Genomic shotgun array: a procedure linking large-scale DNA sequencing with regional transcript mapping

Ling-Hui Li1, Jian-Chiuan Li1, Yung-Feng Lin1, Chung-Yen Lin2, Chung-Yung Chen1 and Shih-Feng Tsai1,3,*

1Division of Molecular and Genomic Medicine, National Health Research Institutes, and 2Division of Biostatistics and Bioinformatics, National Health Research Institutes, 128, Yen-Chiu-Yuan Road, Sec. 2, Taipei 115, Taiwan and 3Institute of Genetics and Genome Research Center, National Yang-Ming University, 155, Li-Non Street, Sec. 2, Taipei 112, Taiwan

Received October 18, 2003; Revised and Accepted January 8, 2004

ABSTRACT

To facilitate transcript mapping and to investigate alterations in genomic structure and gene expression in a defined genomic target, we developed a novel microarray-based method to detect transcriptional activity of the human chromosome 4q22-24 region. Loss of heterozygosity of human 4q22-24 is frequently observed in hepatocellular carcinoma (HCC). One hundred and eighteen well-characterized genes have been identified from this region. We took previously sequenced shotgun subclones as templates to amplify overlapping sequences for the genomic segment and constructed a chromosome-region-specific microarray. Using genomic DNA fragments as probes, we detected transcriptional activity from within this region among five different tissues. The hybridization results indicate that there are new transcripts that have not yet been identified by other methods. The existence of new transcripts encoded by genes in this region was confirmed by PCR cloning or cDNA library screening. The procedure reported here allows coupling of shotgun sequencing with transcript mapping and, potentially, detailed analysis of gene expression and chromosomal copy of the genomic sequence for the putative HCC tumor suppressor gene(s) in the 4q candidate region.

INTRODUCTION

By now the human genome has been completely sequenced (1). The issues on how many genes are encoded by the human genome and what are the functions of these genes have been and continue to be subjects of intensive investigation. Many bioinformatics-based gene prediction methods rely on identification of splicing signals, prediction of transcriptional factor binding sites, prediction of open reading frames, alignment of cDNA sequences, protein homology search, and sequence comparison to existing genes of human or model organisms (2–9). Gene prediction information has proven useful for experimental cloning of unidentified human genes (10–12). Although the accuracy of information-based gene prediction at the nucleotide level, on average, is ~80%, at the exon level is only 45–70% (13,14). As alignment to a known protein or cDNA/EST sequence can significantly improve the performance of the gene prediction procedure, high-throughput experimental protocol to identify expressed genes is a key point in this field. Additionally, all computation-based predicted genes require validation by complementary experimental methods. Several experimental protocols have been designed to discover new transcripts in a high-throughput manner. EST clone sequencing (15) and SAGE (16) are two well-established methods to sample gene transcripts in different tissues at different stages (17). RT–PCR is frequently used to validate the existence of predicted exons (10,12). Microarray is another powerful way to detect the existence and expression of predicted genes (18). Recently, Gingeras and colleagues applied oligo-nucleotide array sets to identify transcriptional activity in chromosomes 21 and 22 (19). They found that, in comparison with what has been predicted computationally, there were many more transcripts identified from their study. Moreover, their findings suggested that gene homologs, low copy number genes or non-coding RNAs were hardly predicted or annotated by the computational methods. Similarly, Cox and colleagues used a chromosomal region expression array (CREA) that covered a 20 Mb region on human chromosome 18q to identify genes regulating HDL cholesterol (20). In that study, they identified two expressed gene fragments whose aligned genomic DNA sequences did not contain any known or predicted gene. Together, these studies support the idea that both experimental approaches and computation-based gene prediction should be applied to comprehensively describe genes in the human genome.

In our laboratory, we have focused on genes transcribed in the human chromosome 4q22-24 region. Human chromosome 4q contains critical regions involved in the tumorigenesis of hepatocellular carcinoma (HCC), findings based on the results...
of comparative genomic hybridization (CGH) and loss of heterozygosity (LOH) studies (21–25). Approximately 30–70% of HCC patients showed LOH of bands 21–25 on chromosome 4q (26–28). Repeatedly, a liver-specific tumor suppressor gene(s) located in this region has been proposed, but so far the culprit gene has not been identified (29). In a previous study, we sequenced more than 12 Mb of BAC clones from 4q22-24, aiming at establishing a complete gene list for the HCC tumor suppressor gene candidate region. Parallel to computational annotation of assembled sequences, we developed a novel experimental approach, genomic shotgun array (GSA), to detect transcriptional activity and identify new transcripts from this region. This method utilizes previously constructed BAC-derived shotgun libraries and a well-covered sequence database for constructing a chromosome-region-specific microarray. Analysis of hybridization results indicated that known genes, including the newly identified ones, were reliably detected by this procedure. The majority of hybridization-positive subclones were either identical or similar to published known genes and EST clones. Moreover, we have identified 166 hybridization-positive subclones that showed no similarity to known genes or EST clones. Two new transcripts were new identified by this GSA method. The established GSA protocol is suitable for transcript mapping and analysis in a chromosomal segment, and the application of GSA to genomic research is also discussed.

MATERIALS AND METHODS

Preparation of probes

All probes (DNA fragments that are fabricated on supported matrix) were generated by PCR amplification using M13–20 forward primer and amine-linked M13 reverse primer. After denaturing reaction at 94°C for 5 min, 40 thermal cycles were carried out at 94°C for 1 min, 55°C for 45 s and 72°C for 4 min. PCR products were purified using a 96-well PCR purification kit (Millipore), and eluted in distilled water. Probes were prepared in 150 mM NaH2PO4 (pH 8.5) solution at a final concentration of 100–150 ng/μl and subjected to arraying on 3D-Link slides (SurModics). Array fabrication was carried out using Cartecian 5200. All post-array treatment was performed according to SurModics’ instructions.

Preparation of labeled targets

Labeled targets (DNA fragments that are used in hybridization solution) were generated by PCR amplification of five cDNA libraries including lung, liver, fetal liver, bone marrow and colon (Clontech). Around 105 p.f.u. from each cDNA library were used in PCR labeling reaction. PCR reaction was done at 94°C for 1 min, 52°C for 45 s and 72°C for 3 min for a total of 40 cycles. A final concentration of 200 nM of d(A,C,G)TP, 40 nM of dTTP and 40 nM of Cy3-dUTP was used in the PCR labeling reaction. Labeled PCR product was purified with BioSpin 6 column (Bio-Rad) and concentrated to 12 μl with Microcon YM30 (Amicon).

Hybridization

Double-stranded DNA on the GSA slides was denatured in 70% formamide/2× SSC at 70°C for 5 min. The slides were dehydrated in 60, 80 and 100% ethanol, sequentially. After air drying, the slides were incubated in 50 mM ethanolamine, 0.1 M Tris (pH 9.0), 0.1% SDS for 15 min at 50°C, and then in 5× SSC, 5× Denhardt’s solution, 0.1% SDS, 0.1 mg/ml denatured salmon sperm DNA for 20 min at 50°C. Hybridization was carried out using 1/10 of labeled targets in hybridization solution of 5× SSC, 0.1% SDS, 0.1 mg/ml salmon sperm DNA and 50% formamide overnight at 65°C. Slides were washed twice in 2× SSC, 0.1% SDS for 5 min at 65°C, then once in 0.2× SSC for 5 min at 60°C, and finally in 0.1× SSC for 1 min at room temperature. After drying, the arrays were scanned using a GenePix 4000A scanner (Axon).

Statistical method for identification of hybridization-positive subclones

Notation of variables. Gene, \( \chi_i, i = 1, \ldots, I; \) negative control, \( \chi_j, j = 1, \ldots, J; \) positive control, \( \chi_k, k = 1, \ldots, K. \)

Normalization between slides: rescale the values for comparison between slides.

For each gene:

\[
\chi_i^{AN} = \frac{\bar{\chi}_i - \bar{\chi}_{NC}}{\bar{\chi}_{PC} - \bar{\chi}_{NC}},
\]

\( \chi_i^{AN} \), after normalization; \( \bar{\chi}_{NC} \), negative control; \( \bar{\chi}_{PC} \), positive control.

Where

\[
\bar{\chi}_{NC} = \frac{\sum_{j=1}^{J} \chi_j}{J} \quad \text{and} \quad \bar{\chi}_{PC} = \frac{\sum_{k=1}^{K} \chi_k}{K}.
\]

Definition of hybridization-positive subclones. For each gene after normalization between slides, if

\[
\chi_i^{AN} \geq \bar{\chi}_{PC}^{AN} - Z_{\alpha} \times SD_{PC}^{AN},
\]

we can define the \( \chi_i \) as belonging to the group of hybridization-positive subclones.

Where

\[
\bar{\chi}_{PC}^{AN} = \frac{\sum_{k=1}^{K} \chi_k^{AN}}{K}, \quad \text{SD}_{PC}^{AN} = \sqrt{\frac{\sum_{k=1}^{K} (\chi_k^{AN} - \bar{\chi}_{PC}^{AN})^2}{(K - 1)}}.
\]

And for \( \alpha = 0.05, Z_{\alpha} = 1.645 \)

For each cDNA sample, subclones identified four times out of six independent experiments as hybridization-positive were selected for further alignment and BLAST search.

Data analysis and BLAST search

Sequences of hybridization-positive subclones were retrieved from a previously established database. Alignment of these subclones to their corresponding BAC clones was done with the PhredPhrap program. BLAST search of nucleotides, ESTs and proteins was done utilizing the NCBI database (build 30). Chromosome location of hybridization-positive subclones and ESTs was assigned using the UCSC Genome Browser database (hg10, December 2001 and hg12, June 2002). The
CDS prediction was performed using GETORF in the EMBOSS package. To discover putative exons within interesting hybridization-positive subclones, we performed BLASTN on human EST databases from GenBank, EMBL and DDBJ, and BLASTX on nr, all non-redundant GenBank CDS translations, RefSeq Proteins, PDB, SwissProt, PIR and PRF. The executive programs and databases were downloaded from NCBI. Microsoft Access was used to manage the results from BLAST hits and putative CDS sequences.

**PCR cloning**

All primers used in the PCR cloning procedure were designed using the Prime3 program. Primers, 5'-CCAAACAATA-AACTTCAAGC-3' and 5'-AGGTCCTGCTTCATAATG-3' were used for cloning CDS5/CDS6 of DN04a01 subclone in a liver cDNA library and first stand cDNA from a liver tissue sample. After hot start for 10 min at 94°C, PCR reaction was carried out for 30 s at 94°C for denaturization, 30 s of annealing at 55.7°C, and 30 s at 72°C for extension, for a total of 50 cycles. Primers, 5'-GAATGCTCCCAGAGGCCA-3' and 5'-AGGCACCTTGTGGAGGCGA-3' were used for cloning the transcript inside the NF-kB1 gene using first strand cDNA of HepG2 as template. PCR reaction was done as described above except that the annealing temperature was changed to 55°C and the extension time was extended to 2 min for either 40 or 45 cycles. PCR products were isolated from agarose gel and subjected to automatic sequencing for confirming their identities.

**cDNA library screening**

AD03b06 Dig-labeled probe was generated using a Dig PCR synthesis kit (Roche) with primers 5'-ACTGATTAACAGTGACACTTCAACAGT-GGGTGGGCA-3' and 5'-CTAGATTCGGGCGATGCTTCCCCACAGTCCA-3'. 150 000 phage clones of liver and lung cDNA libraries, respectively, were screened. Positive phage clones were detected with AP-conjugated anti-Dig antibodies and CSPD substrate (Tropix) following the manufacturer’s instruction. Phage DNA was isolated from NZY agar plate for PCR amplification of inserted cDNA. Sequencing reaction was performed using an ABI 3700 sequencer.

**RESULTS**

**Genomic shotgun array hybridization results of five cDNA libraries**

We established a procedure for detecting transcripts expressed from a defined genomic target. A region of ~1.7 Mb was covered by 5376 recombinant plasmids from 15 BAC clones. At the beginning, we checked whether the hybridization efficiency was affected by duplex formation of arrayed probes. Either amine-linked M13 forward primer or amine-linked M13 reverse primer or both were used in PCR amplification of a subset of these inserts. After coupling to array slides and denaturation, single-stranded DNA probes or double-stranded DNA probes that contained amino-group modification at their 5'-end were covalently linked to slides. Hybridization results indicated that the signals with single-stranded DNA probes were about twice the intensity of double-stranded DNA probes (data not shown). Therefore, in the following experiments, we only used probes with a 5'-end amine group in one strand.

The amplified inserts were arrayed in duplicates together with 24 controls. All probe elements have been sequenced, and the total length of probe sequences was 12.96 Mb, equivalent to 7.65× coverage of the genomic segment. We PCR labeled and hybridized, individually, five samples of bone marrow, colon, liver, fetal liver and lung cDNA libraries to the GSA. A total of 1210 hybridization-positive subclones, whose sequences were readily available, were detected in at least one of the five cDNA libraries. From the hybridization results, we determined that the overall hybridization patterns were similar when the five different cDNA libraries were used as labeled targets, but some difference did exist among the hybridized subclones (data not shown). We also concluded that most positive subclones were not distributed evenly on shotgun fragments of 4q22-24, but rather, formed clusters along the overlapping clone path of 4q22-24. A typical result is shown in Figure 1. This clustering was not caused by uneven distribution of the spotted subclones, as alignment of the arrayed subclones on the finished BAC DNA sequence demonstrated that they fully and evenly covered the entire insert span. Therefore, the results suggested the existence of putative exons. In fact, by sequence comparison we confirmed that most clusters indeed contain exons of known genes (Fig. 1). In contrast, most hybridization-negative subclones were distributed on the intronic regions of known genes and intergenic regions (Fig. 1). We didn’t detect any new transcript within four such regions when we applied the same procedure used to identify new transcripts within the sequences of hybridization-positive subclones (data not shown). Table 1 summarizes the BLAST search results of all positive subclones that hybridized to at least one of the five cDNA libraries. There were 153 exons belonging to 53 known and predicted genes within the studied region. Among the 1210 hybridization-positive subclones, a total of 240...
subclones corresponded to 128 exons out of the 153 known and predicted exons (Table 1). This result indicated that the false-negative rate for reporting an exon was at most 17% (25/153). We then analyzed the other positive subclones with BLAST using the EST database (ftp://ftp.ncbi.nih.gov/blast; NCBI BLAST to nr, EST and proteins). We considered 505 and five subclones that showed an E value less than 1E-80 and an identity >90% as identical to the corresponding ESTs. Together with subclones corresponding to known and predicted genes, we defined these 745 subclones, accounting for 61.6% of total hybridization-positive subclones, as true positives. Of the remaining 465 hybridization-positive subclones, 59 subclones had an E value smaller than 1E-80 and an identity between 80 and 90%, 240 subclones had an E value between 1E-80 and 1E-20 with an identity >80%. We considered these subclones to be homologous to the identified ESTs and the hybridization signals might be from cross hybridization. There were only 166 subclones that did not display any significant homology to the annotated genes or ESTs. This result could be due to non-specific hybridization and was counted as false positive. Alternatively, a plausible explanation, which we are in favor of, is that a fraction of these subclones could correspond to unidentified transcripts. We analyzed these subclones by BLASTX search and demonstrated that some of them contained putative coding sequences that were homologs to existing polypeptide sequences. Protein prediction programs also demonstrated that 62 subclones of this category contained coding sequences of at least 100 amino acids.

Identification of new transcripts

After running protein prediction programs, we selected 17 subclones that contained at least one predicted coding sequences of more than 100 amino acids for further analysis. We grouped these subclones into four clusters according to their physical proximity. The subclone that contained the maximal number of predicted coding sequences and the highest similarity to the known or predicted proteins from each group was selected for primer design. To confirm the existence of new transcripts containing the predicted coding sequences, we performed PCR cloning using primers corresponding to the predicted coding sequences in these selected subclones. An example is shown in Figure 2. Six coding sequences, designated CDS1–CDS6, were predicted in subclone DN04a01 (Fig. 2A). After masking repeated sequences, we designed two pairs of primers to detect the existence of CDS2 and CDS5 or CDS6, respectively. We detected a single band of PCR product using primers specific to CDS5 and CDS6 in the liver cDNA library that was used for microarray hybridization as well as first strand cDNA from a liver tissue sample. As shown in Figure 2B, there was no PCR product detected in normal liver total RNA treated with DNaseI but not with reverse transcription (lane 1). On the other hand, a PCR product was detected in the same sample treated with DNaseI followed by reverse transcription (lane 2). Sequencing confirmed that the PCR product matched exactly to the overlapped region of CDS5 and CDS6. These results indicate that we have identified a new transcript within this DNA fragment rather than detecting false positive due to DNA contamination.

Identification of a new locus of a known gene

In addition to validating the existence of new CDS-containing transcripts by PCR analysis, we also conducted cDNA library screening to confirm the existence of transcripts identified by GSA. Three hybridization-positive subclones, AD03b06, AD03a07 and AD03e10, matched to a 3300 bp region on 4q24 (Fig. 3A), where several exons were predicted by different computational methods (Fig. 3B). Although there was an mRNA assigned in this region, no known genes were located in it at the time the study was conducted (based on the annotation of human genome from NCBI built 30 and UCSC Genome Browser hg10). The structure of predicted exons by different prediction programs and well-characterized mRNA are not identical to one another. In spite of the inconsistency of exon boundaries, overlapping of the three hybridization-positive subclones and the predicted exons and mRNA suggested the existence of a transcript within this region. Therefore, we used AD03b06 as a DNA template to generate a probe and screened liver and lung cDNA libraries (Fig. 3C). We identified hundreds to thousands of positive clones hybridized to AD03b06 probe and picked a total of 60 positives for DNA sequencing. The sequences of positive cDNA clones demonstrated that all of them are from the same gene. The cloned cDNA sequences (1750 bp) were 96% identical to AD03b06 and its corresponding genomic sequence. Sequence annotation of human genome from NCBI built 30 and UCSC Genome Browser hg10). The structure of predicted exons by different prediction programs and well-characterized mRNA are not identical to one another. In spite of the inconsistency of exon boundaries, overlapping of the three hybridization-positive subclones and the predicted exons and mRNA suggested the existence of a transcript within this region. Therefore, we used AD03b06 as a DNA template to generate a probe and screened liver and lung cDNA libraries (Fig. 3C). We identified hundreds to thousands of positive clones hybridized to AD03b06 probe and picked a total of 60 positives for DNA sequencing. The sequences of positive cDNA clones demonstrated that all of them are from the same gene. The cloned cDNA sequences (1750 bp) were 96% identical to AD03b06 and its corresponding genomic sequence without interruption by introns (Fig. 3C). After BLAST search using nucleotide sequence database, we found that in fact it was a known gene, EEF1A1, which is a highly abundant transcript in cells. We learned that, based on FISH experiments, the EEF1A1 gene was located on chromosome 6, and several cross-hybridization signals to EEF1A1 probe were detected on other chromosomes including chromosome 4, but with a less significant signal (30). The EEF1A1 locus identified here by this study is very likely a pseudo-gene instead of an EEF1A1-like gene since there are several nucleic acid changes that introduce stop codons in the coding region. Nevertheless, the result demonstrates the feasibility of the GSA protocol in identifying a candidate gene, even when it is not consistently predicted by computational methods. Using the same procedure, we also discovered another expressed

<table>
<thead>
<tr>
<th>Subjects of BLAST search</th>
<th>Identity</th>
<th>Expect value</th>
<th>Number of clones</th>
<th>%</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known and predicted exons</td>
<td>100%</td>
<td>0</td>
<td>240</td>
<td>19.8</td>
<td>True positives</td>
</tr>
<tr>
<td>ESTs</td>
<td>≥90%</td>
<td>≤1E-80</td>
<td>505</td>
<td>41.8</td>
<td>True positives</td>
</tr>
<tr>
<td>ESTs</td>
<td>&lt;90% and &gt;80%</td>
<td>≤1E-80</td>
<td>59</td>
<td>4.9</td>
<td>Homologs</td>
</tr>
<tr>
<td>ESTs</td>
<td>≥80%</td>
<td>&gt;1E-80 and ≤1E-20</td>
<td>240</td>
<td>19.8</td>
<td>Homologs</td>
</tr>
<tr>
<td>ESTs</td>
<td>No restriction</td>
<td>&gt;1E-20</td>
<td>83</td>
<td>6.8</td>
<td>False positives or new transcripts</td>
</tr>
<tr>
<td>ESTs</td>
<td>0</td>
<td></td>
<td>83</td>
<td>6.8</td>
<td>False positives or new transcripts</td>
</tr>
</tbody>
</table>
transcript, and the corresponding gene (:DAPP1:) was only recently identified by other groups (31). Therefore, we conclude that this GSA screening method is useful for localizing known genes as well as novel transcripts from a defined region on a chromosome.

Identification of new transcripts inside known genes

Among the 505 hybridization-positive subclones that contained sequences identical to ESTs, we found that some were located between two exons of characterized genes. This could be due to alternative splicing of the corresponding genes or the existence of new transcripts within the intronic region of another gene. To test the possibility of alternative splicing, we chose as a target for further study two hybridization-positive subclones, K07b02 and K07d03, which overlapped with each other and are located between exons 16 and 17 of the :NF-κB1: gene. We designed primers corresponding to exons 16 and 17 of the :NF-κB1: gene and performed PCR cloning using a fetal liver cDNA library as DNA template. The result of PCR amplification indicated that there was no additional exon between exons 16 and 17 under several PCR conditions. Therefore, we checked whether the two subclones included a new transcript. Since the two subclones contained sequences identical to two separated ESTs, we proposed that these two ESTs were actually derived from the same transcript. To test this hypothesis, we performed PCR cloning using primers specific to individual ESTs. A single-band PCR product with a molecular weight slightly larger than the size of a two-exon transcript was detected in a fetal liver cDNA library as well as in two HCC cell lines, HepG2 and HA22T/VGH. DNA sequencing confirmed that the two ESTs were included in one transcript and there was no intronic structure in this new transcript (Fig. 4A). To confirm that this result was not due to genomic DNA contamination, we performed the same PCR in total RNA of Hep G2 with and without DNaseI treatment. As shown in Figure 4B, a PCR product was detected in the sample prior to DNaseI treatment after 45 cycles of PCR amplification (lane 2). With DNaseI treatment but without reverse transcription, we did not detect any PCR product under the same PCR conditions (lane 4). In contrast, we did detect PCR product in the sample treated with DNaseI followed by reverse transcription (lanes 5 and 6). This result demonstrated that we identified and cloned a new transcript from within the intronic sequence. Based on these studies, we concluded that GSA can provide a lead to the existence of a new transcript inside well-characterized gene structure.

DISCUSSION

In this study, we demonstrated that transcriptional activity from a genomic segment can be systematically scrutinized by GSA. Moreover, the differential transcription activity among different tissues in a specific chromosome region can be studied. The sensitivity of this method was conservatively estimated to be 83% from the analysis of the hybridization result of known and predicted exons. The specificity was at
least 86%, since only 166 subclones among 1210 hybridization-positive subclones did not show sequence identity to existing ESTs. This method is also an effective means to document expression of predicted exons in a specific tissue. Since computational prediction programs only provide the information about the existence of exons but not tissue distribution, an experimental scheme is needed to validate spatial and temporal expression of any predicted exon. One advantage of GSA is its rich information content generated through multi-layer of genomic coverage and multi-fold of experimentation. Overlapping DNA fragments with or without annotation are included in probe sets. Thus, not only transcriptional activity of known genes but also that of unidentified genes can be detected. Our results indicated that an intronless gene can be detected by the GSA method, while it is less often predicted by computational programs (14). In addition, we also demonstrated that some separated ESTs actually correspond to a single transcript.

Microarray hybridization and cDNA screening indicated that there was a new locus of the EEF1A1 gene on 4q24. It was confirmed to be an intronless gene at this location by sequence comparison. Although it is highly possible to be a pseudo-gene related to EEF1A1, our result demonstrates that GSA, in practice, is a more sensitive method for identifying single-exon genes than computational programs. We also found that most of the hybridization-positive subclones were identical or similar to identified ESTs. Interestingly, some ESTs were localized inside well-characterized genes. PCR reactions were conducted to distinguish the possibility of alternative splicing from the possibility of independent transcripts. One case we demonstrated in this report showed that an independent transcript containing two separated ESTs was located between exons 16 and 17 of the NF-kB1 gene. This result strongly suggests that the so-called intronic regions of known genes could be the location of a new independent transcript. GSA can pinpoint putative new transcripts inside known genes and provides a new direction of gene expression study. Although the function of RNAs transcribed from intronic sequences remains unknown, attention should be directed to these intronic transcripts since they could be alternatively spliced exons, new genes or non-coding RNAs that regulate transcriptional or translational activity of genes (32–34). In this regard, it is noticeable that there are several studies estimating high frequency of alternative splicing events of human genes using EST collections (35–38).

One theoretical disadvantage of GSA is the long length of probes. Since the average insert size of our shotgun library subclones is ~3 kb, two or more predicted exons or coding sequences could be included in a single subclone. It takes extra effort to verify whether there is a new transcript in a hybridization-positive subclone when more than two separated coding sequences are predicted. In this aspect, oligo-nucleotide array sets developed by Gingeras and colleagues could locate exons on genomic DNA more precisely (19).

Figure 3. Identification of a new locus of the EEF1A1 gene. (A) Three hybridization-positive subclones, AD03a07, AD03b06 and AD03e10, are aligned on a 73 kb interval of 4q24. (B) Two slightly different predicted genes and one mRNA were located within this contig. Color boxes represent exons and black solid lines indicate introns. (C) Arrows indicate primers used to amplify DNA probe from subclone AD03b06 that was used in cDNA library screening. Sequence comparison of AD03b06 and isolated clones from cDNA library screening indicated that an EEF1A1 pseudogene is presented as a single-exon gene within AD03b06.
However, due to the short length of oligo-nucleotide probes and repetitive sequences in genomic DNA, cross hybridization or false positive is a problem when using genomic oligo-nucleotide array to detect transcriptional activity. We propose combining both methods to identify new transcripts to improve accuracy and reduce cost. For a first-line screen, GSA is recommended as it can cover a large region with a small number of probes. After filtering out the well-characterized genes and identified ESTs, the remaining hybridization-positive subclones are templates for making probes for an oligo-nucleotide array. Using this two-tier approach, putative transcripts can be initially screened in a high-throughput manner by our GSA protocol. The gene structure can then be effectively determined by a small-size oligonucleotide array.

Finally, one suitable target for GSA application is to establish microarray probe sets for a microbial genome. We envision that a microbial GSA can be used to monitor sequence variation in different isolates, in addition to gene prediction. As whole-genome shotgun sequencing of a microbial genome has become a common practice and a typical bacterial genome can effectively be covered by several thousand probes, we recommend that GSA be applied to this field of research to maximize the utility of microbial genome sequences. Indeed, very recently, Broekhuijsen et al. have reported their results on analysis of genetic variations of four *Francisella tularensis* subspecies using genome-wide DNA microarray (39).

In summary, our study demonstrated that GSA is useful in identifying new transcripts, discovering new loci of known genes, analyzing alternative splicing forms and establishing linkage between ESTs. Since there have been many shotgun libraries generated for genome sequencing, our method provides a powerful and convenient tool to fully utilize these DNA fragments and sequence information. Instead of detecting transcriptional activity of one tissue at a time, GSA can also be used to interrogate gene expression from multiple tissues for comparative analysis. Because of the complete coverage of genomic sequences by overlapping probes in GSA, not only the expression profiles of well-characterized genes can be studied, but also that of new transcripts and/or alternative splicing forms of known genes. In the current study, we have investigated HCC using this protocol and isolated several differentially expressed genes from the 4q22-24 region. We figure that, by coupling transcript identification with genomic sequencing, the GSA method reported in this communication can be applied to research that requires high-throughput gene identification and transcript analysis.

**ACKNOWLEDGEMENTS**

We thank Chih-Yi Huang for preparing probes for microarrays and Keh-Ming Wu for maintaining the BAC-derived genome sequence database. This work was supported by the intramural fund of the National Health Research Institutes, Taipei, Taiwan.
REFERENCES


