning (9). Recently, DNA microarray-based CGH has been adapted for genome-wide detection of chromosomal imbalance (10–12); however, the resolution of the results depends on array densities, and the technology’s requirement to perform a large number of microarray experiments using heterogeneous cancer samples can limit the large-scale identification of tumor amplicons.

The availability of the human genome sequences and >4,000,000 ESTs (13) in the public database has made it possible to not only construct a general transcriptome map but also to perform large-scale comparisons of differential gene expression between normal and tumor tissues. In this study, we constructed human transcriptome maps in 10 different tissue types: brain; breast; colon; kidney; liver; lung; pancreas; prostate; stomach; and testis. In each tissue, we discovered that sets of genes up-regulated in tumors cluster on the chromosomes. Permutation analyses demonstrated that this clustering was nonrandom. Furthermore, 46–74% of these genomic regions with increased tumor expression corresponded to amplicons identified by CGH analysis.

MATERIALS AND METHODS

Human EST and mRNA Collection. The human EST collection consists of all sequences in the EST division of GenBank (14). Quality information contained in the GenBank file was used to trim poor sequences. On the basis of tissue and histology data for library information from National Cancer Institute and Cancer Genome Anatomy Project, each EST was assigned a tissue source value and a disease value. ESTs with unknown or ambiguous tissue sources and ESTs that failed to fall into either the normal or cancer categories were removed. ESTs from libraries labeled as normalized or subtracted were also discarded as they might obscure EST abundance calculation. Although a total of 5487 usable libraries were collected, ~85% of the ESTs were covered by 820 libraries, with an average library size of 3194.

The human mRNA collection, downloaded from NCBI, contains of mRNA sequences mapped onto the assembled human genome from NCBI (release 31, January 2003), as specified in each chromosome’s GenBank file. The chromosomal location of each mRNA was determined from the contig order information plus position information. Pseudogenes, mRNAs labeled “similar to Alu Subfamily,” and mRNAs with anomalies in the location specification were eliminated. Among 20,630 genes with EST coverage, 20,182 nonoverlapping genes were selected for expression profiling analysis.

EST-based Expression Level Calculation and Statistical Comparison. The sequence alignment program Basic Local Alignment Search Tool (15) was used to identify all EST sequences with matches of at least 60 bp with >98% identity to a gene of interest. Matched ESTs were grouped by their source libraries, which were additionally consolidated by library tissue types. For the normal and tumor groups in each tissue category, a DEL value was calculated as the total number of matching EST clones divided by the sum of library sizes and multiplied by 1,000,000.

The Z test was applied to determine whether DEL in tumor samples was statistically higher than DEL in normal samples because this was similar to the comparison of the proportions from two independent populations. For a given gene, the common relative abundance (PHAT) was computed in all libraries by taking the sum of clone counts for the gene over all libraries and dividing by the total number of clones over all libraries. The relative abundance for the gene in tumor (PHAT_tumor) and in normal (PHAT_normal) was also calculated. The test statistic Z-score was calculated as follows:

\[ Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{(1 \times \hat{P}_1 \times (1 - \hat{P}_1) + (1 \times \hat{P}_2 \times (1 - \hat{P}_2)))}} \]

where \( N_{tumor} \) and \( N_{normal} \) are the total numbers of EST clones derived from tumor and normal tissues.

Computational Detection of RITEs. To identify genomic regions where clusters of genes are up-regulated in tumor, we devised a TMZ and an SJ. TMZ was calculated using a 500-kb window moving along the chromosome at 100-kb intervals and defined as:


Received 11/26/02; revised 6/12/03; accepted 7/8/03.

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.

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

2 The first two authors contributed equally to this work.

3 To whom requests for reprints should be addressed, at 1 DNA Way, MS93, South San Francisco, CA 94080. Phone: (650) 225-4293; Fax: (650) 225-5389; E-mail: zemin@ gene.com.

4 The abbreviations used are: FISH, fluorescence in situ hybridization; CGH, comparative genomic hybridization; EST, expressed sequence tag; NCBI, National Center for Biotechnology Information; DEL, digital expression level; TMZ, trimmed mean of Z-score; SI, sensing index; RITE, region of increased tumor expression; mb, million bp.
where \( Z_i \) is the Z-score of the \( i \)th gene in a 500-kb window, max and min are the maximum and minimum Z-scores in the window, and \( n \) is the number of genes in the window. The SI was empirically devised to give weight to regions where consecutive genomic windows showed high TMZ values. It was calculated as \( SI = \frac{(m - g) \times \sqrt{n \times L}}{n - 2} \) where \( g \) is the mean of TMZ for all windows in the genome, \( m \) is the TMZ mean of consecutive windows where TMZ is greater than \( g \), \( n \) is the number of consecutive windows where TMZ is greater than \( g \), and \( L \) is the length of the window in mb. If the trimmed mean of the window was less than the genomic mean, SI was simply calculated as \( m - g \.

Using this method, all of the consecutive windows within a genomic region had the same SI value when their TMZ values were greater than \( g \). When a RITE was identified, we refined its boundaries by trimming off genes on each side where Z-scores were \( >4g \). Because of this trimming process, negative-sensing indices often appeared adjacent to positive indices as genes with negative Z-scores were removed from the edges.

**RESULTS AND DISCUSSION**

**Identification of Chromosome Regions of Increased Tumor Expression.** From the GenBank and NCBI databases, we retrieved 1.45 million ESTs from normal and 1.91 million from tumor tissues (Table 1). To quantify the tissue expression level for each gene, the gene’s normalized EST abundance was computed and expressed as the digital expression level (DEL = EST abundance/million ESTs). A linear correlation was found between DEL and normalized mRNA level measured by real-time PCR analysis (unpublished data).

The DEL values for each gene were plotted along the chromosome according to its genomic coordinates, producing transcriptome maps for both normal and tumor tissues in each tissue type (Fig. 1). In specific regions, genes with proximal locations appeared to be co-up-regulated in tumor tissues. One such example is a 2.6-mb region corresponding to 17q21 where 38 genes are expressed at a higher level in breast tumor samples as compared with normal (Fig. 1). This region includes the \( \text{ERBB2} \) gene and is known to be amplified in \( \approx30\% \) of breast cancer patients (16, 17). Additionally, this region roughly corresponds to breast tumor amplicons in size (18). As we identified other similar tumor-related regions, we named these RITEs.

To determine whether RITEs merely represent isolated cases or common features on chromosome, a computational approach was designed to identify all RITEs in 10 organs with abundant EST coverage (Table 1): brain; breast; colon; kidney; liver; lung; pancreas; prostate; stomach; and testis. Using each gene’s EST distribution data in normal and tumor samples, a two-proportion Z test was performed, measuring the significance of the gene’s up-regulation in tumor. To systematically identify RITEs and to reduce noise caused by individual genes, the human genome was scanned in a 500-kb window with a 100-kb offset and the TMZs for the genes in each window was calculated. An SI was then empirically devised that captures consecutive windows with a TMZ above the genome-wide mean. The SI distribution along the genome is not uniform, and the regions with high SI could represent RITEs. As an example, chromosome 17 for breast data demonstrates this nonuniform distribution (Fig. 2, also see Supplementary Figs. 1–20). Guided by known examples like the \( \text{ERBB2} \) region, we empirically set the SI cutoff as 2.0. Our algorithm revealed 75 putative RITEs for breast tumor. Manual inspection of these putative RITEs confirmed gene clusters with higher expression in tumors. Similarly, we also identified 36–110 RITEs in each of the 9 other organs (Fig. 3). On average, a RITE region spans \( \approx1.3 \) mb genomic region, covering 21 genes.

**Clustering of Tumor Up-Regulated Genes Is Not Random.** To address the concern that the detected RITEs simply result from individual tumor-specific genes that happen to reside close to one another along chromosomes by chance, a series of permutation analyses were performed where each gene’s chromosomal coordinates were randomly shuffled. Ten thousand permuted breast data sets were...

**Table 1 Number of usable ESTs in public database**

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Normal</th>
<th>Tumor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain</td>
<td>231,788</td>
<td>144,183</td>
</tr>
<tr>
<td>Breast</td>
<td>64,056</td>
<td>116,821</td>
</tr>
<tr>
<td>Colon</td>
<td>22,556</td>
<td>134,081</td>
</tr>
<tr>
<td>Kidney</td>
<td>15,973</td>
<td>55,896</td>
</tr>
<tr>
<td>Liver</td>
<td>45,644</td>
<td>48,215</td>
</tr>
<tr>
<td>Lung</td>
<td>56,620</td>
<td>119,143</td>
</tr>
<tr>
<td>Pancreas</td>
<td>46,140</td>
<td>66,025</td>
</tr>
<tr>
<td>Prostate</td>
<td>40,831</td>
<td>61,041</td>
</tr>
<tr>
<td>Stomach</td>
<td>24,311</td>
<td>118,025</td>
</tr>
<tr>
<td>Testis</td>
<td>35,279</td>
<td>42,896</td>
</tr>
<tr>
<td>Others</td>
<td>864,693</td>
<td>1,001,577</td>
</tr>
<tr>
<td>Total</td>
<td>1,447,891</td>
<td>1,907,903</td>
</tr>
</tbody>
</table>

\(^a\) Number of EST as of February 2003.
analyzed. Under permuted conditions, the observed number of RITEs was 2.1 ± 1.5, a dramatic reduction from the 77 RITEs detected under the native condition. Similar analyses were also performed on the 9 other tissues. In all cases, the observed RITEs were almost completely abolished, with the average number of RITEs at least 30-fold lower under permutation conditions (Fig. 3).

Relationship between RITEs and Tumor Amplicons. The presence of these RITEs can be attributed to either genomic alteration or transcriptional control. To examine the relationship between RITEs and known tumor amplicons, we compared the cytogenetic marker for each RITE with known CGH data. In brain, breast, liver and lung tissues where CGH data are most abundant, 46–74% of RITEs correspond to amplicons identified by CGH analysis (Refs. 19, 20; Table 2). This is consistent with the finding by microarray analyses that 62% of high-level DNA amplifications in breast samples are associated with at least moderately elevated mRNA levels (4). Conversely, 55–65% of amplicons in the above 4 tissues determined by CGH analysis (19, 20) were identified by RITE analyses. As this comparison was done in the context of 320 all available NCBI cytogenetic markers at the resolution given in Table 2, Fisher’s exact test showed significant correlation (P < 0.001) between the cytogenetic markers of known amplicons and RITEs for each of the four tissue organs, suggesting that these RITEs could result from genomic DNA amplification in tumor.

Despite the observed correlation, there are several important differences between RITEs and tumor amplicons. Most importantly, because EST clones come from multiple individual tissue libraries, RITEs are representations of the cancer patient population rather than individual patients. CGH analyses, on the other hand, are usually performed on individual patients and therefore reveal DNA copy numbers in those individual samples, and the results could be more specific and sensitive. Although we could in theory perform RITE analysis using EST data from a single library or from a subset of libraries from a specific tumor type (e.g., oligodendroglioma), our statistical analysis will suffer because of the much smaller number of ESTs. In addition, because not all DNA amplifications result in detectable elevation of mRNA level, expression-based analysis may not always detect tumor amplicons even with sufficient EST data.

An alternative mechanism for the presence of RITEs could be coordinated gene expression at the transcriptional level, as evidence for large domains of similarly expressed genes in eukaryotic genomes exists (21, 22). A preliminary analysis of the upstream DNA sequences failed to identify promoter elements common to all genes in the RITE regions (data not shown); however, we cannot rule out the possibility that these regions are organized in some form of chromosomal scaffold that enables coordinated expression in tumors.

Our transcriptome analyses revealed a few other interesting observations. Notably, although many of the RITE regions are recurrently found in multiple tissues, some RITEs appear to be tissue specific. For instance, 11q13 is a RITE in 8 of the 10 tissues tested (Table 2, Supplementary Table 1, and Supplementary Fig. 11), and it is also one of the most frequently amplified regions in human cancer (23). On the other hand, 10q11 appears to be specific to brain tumors (Supplementary Fig. 10). Furthermore, RITEs are more frequently observed on certain chromosomes. As an example, chromosome 19 appears to be more active than chromosome 13 in tumors (Supplementary Figs. 13

Table 2. List of identified RITEs in brain, breast, liver, and lung

<table>
<thead>
<tr>
<th>CHR</th>
<th>Brain</th>
<th>Breast</th>
<th>Liver</th>
<th>Lung</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1p34, 1p36, 1q21, 1q31</td>
<td>1p34, 1p33, 1q21, 1q41, 1q42</td>
<td>1p34, 1q21, 1q42</td>
<td>1p33, 1p34, 1p36, 1q21, 1q42</td>
</tr>
<tr>
<td>2</td>
<td>2p23, 2p21, 2q35, 2q36, 2q37</td>
<td>2p21, 2p24, 2q11, 2q37</td>
<td>2p11, 2p25, 2q21, 2q35, 2q36</td>
<td>2p12, 2q11, 2q13, 2q21, 2q37</td>
</tr>
<tr>
<td>3</td>
<td>3p21, 3p22, 3q21, 3q28</td>
<td>3q25, 3q26, 3q29</td>
<td>3q21</td>
<td>3q21, 3q23, 3q27</td>
</tr>
<tr>
<td>4</td>
<td>4q13, 4q21, 4q34</td>
<td>4p16</td>
<td>4p16</td>
<td>5p15, 5q35</td>
</tr>
<tr>
<td>5</td>
<td>5p15, 5q35</td>
<td>6p21</td>
<td>6p21</td>
<td>6p21</td>
</tr>
<tr>
<td>6</td>
<td>6p21, 6q24, 6q16</td>
<td>6p21</td>
<td>6p21</td>
<td>6p21</td>
</tr>
<tr>
<td>7</td>
<td>7p22, 7p12, 7p13, 7q22, 7q36</td>
<td>7q11, 7q22, 7q32, 7q34</td>
<td>7p13, 7p12</td>
<td>7p22, 7p12, 7q11, 7q22</td>
</tr>
<tr>
<td>8</td>
<td>8p21, 8q24</td>
<td>8q24</td>
<td>8q24</td>
<td>8q24</td>
</tr>
<tr>
<td>9</td>
<td>9p21, 9q22, 9q34</td>
<td>9q32, 9q33</td>
<td>9p13, 9q34</td>
<td>9q22, 9q34</td>
</tr>
<tr>
<td>10</td>
<td>10q11, 10q22</td>
<td>10q15, 10q26</td>
<td>10q22, 10q24</td>
<td>10q15</td>
</tr>
<tr>
<td>11</td>
<td>11p15, 11q12, 11q13</td>
<td>11q12, 11q13</td>
<td>11q12, 11q13</td>
<td>11q12, 11q13</td>
</tr>
<tr>
<td>12</td>
<td>12p13, 12q13</td>
<td>12p11, 12q13, 12q24</td>
<td>12p13, 12q13</td>
<td>12p13, 12q13, 12q24</td>
</tr>
<tr>
<td>13</td>
<td>13q14</td>
<td>13q14</td>
<td>13q14</td>
<td>13q14</td>
</tr>
<tr>
<td>14</td>
<td>14q32</td>
<td>14q32</td>
<td>14q32</td>
<td>14q32</td>
</tr>
<tr>
<td>15</td>
<td>15q14, 15q15</td>
<td>15q14, 15q22, 15q23, 15q24</td>
<td>15q22, 15q24, 15q25, 15q26</td>
<td>15q22, 15q24, 15q25, 15q26</td>
</tr>
<tr>
<td>16</td>
<td>16p13, 16p12, 16p11, 16q12, 16q24</td>
<td>16p13, 16p12, 16p11, 16q24</td>
<td>16p13, 16p12, 16p11, 16q22</td>
<td>16p13, 16p12, 16p11, 16q12, 16q22</td>
</tr>
<tr>
<td>17</td>
<td>17p21, 17q23, 17q25</td>
<td>17p13, 17p11, 17q21, 17q23, 17q25</td>
<td>17p13, 17q12, 17q21, 17q25</td>
<td>17p11, 17q11, 17q21, 17q25</td>
</tr>
<tr>
<td>18</td>
<td>19p13, 19p12, 19q13</td>
<td>19p13, 19p12, 19p11, 19q13</td>
<td>19p13, 19q13</td>
<td>19p13, 19p12, 19q13</td>
</tr>
<tr>
<td>20</td>
<td>20p13, 20q11, 20q13</td>
<td>20p11, 20q13</td>
<td>20p13, 20q11, 20q13</td>
<td>20p13, 20q11, 20q13</td>
</tr>
<tr>
<td>21</td>
<td>21p22, 21q22</td>
<td>21q22</td>
<td>21q22</td>
<td>21q22</td>
</tr>
<tr>
<td>22</td>
<td>22q12, 22q13</td>
<td>22q12, 22q13</td>
<td>22q13</td>
<td>22q12</td>
</tr>
<tr>
<td>X</td>
<td>Xp11, Xq11, Xq12, Xq13, Xq28</td>
<td>Xp22, Xq12, Xq13, Xq22, Xq28</td>
<td>Xp11, Xq28</td>
<td>Xp11, Xq28</td>
</tr>
</tbody>
</table>

* Cytogenetic markers for each RITE were based on the ISCN800_abc lookup file from NCBI. The markers are shown with the above resolution for easy comparison with other studies. RITEs matched to the same markers are only shown once. RITEs marked in bold correspond to tumor amplicon regions previously identified by CGH analyses.
and 18). It is unclear whether this is caused by different gene densities or by unbalanced distributions of oncogenes along different chromosomes. Moreover, based on detailed analyses of a few recurrent RITEs, oncogenes are usually present in the clusters of tumor up-regulated genes. ERBB2, c-myc, and a recently found oncogene TAOS1 (24), for examples, are present in 17q21, 8q24, and 11q13, respectively. Identification of RITEs could therefore help in finding additional oncogenes. Preliminary experimental analysis of the RITE region at 9q13 confirmed that this region not only was amplified at the DNA level in a number of pancreatic cancer cell lines but also contained a putative oncogene with protein kinase activities (D. Davis, unpublished data).

In summary, we observed a large number of nonrandom clusters of genes along the genome with increased expression in tumors. Although experimental validation is needed, it is likely many of RITEs are the consequence of allelic imbalance in tumor cells. Although microarray-based CGH analysis is effective in finding tumor-specific amplicons in individual patient samples, the EST-based transcriptional analysis has the advantage of rapidly identifying potential amplicons present in a large number of cancer types and patients. The availability of these RITE regions would be helpful in the study of mechanisms of aberrant DNA amplification in tumors and in the identification of novel target genes for therapeutic intervention.

ACKNOWLEDGMENTS

We thank David Davis (Department of Molecular Biology, Genentech Inc.) for sharing the preliminary data for analysis of the 9q13 RITE region, and Allison Waugh for critical comments.

REFERENCES

Genome-wide Identification of Chromosomal Regions of Increased Tumor Expression by Transcriptome Analysis

Yan Zhou, Shiuh-Ming Luoh, Yan Zhang, et al.


Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/63/18/5781

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2004/05/19/63.18.5781.DC1

Cited articles
This article cites 22 articles, 9 of which you can access for free at:
http://cancerres.aacrjournals.org/content/63/18/5781.full.html#ref-list-1

Citing articles
This article has been cited by 10 HighWire-hosted articles. Access the articles at:
/content/63/18/5781.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.