DIFFERENTIALLY REGULATED HEPATOCELLULAR CARCINOMA GENES AND USES THEREOF

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims the benefit of U.S. provisional application number 60/475,508, filed June 4, 2003, which is hereby incorporated by reference.

FIELD OF THE INVENTION

The invention is in the field of diagnostics and therapeutics for cancer. More specifically, the invention is in the field of diagnostics and therapeutics for hepatocellular carcinoma.

BACKGROUND OF THE INVENTION

Hepatocellular carcinoma (HCC) is the most common primary malignant tumor of the liver that accounts for more than 70% of liver cancers worldwide (Parkin et al., 1999). Many risk factors have been associated with the development of HCC, including hepatitis B (HBV) and hepatitis C (HCV) viral infection, cirrhosis, male gender, exposure to toxins, etc. Death generally occurs due to liver failure associated with cirrhosis and/or rapid outgrowth of multiple nodules. Approximately 0.25-1 million new cases of HCC are diagnosed each year, and the cancer is especially prevalent in Southeast Asia, China, and sub-Saharan Africa. While surgical resection is considered to be the main curative treatment, only 10-15% of cases are suitable for surgery at the time of presentation. This is because either the disease is detected at an advanced stage at presentation or the underlying poor liver functional reserve precluded surgical intervention.

Diagnosis of HCC has included detection of the presence of a liver mass on radiological investigations and the detection of elevated serum alpha fetoprotein (AFP) levels (Yu and Keeffe, 2003). However, elevation of AFP is not exclusive to HCC and has been observed in benign hepatic disease, such as liver cirrhosis, and other cancers such as germ cell cancer (Bosl and Head, 1994). Treatment of HCC has included interferon therapy and antiviral drugs, but the results have proved

unpredictable and the effectiveness may be limited (Lee, 1997; Yu and Keeffe, 2003). Microarrays have been used to address changes in gene expression of HCC (Chen et al, 2002, Okabe et al, 2001; Honda et al, 2001; Shirota et al, 2001; Tackels-Horne et al, 2001; Xu et al, 2001a; Xu et al, 2001b). However, these reports were restricted to the tissue samples selected for each study and exhibited wide variation in the results, thus limiting the potential significance and utility of the data.

SUMMARY OF THE INVENTION

The invention provides in part molecular markers for hepatocellular carcinoma (HCC) that may be used for HCC diagnosis, to assess HCC progression or regression, or the efficacy and/or toxicity of HCC therapeutics, and/or to identify candidate compounds for HCC therapy, with high predictive accuracy.

In one aspect, the invention provides a composition including an addressable collection of two or more nucleic acid molecules, or polypeptides encoded by these nucleic acid molecules, that are differentially expressed in hepatocellular carcinoma, where the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4 or complements, fragments, variants, or analogs thereof. The composition may include all of the nucleic acid molecules, or their encoded polypeptides, set forth in any one or more of Tables 1 through 4 or complements, fragments, variants, or analogs thereof, or any subset of these nucleic acid molecules or polypeptides. The nucleic acid molecules or polypeptides may be differentially expressed between hepatocellular carcinoma tissue and non-tumor tissue. The nucleic acid molecules or the polypeptides may be attached to a solid support. The compositions may be used in the preparation of a medicament for diagnosis or therapy of hepatocellular carcinoma.

In other aspects, the invention provides a method of diagnosing hepatocellular carcinoma in a subject by obtaining a sample from the subject and detecting the level of expression of two or more nucleic acid molecules or expression products thereof in the sample, where the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4 or complements, fragments, variants, or analogs thereof.

In other aspects, the invention provides a method of monitoring the progression of hepatocellular carcinoma in a subject by obtaining a sample from the subject and detecting the level of expression of two or more nucleic acid molecules or expression products thereof in the sample, where the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4, or complements, fragments, variants, or analogs thereof. The sample may be obtained at two or more time points. The method may further include comparing the level of expression of the nucleic acid molecules or expression products at two or more time points.

In other aspects the invention provides a method of monitoring the efficacy of a hepatocellular carcinoma therapy in a subject by administering the therapy to the subject, obtaining a sample from the subject, and detecting the level of expression of two or more nucleic acid molecules or expression products thereof in the sample, where the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4, or complements, fragments, variants, or analogs thereof. The therapy may be administered at two or more administration time points. The sample may be obtained at two or more sampling time points. The method may further include comparing the level of expression of the nucleic acid molecules or expression products at two or more administration time points, and/or at two or more sampling time points.

In other aspects, the invention provides a method of screening a compound for treating hepatocellular carcinoma by contacting a sample with a test compound and detecting the level of expression of two or more nucleic acid molecules or expression products thereof in the sample, where the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4, or complements, fragments, variants or analogs thereof.

In alternate embodiments of the various aspects, the sample may be liver or serum, or may be suspected of being cancerous, or may be non-cancerous. The methods may further include comparing the level of expression of the nucleic acid molecules or expression products thereof in a non-cancerous sample and in a sample suspected of being cancerous. Differential expression of the nucleic acid molecules or expression products thereof may be indicative of hepatocellular carcinoma, or of

progression of hepatocellular carcinoma, or of the efficacy of the hepatocellular carcinoma therapy. The subject may be suspected of having hepatocellular carcinoma. The subject may be a human.

In alternate embodiments of the various aspects, the method may further include comparing the level of expression of two or more nucleic acid molecules or expression products thereof with a standard, or further include preparing a gene expression profile. The method may be a high throughput method.

In other aspects, the invention provides a solid support including two or more nucleic acid molecules or polypeptides encoded by these nucleic acid molecules that are differentially expressed in hepatocellular carcinoma, where the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in Tables 1 through 4 or complements, fragments, variants, or analogs thereof. The nucleic acid molecules may consist essentially of all the nucleic acid molecules set forth in any one or more of Tables 1 through 4, and/or be differentially expressed between hepatocellular carcinoma tissue and non-tumor tissue. The polypeptides may consist essentially of the polypeptides encoded by all the nucleic acid molecules set forth in any one or more of Tables 1 through 4, and/or be differentially expressed between hepatocellular carcinoma tissue and non-tumor tissue. The nucleic acid molecules or the polypeptides may be covalently or non-covalently attached to the solid support (e.g., a microarray).

In other aspects, the invention provides a database including information identifying the expression level in liver tissue (e.g., cancerous or non-cancerous tissue) of two or more nucleic acid molecules or expression products thereof, where the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4, or complements, fragments, variants, or analogs thereof.

A "composition" as used herein includes a plurality of the nucleic acid molecules described herein, including complements, analogs, variants, and fragments thereof. A composition as used herein also includes a plurality of polypeptides encoded by the nucleic acid molecules described herein, and complements, analogs, variants, and fragments thereof. A composition as used herein also includes a plurality of polypeptides capable of specifically binding to the polypeptides or nucleic acid

molecules described herein (e.g., antibodies). The composition may include any combination of the nucleic acid molecules described herein, including complements, analogs, variants, and fragments thereof, or polypeptides encoded by these nucleic acid molecules. Accordingly, the composition may include 2, 3, 4, 5, 6, 7, 8, 9, 10, etc. up to all of the nucleic acid molecules or polypeptides described herein, e.g., in any one or more of the Tables or Figures herein. In some embodiments, the composition may include subsets of the nucleic acid molecules or polypeptides described herein, e.g., in any one or more of the Tables or Figures herein, for example, subsets groups by protein function or characteristics, e.g., proteins involved in the ubiquitination pathway, or proteins localized to a particular cellular compartment. These nucleic acid molecules or polypeptides may for example be used with a substrate (e.g, a solid substrate or a liquid substrate) in a variety of applications, including the diagnosis of HCC, or monitoring the progression of HCC.

By "addressable collection" is meant a combination of nucleic acid molecules or polypeptides capable of being detected by, for example, the use of hybridization techniques or antibody binding techniques or by any other means of detection known to those of ordinary skill in the art.

The terms "nucleic acid" or "nucleic acid molecule" encompass both RNA (plus and minus strands) and DNA, including cDNA, genomic DNA, and synthetic (e.g., chemically synthesized) DNA. The nucleic acid may be double-stranded or single-stranded. Where single-stranded, the nucleic acid may be the sense strand or the antisense strand. A nucleic acid molecule may be any chain of two or more covalently bonded nucleotides, including naturally occurring or non-naturally occurring nucleotides, or nucleotide analogs or derivatives. By "RNA" is meant a sequence of two or more covalently bonded, naturally occurring or modified ribonucleotides. One example of a modified RNA included within this term is phosphorothioate RNA. By "DNA" is meant a sequence of two or more covalently bonded, naturally occurring or modified deoxyribonucleotides. By "cDNA" is meant complementary or copy DNA produced from an RNA template by the action of RNA-dependent DNA polymerase (reverse transcriptase). Thus a "cDNA clone" means a duplex DNA sequence complementary to an RNA molecule of interest, carried in a cloning vector. An "oligonucleotide" as used herein is a single stranded molecule

which may be used in hybridization or amplification technologies. In general, an oligonucleotide may be any integer from about 15 to about 100 nucleotides in length, but may also be of greater length. A "probe" or "primer" is a single-stranded DNA or RNA molecule of defined sequence that can base pair to a second DNA or RNA molecule that contains a complementary sequence (the target). The stability of the resulting hybrid molecule depends upon the extent of the base pairing that occurs, and is affected by parameters such as the degree of complementarity between the probe and target molecule, and the degree of stringency of the hybridization conditions. The degree of hybridization stringency is affected by parameters such as the temperature, salt concentration, and concentration of organic molecules, such as formamide, and is determined by methods that are known to those skilled in the art. Probes or primers specific for the nucleic acid sequences described herein, or portions thereof, may vary in length by any integer from at least 8 nucleotides to over 500 nucleotides, including any value in between, depending on the purpose for which, and conditions under which, the probe or primer is used. For example, a probe or primer may be 8, 10, 15, 20, or 25 nucleotides in length, or may be at least 30, 40, 50, or 60 nucleotides in length, or may be over 100, 200, 500, or 1000 nucleotides in length. Probes or primers specific for the nucleic acid molecules described herein may have greater than any integer between 20-30% sequence identity, or at least any integer between 55-75% sequence identity, or at least any integer between 75-85% sequence identity, or at least any integer between 85-99% sequence identity, or 100% sequence identity to the nucleic acid sequences described herein. Probes or primers can be detectably-labeled, either radioactively or non-radioactively, by methods that are known to those skilled in the art. Probes or primers can be used for methods involving nucleic acid hybridization, such as nucleic acid sequencing, nucleic acid amplification by the polymerase chain reaction, single stranded conformational polymorphism (SSCP) analysis, restriction fragment polymorphism (RFLP) analysis, Southern hybridization, northern hybridization, in situ hybridization, electrophoretic mobility shift assay (EMSA), microarray, and other methods that are known to those skilled in the art. Probes or primers may be derived from genomic DNA or cDNA, for example, by amplification, or from cloned DNA segments, or may be chemically synthesized.

The "expression product" of a nucleic acid molecule may be any polypeptide encoded by that nucleic acid molecule. Generally, the polypeptide is capable of being expressed.

A "protein," "peptide" or "polypeptide" is any chain of two or more amino acids, including naturally occurring or non-naturally occurring amino acids or amino acid analogues, regardless of post-translational modification (e.g., glycosylation or phosphorylation). An "amino acid sequence", "polypeptide", "peptide" or "protein" of the invention may include peptides or proteins that have abnormal linkages, cross links and end caps, non-peptidyl bonds or alternative modifying groups. Such modified peptides are also within the scope of the invention. The term "modifying group" is intended to include structures that are directly attached to the peptidic structure (e.g., by covalent coupling), as well as those that are indirectly attached to the peptidic structure (e.g., by a stable non-covalent association or by covalent coupling to additional amino acid residues, or mimetics, analogues or derivatives thereof, which may flank the core peptidic structure). For example, the modifying group can be coupled to the amino-terminus or carboxy-terminus of a peptidic structure, or to a peptidic or peptidomimetic region flanking the core domain. Alternatively, the modifying group can be coupled to a side chain of at least one amino acid residue of a peptidic structure, or to a peptidic or peptido- mimetic region flanking the core domain (e.g., through the epsilon amino group of a lysyl residue(s), through the carboxyl group of an aspartic acid residue(s) or a glutamic acid residue(s), through a hydroxy group of a tyrosyl residue(s), a serine residue(s) or a threonine residue(s) or other suitable reactive group on an amino acid side chain). Modifying groups covalently coupled to the peptidic structure can be attached by means and using methods well known in the art for linking chemical structures, including, for example, amide, alkylamino, carbamate or urea bonds. Peptides according to the invention may include peptides encoded by the nucleic acid molecules of Tables 1 through 4 or complements or analogs thereof.

By "differential expression" or "differentially expressed" is meant increased, upregulated or present, or decreased, downregulated or absent, gene expression as detected by the absence, presence, or change (up or down) in the amount of transcribed messenger RNA or translated protein in a sample. For example, the

change may be detected by comparison of the change in gene expression level between a HCC sample and a non-tumor sample. The absolute amount of change of gene expression is not important, as long as the amount of change is reproducible, and measurable. In some embodiments, the change (up or down) in the amount of transcribed messenger or translated protein may be at least 1-fold or at least 1.5-fold or may be over 2.0, 2.5., 3.0, 3.5, 4.0, 4.5, or 5.0-fold. In some embodiments, the change in the amount of transcribed messenger or translated protein may be 40%, 50%, 60%, 70%, 80%, 90%, or 100%.

By "detecting" it is intended to include determining the presence or absence, or quantifying the amount, of a nucleic acid molecule or polypeptide of the invention a substance. The term thus refers to the use of the materials, compositions, and methods of the present invention for qualitative and quantitative determinations. For example, detecting an increase in gene expression levels may include quantifying a change of any value between 10% and 90%, or of any value between 30% and 60%, or over 100%, of any of the nucleic acid molecules or polypeptides of the invention when compared to a control. In other embodiments, detecting an increase in gene expression levels may include quantifying a change of any value between 1 to 5 fold or more of any of the nucleic acid molecules or polypeptides of the invention when compared to a control.

"Hepatocellular carcinoma" is cancer that arises from hepatocytes, the major cell type of the liver. It is a form of adenocarcinoma, and is the most common type of liver tumor. "Non-tumor" tissue refers to tissue or cells that are non-cancerous. In some embodiments, non-tumor tissue may include tissue or cells from a subject having a liver disorder, such as HBV or HCV infection, cirrhosis, exposure to aflatoxins, etc. The phrase "suspected of being cancerous" as used herein means a HCC tissue sample believed by one of ordinary skill in the art to contain HCC cells. By "non-cancerous" or "non-tumor" is meant a tissue sample demonstrated by standard diagnostic or other techniques (e.g., histologic staining, microscopic analysis, immunoassay, etc.) to contain no HCC cells or evidence of HCC.

A "sample" can be any organ, tissue, cell, or cell extract isolated from a subject, such as a sample isolated from a mammal having a hepatocellular carcinoma or isolated from a mammal not having a hepatocellular carcinoma or a tumor. For

example, a sample can include, without limitation, tissue such as liver tissue (e.g., from a biopsy or autopsy), cells, peripheral blood, whole blood, red cell concentrates, platelet concentrates, leukocyte concentrates, blood cell proteins, blood plasma, platelet-rich plasma, a plasma concentrate, a precipitate from any fractionation of the plasma, a supernatant from any fractionation of the plasma, blood plasma protein fractions, purified or partially purified blood proteins or other components, serum, semen, mammalian colostrum, milk, urine, stool, saliva, placental extracts, amniotic fluid, a cryoprecipitate, a cryosupernatant, a cell lysate, mammalian cell culture or culture medium, products of fermentation, ascitic fluid, proteins present in blood cells, solid tumours isolated from a mammal with a hepatocellular carcinoma, or any other specimen, or any extract thereof, obtained from a patient (human or animal), test subject, or experimental animal. A sample may also include, without limitation, products produced in cell culture by normal, non-tumor, or transformed cells (e.g., via recombinant DNA technology). A "sample" may also be a cell or cell line created under experimental conditions, that are not directly isolated from a subject. A sample can also be cell-free, artificially derived or synthesised. In some embodiments, samples refer to liver tissue or cells. In some embodiments, the liver tissue may be from a subject having a hepatocellular carcinoma; a subject infected with a hepatitis virus; a subject having a liver disorder e.g., cirrhosis, or a subject having a normal liver e.g., not diagnosed with or suspected of having a liver disorder.

As used herein, a subject may be a human, non-human primate, rat, mouse, cow, horse, pig, sheep, goat, dog, cat, etc. The subject may be a clinical patient, a clinical trial volunteer, an experimental animal, etc. The subject may be suspected of having HCC, be diagnosed with HCC, or be a control subject that is confirmed to not have HCC. Diagnostic methods for HCC and the clinical delineation of HCC diagnoses are known to those of ordinary skill in the art, and include biopsy including radiological biopsy by means of a radiological scan, laproscopy, or open surgical biopsy.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a plot showing natural patterns of gene expression differences between HCC tumor and non-tumor liver tissue specimens based on unsupervised

clustering. The plot shows the variance of expression value for each of the gene features across all the HCC tumor and non-tumor liver tissue specimens. The dotted line indicates the 500 most variable gene features.

Figure 2 is a multidimensional scaling plot showing significant gene differential expression between HCC tumor and non-tumor liver tissues, and comparison with liver cancer cell lines ($P<1x10^{-6}$, approximately 1.5-fold change). The plot illustrates the ability of the 218 outlier genes to separate HCC tumor specimens (black circles) from non-tumor liver tissue specimens (dark gray circles). The plot also shows how different liver cancer cell lines (light gray circles) are from the clinical tissue samples.

Figures 3A-B characterize differentially expressed genes in HCC tumor specimens ($P<1x10^{-6}$, approximately 1.5-fold change). Figure 3A is a bar graph showing the chromosomal distribution of the 218 outlier genes. The dark colored and light shaded bars represent genes that are at least 1.5-fold up- and downregulated, respectively, in HCC tumors relative to non-tumor livers. Figure 3B is a bar graph showing the functional characterization of the outlier genes based on Gene Ontology and published works.

Figure 4 is a bar graph showing the expression of BMI-1 in HCC tumors as determined by cDNA microarray analysis. The data are presented as the level of expression (log base 2) in each HCC tumor specimen with respect to the corresponding non-tumor liver sample.

Figures 5A-D show real-time RT-PCR analysis of IGFBP3, ERBB3, ERBB2 and EGFR in HCC tumor samples. The gene expression patterns for (A) IGFBP3, (B) ERBB3, (C) ERBB2 and (D) EGFR in all the 37 HCC tumor samples and their corresponding non-tumor liver tissue specimens were examined. All data was normalized to the amount of 'housekeeping' gene PBGD and are presented as relative fold expression change (log base 2 ratio) in HCC tumor specimens with respect to their corresponding non-tumor liver counterpart. Positive value depicts higher expression level, while negative value depicts lower expression level in the tumor relative to the non-tumor specimen.

Figure 6 lists a panel of genes analyzed using real-time and semi-quantitative RT-PCR analyses, and indicating whether the analysis was conducted in non-tumor human tissues, and in clinical tissue samples or HCC cell lines or both.

Figures 7A-C show gene expression analysis of ARMET. Semi-quantitative RT-PCR analysis (A) of non-tumor tissues and HCC cell lines and real time RT-PCR analysis of non-tumor tissues (B) and patient samples (C) was performed. GAPDH expression level was used as the control in the analyses of HCC cell lines vs. non-tumor human tissues (A). The dotted line in (B) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figures 8A-C show gene expression analysis of BMI-1. Semi-quantitative RT-PCR analysis (A) of non-tumor tissues and HCC cell lines and real time RT-PCR analysis of non-tumor tissues (B) and patient samples (C) was performed. GAPDH expression level was used as the control in the analyses of HCC cell lines vs. non-tumor human tissues (A). The dotted line in (B) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figures 9A-C show gene expression analysis of CRHBP. Semi-quantitative RT-PCR analysis (A) of non-tumor tissues and HCC cell lines and real time RT-PCR analysis of non-tumor tissues (B) and patient samples (C) was performed. GAPDH expression level was used as the control in the analyses of HCC cell lines vs. non-tumor human tissues (A). The dotted line in (B) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figures 10A-C show gene expression analysis of CSTB. Semi-quantitative RT-PCR analysis (A) of non-tumor tissues and HCC cell lines and real time RT-PCR analysis of non-tumor tissues (B) and patient samples (C) was performed. GAPDH expression level was used as the control in the analyses of HCC cell lines vs. non-tumor human tissues (A). The dotted line in (B) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figures 11A-C show gene expression analysis of DPT. Semi-quantitative RT-PCR analysis (A) of non-tumor tissues and HCC cell lines and real time RT-PCR analysis of non-tumor tissues (B) and patient samples (C) was performed. GAPDH expression level was used as the control in the analyses of HCC cell lines vs. non-

tumor human tissues (A). The dotted line in (B) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figures 12A-B show gene expression analysis of ERBB3. Real time RT-PCR analysis of non-tumor tissues (A) and patient samples (B) was performed. The dotted line in (A) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figures 13A-B show gene expression analysis of EZH2. Real time RT-PCR analysis of non-tumor tissues (A) and patient samples (B) was performed. The dotted line in (A) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figures 14A-B show gene expression analysis of GPC3. Real time RT-PCR analysis of non-tumor tissues (A) and patient samples (B) was performed. The dotted line in (A) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figures 15A-B show gene expression analysis of HDGF. Real time RT-PCR analysis of non-tumor tissues (A) and patient samples (B) was performed. The dotted line in (A) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figures 16A-B show gene expression analysis of MDK. Real time RT-PCR analysis of non-tumor tissues (A) and patient samples (B) was performed. The dotted line in (A) indicates mean expression value of four non-tumor liver tissues i.e., Fetal/F, Fetal/M, Adult/F, Adult/M.

Figure 17 shows gene expression analysis of D123. Semi-quantitative RT-PCR analysis of non-tumor tissues and HCC cell lines was performed. GAPDH expression level was used as the control in the analyses of HCC cell lines vs. non-tumor human tissues.

Figure 18 shows gene expression analysis of FLJ10326. Semi-quantitative RT-PCR analysis of non-tumor tissues and HCC cell lines was performed. GAPDH expression level was used as the control in the analyses of HCC cell lines vs. non-tumor human tissues.

Figure 19 shows gene expression analysis of ICA-1A. Semi-quantitative RT-PCR analysis of non-tumor tissues and HCC cell lines was performed. GAPDH

expression level was used as the control in the analyses of HCC cell lines vs. non-tumor human tissues.

Figure 20 shows gene expression analysis of LASP1. Semi-quantitative RT-PCR analysis of non-tumor tissues and HCC cell lines was performed. GAPDH expression level was used as the control in the analyses of HCC cell lines vs. non-tumor human tissues.

Figure 21 shows gene expression analysis of PODXL. Semi-quantitative RT-PCR analysis of non-tumor tissues and HCC cell lines was performed. GAPDH expression level was used as the control in the analyses of HCC cell lines vs. non-tumor human tissues.

DETAILED DESCRIPTION OF THE INVENTION

Phenotypic changes in cancer may be due to cellular changes at the nucleotide level. Thus, some genes may be expressed, overexpressed, or under-expressed in tumor cells relative to non-tumor cells. However, a wide variation exists in gene expression patterns among cancer patients, including HCC patients. Therefore, examining the regulation or expression of a single gene or target, or even of multiple genes or targets whose regulation or expression vary across different HCC tumors, may be insufficient for accurate diagnosis or treatment of HCC or for screening of HCC therapeutics. Selecting a set of differentially expressed HCC genes, nucleic acid molecules, and/or polypeptides, assists in predictable and accurate diagnosis and therapy, and design of efficacious therapeutics.

The invention provides, in part, nucleic acid molecules and polypeptides that are differentially expressed in HCC cells, when compared to non-HCC tissue, e.g., liver or serum. Thus, the invention provides, in part, molecular markers for HCC derived from the analysis of global changes in gene expression ("gene expression profiles") between HCC tissue and non-HCC tissue. More specifically, cDNA microarrays were used to examine the global cellular changes in matched pairs of HCC tumor and non-tumor tissues of patients diagnosed with HCC. In addition, gene expression patterns between primary HCC tumors and liver cancer cell lines were examined for possible biological variation.

The nucleic acid molecules or polypeptides provided by the invention, as well as subsets thereof, serve as molecular markers that may be used for example for HCC diagnosis; to assess HCC progression or regression; to access the efficacy and/or toxicity of HCC therapeutics; and/or to identify candidate compounds for HCC therapy, with high predictive accuracy. The genes lists identified permit rapid, simple, and reproducible screening of a variety of HCC samples by, for example, nucleic acid microarray hybridization or protein expression technology to determine the expression of the specific genes, or by other means such as differential display, gel electrophoresis, genome mismatch scanning, representational discriminate analysis, clustering, transcript imaging, etc. used singly or in combination. Thus, the selected nucleic acid molecules or polypeptides of the invention define standard and reproducible differential expression patterns against which to compare the expression pattern in a variety of tissue or cells, e.g., HCC tissue or cells or serum, obtained by biopsy, autopsy, or from cell lines and/or in vitro treatment or assays. The selected nucleic acid molecules or polypeptides of the invention and subsets thereof provide reliable detection of HCC cells or tissue, with reduction or elimination of false positives or false negatives. In some embodiments, the invention provides composite sets of discriminator genes for use as general or global HCC tumor markers. In some embodiments, the nucleic acid molecules or polypeptides of the invention may be used to assess the suitability of a HCC cell line for use as a model for HCC, as gene expression profiles may vary between primary HCC tumors and HCC cell lines.

Various alternative embodiments and examples of the invention are described herein. These embodiments and examples are illustrative and should not be construed as limiting the scope of the invention.

Nucleic Acid Molecules, Polypeptides, And Test Compounds

Compounds according to the invention include, without limitation, molecules substantially identical to the nucleic acid molecules of Tables 1 through 4 (e.g., BMI-1, ARMET, CRHBP, CSTB, DPT, ERBB3, EZH2, GPC3, HDGF, MDK, D123, FLJ10326, ICAP-1A, LASP1, PODXL) and complements, analogs, fragments, and variants thereof, including, for example, the polypeptides described herein that are encoded by the nucleic acid molecules of Tables 1 through 4, as well as homologs and

fragments thereof. In some embodiments of the invention, compounds of the invention include antibodies that specifically bind to polypeptides encoded by the nucleic acid molecules of Tables 1 through 4. An antibody "specifically binds" an antigen when it recognises and binds the antigen, for example, a polypeptide encoded by any of the nucleic acid molecules described herein, but does not substantially recognise and bind other reference molecules in a sample, for example, a polypeptide that is encoded by a nucleic acid molecule that is not substantially identical to any of the nucleic acid molecules described herein. Such an antibody has, for example, an affinity for the antigen which is at least 10, 100, 1000 or 10000 times greater than the affinity of the antibody for another reference molecule in a sample.

A "substantially identical" sequence is an amino acid or nucleotide sequence that differs from a reference sequence only by one or more conservative substitutions, as discussed herein, or by one or more non-conservative substitutions, deletions, or insertions located at positions of the sequence that do not destroy the biological function of the amino acid or nucleic acid molecule, or that do not destroy the detectability (e.g., by hybridization or specific binding) of the amino acid or nucleic acid molecule. Such a sequence can be any integer from 10% to 99%, or more generally at least 10%, 20%, 30%, 40%, 50, 55% or 60%, or at least 65%, 75%, 80%, 85%, 90%, or 95%, or as much as 96%, 97%, 98%, or 99% identical when optimally aligned at the amino acid or nucleotide level to the sequence used for comparison using, for example, the Align Program (Myers and Miller, CABIOS, 1989, 4:11-17) or FASTA. For polypeptides, the length of comparison sequences may be at least 2, 5, 10, or 15 amino acids, or at least 20, 25, or 30 amino acids. In alternate embodiments, the length of comparison sequences may be at least 35, 40, or 50 amino acids, or over 60, 80, or 100 amino acids. For nucleic acid molecules, the length of comparison sequences may be at least 5, 10, 15, 20, or 25 nucleotides, or at least 30, 40, or 50 nucleotides. In alternate embodiments, the length of comparison sequences may be at least 60, 70, 80, or 90 nucleotides, or over 100, 200, or 500 nucleotides. Sequence identity can be readily measured using publicly available sequence analysis software (e.g., Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 1710 University Avenue, Madison, Wis. 53705, or BLAST software available from the National Library of Medicine, or

as described herein). Examples of useful software include the programs Pile-up and PrettyBox. Such software matches similar sequences by assigning degrees of homology to various substitutions, deletions, substitutions, and other modifications.

Alternatively, or additionally, two nucleic acid sequences may be "substantially identical" if they hybridize under high stringency conditions. In some embodiments, high stringency conditions are, for example, conditions that allow hybridization comparable with the hybridization that occurs using a DNA probe of at least 500 nucleotides in length, in a buffer containing 0.5 M $NaHPO_4$, pH 7.2, 7% SDS, 1 mM EDTA, and 1% BSA (fraction V), at a temperature of 65°C, or a buffer containing 48% formamide, 4.8x SSC, 0.2 M Tris-Cl, pH 7.6, 1x Denhardt's solution, 10% dextran sulfate, and 0.1% SDS, at a temperature of 42°C. (These are typical conditions for high stringency northern or Southern hybridizations.) Hybridizations may be carried out over a period of about 20 to 30 minutes, or about 2 to 6 hours, or about 10 to 15 hours, or over 24 hours or more. High stringency hybridization is also relied upon for the success of numerous techniques routinely performed by molecular biologists, such as high stringency PCR, DNA sequencing, single strand conformational polymorphism analysis, and in situ hybridization. In contrast to northern and Southern hybridizations, these techniques are usually performed with relatively short probes (e.g., usually about 16 nucleotides or longer for PCR or sequencing and about 40 nucleotides or longer for in situ hybridization). The high stringency conditions used in these techniques are well known to those skilled in the art of molecular biology, and examples of them can be found, for example, in Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, New York, N.Y., 1998, which is hereby incorporated by reference.

A "variant" is a nucleic acid molecule that is a recognized variation of a nucleic acid molecule or expression product thereof. Splice variants may be determined for example by using computer programs, e.g, BLAST. Allelic variants have in general a high percent identity to the nucleic acid molecule of interest. "Single nucleotide polymorphism" (SNP) refers to a change in a single base as a result of a substitution, insertion or deletion. The change may be conservative (purine for purine) or non-conservative (purine to pyrimidine) and may or may not result in a change in an encoded amino acid. An "analog" is a nucleic acid molecule or polypeptide that

has been subjected to a chemical modification. Nucleic acid analogs can include substitution of a non-traditional base such as queosine or of an analog such as hypoxanthine, or other substitutions known in the art. Analogs in general retain the biological activities of the naturally occurring molecules but may confer advantages such as longer lifespan or enhanced activity. By "complementary" or "complement" is meant that two nucleic acids, e.g., DNA or RNA, contain a sufficient number of nucleotides which are capable of forming Watson-Crick base pairs to produce a region of double-strandedness between the two nucleic acids. Thus, adenine in one strand of DNA or RNA pairs with thymine in an opposing complementary DNA strand or with uracil in an opposing complementary RNA strand. It will be understood that each nucleotide in a nucleic acid molecule need not form a matched Watson-Crick base pair with a nucleotide in an opposing complementary strand to form a duplex. A nucleic acid molecule is "complementary" to another nucleic acid molecule, or is a "complement" of that other nucleic acid molecule, if it hybridizes, under conditions of high stringency, with the second nucleic acid molecule. The "complement" of a nucleic acid molecule of Tables 1 through 4 may in some embodiments include a nucleic acid molecule that is complementary over the full length of the sequence of a nucleic acid molecule of Tables 1 through 4. A "fragment" may be any portion of a nucleic acid molecule or polypeptide as described herein that is capable of being differentially expressed or detected in an assay or screening method according to the invention.

Various genes and nucleic acid sequences of the invention may be recombinant sequences. The term "recombinant" means that something has been recombined, so that when made in reference to a nucleic acid construct the term refers to a molecule that is comprised of nucleic acid sequences that are joined together or produced by means of molecular biological techniques. The term "recombinant" when made in reference to a protein or a polypeptide refers to a protein or polypeptide molecule which is expressed using a recombinant nucleic acid construct created by means of molecular biological techniques. The term "recombinant" when made in reference to genetic composition refers to a gamete or progeny with new combinations of alleles that did not occur in the parental genomes. Recombinant nucleic acid constructs may include a nucleotide sequence which is ligated to, or is

manipulated to become ligated to, a nucleic acid sequence to which it is not ligated in nature, or to which it is ligated at a different location in nature. Referring to a nucleic acid construct as 'recombinant' therefore indicates that the nucleic acid molecule has been manipulated using genetic engineering, i.e. by human intervention.

Recombinant nucleic acid constructs may for example be introduced into a host cell by transformation. Such recombinant nucleic acid constructs may include sequences derived from the same host cell species or from different host cell species, which have been isolated and reintroduced into cells of the host species. Recombinant nucleic acid construct sequences may become integrated into a host cell genome, either as a result of the original transformation of the host cells, or as the result of subsequent recombination and/or repair events.

As used herein, "heterologous" in reference to a nucleic acid or protein is a molecule that has been manipulated by human intervention so that it is located in a place other than the place in which it is naturally found. For example, a nucleic acid sequence from one species may be introduced into the genome of another species, or a nucleic acid sequence from one genomic locus may be moved to another genomic or extrachromasomal locus in the same species. A heterologous protein includes, for example, a protein expressed from a heterologous coding sequence or a protein expressed from a recombinant gene in a cell that would not naturally express the protein.

A compound is "substantially pure" when it is separated from the components that naturally accompany it. Typically, a compound is substantially pure when it is at least 10%, 20%, 30%, 40%, 50%, or 60%, more generally 70%, 75%, 80%, or 85%, or over 90%, 95%, or 99% by weight, of the total material in a sample. Thus, for example, a polypeptide that is chemically synthesised, produced by recombinant technology, isolated by known purification techniques, will be generally be substantially free from its naturally associated components. A substantially pure compound can be obtained, for example, by extraction from a natural source; by expression of a recombinant nucleic acid molecule encoding a polypeptide compound; or by chemical synthesis. Purity can be measured using any appropriate method such as column chromatography, gel electrophoresis, HPLC, etc. A nucleic acid molecule is substantially pure or "isolated" when it is not immediately contiguous with (i.e.,

covalently linked to) the coding sequences with which it is normally contiguous in the naturally occurring genome of the organism from which the DNA of the invention is derived. Therefore, an "isolated" gene or nucleic acid molecule is intended to mean a gene or nucleic acid molecule which is not flanked by nucleic acid molecules which normally (in nature) flank the gene or nucleic acid molecule (such as in genomic sequences) and/or has been completely or partially purified from other transcribed sequences (as in a cDNA or RNA library). For example, an isolated nucleic acid of the invention may be substantially isolated with respect to the complex cellular milieu in which it naturally occurs. In some instances, the isolated material will form part of a composition (for example, a crude extract containing other substances), buffer system or reagent mix. In other circumstance, the material may be purified to essential homogeneity, for example as determined by PAGE or column chromatography such as HPLC. The term therefore includes, e.g., a recombinant nucleic acid incorporated into a vector, such as an autonomously replicating plasmid or virus; or into the genomic DNA of a prokaryote or eukaryote, or which exists as a separate molecule (e.g., a cDNA or a genomic DNA fragment produced by PCR or restriction endonuclease treatment) independent of other sequences. It also includes a recombinant nucleic acid which is part of a hybrid gene encoding additional polypeptide sequences. Preferably, an isolated nucleic acid comprises at least about 40%, 50%, 60%, 70%, 80%, 90%, 95%, or 99% (on a molar basis) of all macromolecular species present. Thus, an isolated gene or nucleic acid molecule can include a gene or nucleic acid molecule which is synthesized chemically or by recombinant means. Recombinant DNA contained in a vector are included in the definition of "isolated" as used herein. Also, isolated nucleic acid molecules include recombinant DNA molecules in heterologous host cells, as well as partially or substantially purified DNA molecules in solution. In vivo and in vitro RNA transcripts of the DNA molecules of the present invention are also encompassed by "isolated" nucleic acid molecules. Such isolated nucleic acid molecules are useful in the manufacture of the encoded polypeptide, as probes for isolating homologous sequences (e.g., from other mammalian species), for gene mapping (e.g., by in situ hybridization with chromosomes), or for detecting expression of the gene in tissue (e.g., human tissue, such as peripheral blood), such as by Northern blot analysis.

Polypeptide compounds can be prepared by, for example, replacing, deleting, or inserting an amino acid residue at any position of a peptide or a peptide analog, for example, a peptide as described herein, with other conservative amino acid residues, i.e., residues having similar physical, biological, or chemical properties. It is well known in the art that some modifications and changes can be made in the structure of a polypeptide without substantially altering the biological function of that peptide, to obtain a biologically equivalent polypeptide. In one aspect of the invention, polypeptides of the present invention also extend to biologically equivalent peptides that differ from a portion of the sequence of the polypeptides of the present invention by conservative amino acid substitutions. As used herein, the term "conserved amino acid substitutions" refers to the substitution of one amino acid for another at a given location in the peptide, where the substitution can be made without substantial loss of the relevant function. In making such changes, substitutions of like amino acid residues can be made on the basis of relative similarity of side-chain substituents, for example, their size, charge, hydrophobicity, hydrophilicity, and the like, and such substitutions may be assayed for their effect on the function of the peptide by routine testing. Conservative changes can also include the substitution of a chemically derivatised moiety for a non-derivatised residue, by for example, reaction of a functional side group of an amino acid. Peptides or peptide analogs can be synthesised by standard chemical techniques, for example, by automated synthesis using solution or solid phase synthesis methodology. Automated peptide synthesisers are commercially available and use techniques well known in the art. Peptides and peptide analogs can also be prepared using recombinant DNA technology using standard methods such as those described in, for example, Sambrook, *et al.* (Molecular Cloning: A Laboratory Manual. 2nd ed., Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989) or Ausubel *et al.* (Current Protocols in Molecular Biology, John Wiley & Sons, 1994). Computer programs such as LASERGENE software (DNASTAR, Madison Wis.), MACVECTOR software (Genetics Computer Group, Madison Wis.) and RasMol software (www.umass.edu/microbio/rasmol) may be used to determine which and how many amino acid residues in a particular portion of the protein may be

substituted, inserted, or deleted without abolishing biological or immunological activity.

Monitoring changes in gene expression may also be advantageous when screening candidate HCC therapeutics. Often candidate compounds are screened and prescreened for the ability to interact with a major target without regard to other effects they may have on cells or in the subject to be treated, such as toxicity, which prevent the development and use of the potential compound. Thus, the methods of the invention may be used to identify candidate compounds suitable for HCC therapy.

In general, candidate or test compounds are identified from large libraries of both natural products or synthetic (or semi-synthetic) extracts or chemical libraries according to methods known in the art. Those skilled in the field of drug discovery and development will understand that the precise source of test extracts or compounds is not critical to the methods of the invention. Accordingly, virtually any number of chemical extracts or compounds can be screened using the exemplary methods described herein. Examples of such extracts or compounds include, but are not limited to, plant-, fungal-, prokaryotic- or animal-based extracts, fermentation broths, and synthetic compounds, as well as modification of existing compounds. Numerous methods are also available for generating random or directed synthesis (e.g., semi-synthesis or total synthesis) of any number of chemical compounds, including, but not limited to, saccharide-, lipid-, peptide-, and nucleic acid-based compounds. Synthetic compound libraries are commercially available. Alternatively, libraries of natural compounds in the form of bacterial, fungal, plant, and animal extracts are commercially available from a number of sources, including Biotics (Sussex, UK), Xenova (Slough, UK), Harbor Branch Oceanographic Institute (Ft. Pierce, FL, USA), and PharmaMar, MA, USA. Furthermore, if desired, any library or compound is readily modified using standard chemical, physical, or biochemical methods. Candidate compounds useful for treating HCC may also be identified by assessing variations in the expression of one or more HCC markers, from Tables 1 through 4, prior to and after contacting HCC cells or tissues with candidate pharmacological agents for the treatment of HCC. The cells may be grown in culture (e.g. from a HCC cell line), or may be obtained from a subject, (e.g. in a clinical trial of candidate pharmaceutical agents to treat HCC). Alterations in expression of one or more of

HCC nucleic acid markers (drug targets), in HCC cells or tissues tested before and after contact with a candidate pharmacological agent to treat HCC, indicate progression, regression, or stasis of the HCC thereby indicating efficacy of candidate agents and concomitant identification of candidate compounds for therapeutic use in HCC. Candidate compounds may also be screened for toxicity, specificity, etc.

When a crude extract is found to modulate expression levels of any of the nucleic acid molecules or polypeptides of the invention, further fractionation of the positive lead extract is necessary to isolate chemical constituents responsible for the observed effect. Thus, the goal of the extraction, fractionation, and purification process is the careful characterization and identification of a chemical entity within the crude extract having the modulatory activities. The same assays described herein for the detection of activities in mixtures of compounds can be used to purify the active component and to test derivatives thereof. Methods of fractionation and purification of such heterogeneous extracts are known in the art. If desired, compounds shown to be useful agents for treatment are chemically modified according to methods known in the art. Compounds identified as being of therapeutic, prophylactic, diagnostic, or other value may be subsequently analyzed using HCC cell lines or a animal model for HCC.

Arrays, Microarrays, Libraries, Databases, And Kits

In one aspect, the invention provides nucleic acid or polypeptide arrays and biological assays thereof. Arrays refer to ordered arrangements of at least two nucleic acid molecules or polypeptides on a substrate, which can be any rigid or semi-rigid support to which two nucleic acid molecules or polypeptides may be attached. In some embodiments, a substrate may be a liquid medium. Substrates include membranes, filters, chips, slides, wafers, fibers, beads, gels, capillaries, plates, polymers, and microparticles etc. Because the nucleic acid molecules or polypeptides are located at specified locations on the substrate, the hybridization or binding patterns and intensities create a unique expression profile, which can be interpreted in terms of expression levels of particular genes and can be correlated with HCC progression, regression, therapy, etc.

High density nucleic acid or polypeptide arrays are also referred to as "microarrays," and may for example be used to monitor the presence or level of

expression of a large number of genes or polypeptides or for detecting sequence variations, mutations and polymorphisms. Arrays and microarrays generally require a solid support (for example, nylon, glass, ceramic, plastic, silica, aluminosilicates, borosilicates, metal oxides such as alumia and nickel oxide, various clays, nitrocellulose, etc.) to which the nucleic acid molecules or polypeptides are attached in a specified 2-dimensional arrangement, such that the pattern of hybridization or binding to a probe is easily determinable. In some embodiments, at least one of the nucleic acid molecules or polypeptides is a control, standard, or reference molecule, such as a housekeeping gene or portion thereof (e.g., PBGD, GAPDH), that may assist in the normalization of expression levels or assist in the determining of nucleic acid quality and binding characteristics; reagent quality and effectiveness; hybridization success; analysis thresholds and success, etc.

Nucleic acid molecules or polypeptide probes may be derived from compounds as described herein for example in Tables 1 through 4, and the compositions of the invention may be used as elements on a microarray to analyze gene expression profiles. For the purpose of such arrays, "nucleic acids" may include any polymer or oligomer of nucleosides or nucleotides (polynucleotides or oligonucleotides), which include pyrimidine and purine bases, preferably cytosine, thymine, and uracil, and adenine and guanine, respectively. A variety of methods are known for making and using microarrays, as for example disclosed in Cheung, V.G., et al. 1999; Lipshutz, R..J., et al. 1999; Bowtell, D.D.L., 1999; and, Schweitzer, B., 2002; G. MacBeath and S. L. Schreiber, 2000.; all of which are incorporated herein by reference. In some embodiments, the microarray substrate may be coated with a compound to enhance synthesis of the nucleic acid molecule on the substrate as disclosed in, for example, U.S. Pat. No. 4,458,066. In some embodiments, probes may be synthesized directly on the substrate in a predetermined ordered arrangement. Methods for storing, querying and analyzing microarray data have for example been disclosed in, for example, United States Patent No. 6,484,183; United States Patent No. 6,188,783; and Holloway, A.J., 2002; each of which is incorporated herein by reference. In an alternative aspect, the invention provides nucleic acid or polypeptide microarrays including a number of distinct and selected nucleic acid or polypeptide array sequences of the invention. The number of distinct sequences may for example

be any integer between 2 and 1 x 10^5, such as at least 10^2, 10^3, 10^4, or 10^5. The size of the distinct sequences may vary depending on the intended use, and can be determined by a skilled person. For example, the nucleic acid sequences may range from 15 to 5000 bases or more, or any integer between this range.

Microarrays may also be used to examine the expression of all the genes in a tissue or cell such as a liver cell or a HCC cell. Thus, the nucleic acid molecules of Tables 1 through 4 may be attached to a solid support, hybridized with single stranded detectably-labeled cDNAs (corresponding to a "complementary" orientation), and quantified using an appropriate method such that a signal is detected at each location at which hybridization has taken place. The intensity of the signal would then reflect the amount of gene expression. Similarly, protein microarrays may be used according to methods known in the art. Comparison of results from different cells or tissue, for example, hepatocellular carcinoma cells or tissue, hepatitis virus infected cells or tissue, non-tumor cells or tissue, normal cells or tissue, cirrhotic liver cells or tissue, or any combination thereof would elucidate differing levels of expression of specified genes from the different sources.

In one aspect of the invention, libraries may be constructed of bacterial strains each of which bears a plasmid expressing a different nucleic acid molecule of any one or more of Tables 1 through 4 under control of an inducible promoter. ORFs are amplified using PCR and cloned into a vector that enables their expression as N-terminal his-tagged polypeptides. These amplicons are also used to construct hybridization microarrays and enable targeted gene disruption, reducing expenses. A suitable expression host (e.g. *E. coli*) is selected, and genes encoding particular biochemical activities are identified by screening arrayed pools of his-tagged proteins as described previously (Martzen, M.R., et al., 1999).

The invention also provides databases including the nucleic acid and polypeptide sequences described herein, as well as gene expression information in various cancerous and non-cancerous liver and liver cell line samples. Such databases may be used to access information that may aid in diagnosis, prognosis, or other HCC-related methods of the invention. A database as used herein includes any electronic form of the compounds (e.g., nucleic acid and polypeptide sequences) of

the invention, and information regarding these compounds, and includes computer readable media and any suitable form for storing the information.

The invention also provides kits including for example one or more of the nucleic acid molecules or polypeptides of the invention (or complements, analogs, variants, or fragments thereof), an appropriate buffer, appropriate reagents for detection, and appropriate controls. For example, a kit may include probes or primers (which may or may not be detectably labeled) suitable for hybridization or amplification, or may include antibodies or ligands suitable for specific binding. A kit may also include written or electronic instructions.

Diagnostic and Other Uses

A wide variety of detectable labels and conjugation techniques are known by those skilled in the art and may be used in various nucleic acid molecule and polypeptide assays to diagnose HCC. The nucleic acid molecules, proteins, antibodies and other compounds according to the invention may be labeled for purposes of assay by joining them, either covalently or noncovalently, with a detectable label. By "detectably labeled" is meant any means for marking and identifying the presence of a molecule, e.g., an oligonucleotide probe or primer, a gene or fragment thereof, a cDNA molecule, or a polypeptide. Methods for detectably-labeling a molecule are well known in the art and include, without limitation, radioactive labeling (e.g., with an isotope such as ^{32}P or ^{35}S) and nonradioactive labelling such as, enzymatic labeling (for example, using horseradish peroxidase or alkaline phosphatase), chemiluminescent labeling, fluorescent labeling (for example, using fluorescein), bioluminescent labeling, or antibody detection of a ligand attached to the probe. Also included in this definition is a molecule that is detectably labeled by an indirect means, for example, a molecule that is bound with a first moiety (such as biotin) that is, in turn, bound to a second moiety that may be observed or assayed (such as fluorescein-labeled streptavidin). Labels also include digoxigenin, luciferases, and aequorin. Synthesis of labeled molecules performed by using labels such as ^{32}P-dCTP, Cy3-dCTP or Cy5-dCTP or ^{35}S-methionine. Compounds according to the invention may also be directly labeled by chemical

conjugation to amines, thiols and other groups present in the molecules using reagents such as BIODIPY or FITC (Molecular Probes, Eugene, OR, USA).

Compounds, compositions, and reagents according to the invention may be used to detect and quantify differential gene expression; absence, presence, or excess expression of nucleic acid molecules (e.g., mRNAs) or polypeptides; or to monitor nucleic acid molecule (e.g., mRNA) or polypeptide levels during therapeutic intervention in subjects with HCC. The compounds, compositions, and reagents according to the invention can also be utilized as markers of HCC treatment efficacy over a period ranging from days to months to years. The diagnostic assays may use hybridization, amplification, ligand binding, or antibody technologies to compare gene expression in a biological sample from a subject to reference samples or standards, or to cancerous and non-cancerous samples from the subject, in order to detect altered gene expression. Qualitative or quantitative methods for this comparison are known in the art, and any suitable method may be used.

In order to provide a basis for the diagnosis of HCC, a non-tumor or standard gene expression profile may be established. This may be accomplished by combining a biological sample taken from normal or non-tumor subjects or from non-cancerous tissue from a subject with HCC, with a probe under conditions for hybridization or amplification. Standard hybridization may be quantified by comparing the values obtained using non-tumor subjects or non-cancerous tissue with values from an experiment in which a known amount of a substantially purified target sequence is used. Standard values obtained in this manner may be compared with values obtained from samples from patients who are symptomatic for HCC. Deviation from standard values toward those associated with HCC is used to diagnose HCC. Such assays may also be used to monitor the efficacy of a particular HCC therapy in animal studies, in clinical trials, or to monitor the treatment of an individual patient or groups of patients. Once the presence of HCC is established in a subject and a treatment protocol is initiated, assays according to the invention may be repeated on a regular basis to determine if the level of expression in the subject begins to approximate that which is observed in a non-tumor subject, and to monitor the progression of HCC in the subject. The results obtained from successive assays may be used to show the efficacy of treatment over a period ranging from several days to months.

Compounds, compositions, and reagents (e.g., microarrays) according to the invention may be used to monitor the progression or regression of HCC. The differences in gene expression between healthy and diseased tissues or cells can be assessed and cataloged. By analyzing changes in patterns of gene expression, HCC can be diagnosed at earlier stages before the subject is symptomatic. Similarly, by analyzing gene expression profiles and changes therein, prognoses may be formulation, and therapies may be designed. Progression or regression of HCC may be determined by comparison of two or more different HCC samples taken at multiple different times from a subject (e.g., at least 2, 3, 4, or 5 or more time points) over the course of days to months. For example, progression or regression may be evaluated by assessments of expression of sets of two or more, or as many as all, of the nucleic acid molecules of Tables 1 through 4 in a HCC tissue sample from a subject before, during, and following treatment for HCC.

Compounds, compositions, and reagents (e.g., microarrays) according to the invention can also be used to monitor the efficacy of a therapy. For therapies with known side effects, compounds, compositions, and reagents (e.g., microarrays) according to the invention may be employed to improve the therapeutic regimen. For example, dosages that causes changes in gene profiles that represent efficacious treatment may be determined, and expression profiles associated with the onset of undesirable side effects may be avoided. This approach may be more sensitive and rapid than waiting for the subject to show inadequate improvement, or to manifest side effects, before altering the course of treatment. In another aspect of the invention, pre- and post-treatment alterations in expression of two or more sets of HCC nucleic acid molecules in HCC cells or tissues may be used to assess treatment parameters including, but not limited to: dosage, method of administration, timing of administration, and combination with other known treatments for HCC.

In some aspects, any one or more of the compounds provided herein may be used in therapeutic applications. For example, selected compounds provided herein may be used as therapeutic targets for the identification of agents, that modulate their expression levels and/or activity, that may be used to treat HCC.

EXAMPLES

Experimental Procedures

RNA isolation, RNA amplification and cDNA microarray hybridization

Paired samples of tumor and corresponding non-tumor tissues were obtained from resected liver specimens from thirty-seven (37) patients who had been diagnosed with hepatitis B virus (HBV)-associated HCC and had undergone curative liver resection. A validation tissue set composed of 58 liver biopsy samples from an independent cohort of twenty-nine (29) patients, who also had HBV-associated HCC and had undergone liver resection, was used. Informed consent from the patient and institutional research and ethics committee approval were obtained. Tissues were snap frozen in liquid nitrogen and stored at -150°C. A small section of each specimen was sampled and total RNA was isolated from tissues using TRIZOL® reagent (Life Technologies, Bethesda, MD, USA) according to the manufacturer's instructions. The integrity of the RNA specimen was verified by gel electrophoresis.

The human liver cancer cell lines used in this study were: PLC/PRF/5, HA22T,Huh1, Huh4, Tong, Hep3B, SNU182, SNU449, SNU475, HepG2, Huh6, Huh7, SKHep1, and Mahlavu. All cell lines were cultured under conditions recommended by the American Type Culture Collection (VA, USA). Total RNA was extracted using TRIZOL® reagent (Life Technologies, Bethesda, MD, USA) according to the manufacturer's instructions.

RNA was linearly amplified using a procedure modified from Eberwine and coworkers (Eberwine et al, 1992). Briefly, total RNA was reverse transcribed using a 63-nucleotide synthetic primer containing the T7 polymerase binding site 5'-GGCCAGTGAATTGTAATACGACTCACTATAGGGAGGCGG$(T)_{24}$-3'. Full-length double-stranded cDNA synthesis was accomplished in the presence of *E. coli* DNA polymerase I, DNA ligase and RNase H. The cDNA was made blunt-ended with T4 DNA polymerase, and purified by extraction in a mixture of phenol, chloroform and isoamyl alcohol, and precipitation in the presence of ammonium acetate and ethanol. Purified double-stranded cDNA was then transcribed with T7 polymerase (T7 Megascript® Kit, Ambion) to yield linearly amplified antisense RNA, which was subsequently purified with RNeasy® mini-columns (Qiagen). Human universal reference RNA (Stratagene, La Jolla, CA), including total RNA from 10

different human cell lines, was amplified and used as the reference for cDNA microarray analysis.

Approximately 9000 human cDNA features (Incyte Genomics, Palo Alto, CA, USA) were spotted onto poly-L-lysine coated slides using OmniGrid® arrayer (GeneMachines). Probes were generated from the amplified RNA material and hybridized to the chip as described elsewhere (Sotiriou et al, 2002). Briefly, 4 μg amplified RNA was reverse-transcribed using random hexamers and directly labeled with Cy3-conjugated dUTP (reference RNA) or Cy5-conjugated dUTP (sample RNA). Hybridization was performed in the presence of 25% formamide and 5X SSC for 16h at 42°C. Slides were scanned with an Axon 4000b laser scanner (Axon Instruments) after washing and drying. To minimize the effects of labeling biases, reciprocal dye swap labeling experiments were performed for each sample.

Data analysis

The 37 paired HCC tumor and non-tumor liver samples, and liver cancer cell lines were processed on the microarray on two separate prints, and the validation tissue set was processed on a third print. Raw data was analyzed on GenePix analysis software version 3.0 (Axon Instruments, Burlingame, CA, USA) and uploaded to a relational database maintained by the Center for Information Technology at the National Institutes of Health (ie. MADB). The cDNA clones used for the microarray are represented by their UniGene identifiers. For each array, the logarithmic expression ratio for a spot on each array was normalized by subtracting the median logarithmic ratio for the same array. Data was filtered to exclude spots with a size of less than 25 μm and any poor quality or missing spots. Since the correlation of the overall data from reciprocal labeling was good, values obtained from reciprocal labeling experiments were averaged. In addition, any gene features that were found to be absent from the data in more than 50% of patient samples in either set of arrays were excluded, and gene features that were common in data from the array print sets were retained. Application of these filters resulted in the inclusion of 8716 of the total 9127 features in subsequent analysis. Statistical comparison of genes between HCC tumors and non-tumors was performed by the Wilcoxon rank-sum non-parametric test. To evaluate gene expression patterns, hierarchical clustering

using one minus Pearson's correlation metric and average linkage (Eisen et al, 1998) and multidimensional scaling was performed on normalized data (mean equals zero, standard deviation equals one). Functional characterization of genes was based on Gene Ontology (The Gene Ontology Consortium, 2000) and other published works known to those of ordinary skill in the art.

The quality of a set of selected gene features to be used as potential markers was measured by estimating the probability that its observed performance, in terms of number of misclassified tissue samples, could occur by chance alone. This was achieved by performing a series of Monte Carlo simulations (Davison and Hinkley, 1997) upon the expression data of the selected genes. In each simulation, the tissues' labels were randomly permuted and the number of misclassifications was noted. A total of one million runs of Monte Carlo simulations were performed. The reported P-value (denoted as P_a) is the fraction of times the permutations generated as few misclassifications as, or fewer than, the original labeling. To determine whether the set of genes observed to have a good performance as tumor discriminators, could appear merely by chance, different Monte Carlo simulations were carried out. In each simulation, an equivalent number of gene features was randomly picked from a designated large population of features, and the performance of the random gene set was evaluated by the number of tissue samples that were misclassified. A total of 10,000 runs of Monte Carlo simulations were performed for each evaluation. The P-value (P_b) is the fraction of times the random gene set performed as good as, or better than, the performance of the selected gene set.

The significance of the number of observed overlapping genes after intersection of the important gene lists, derived as described herein, with gene lists reported previously was approximated by measuring the probabilities that such overlap could occur by chance alone. A separate series of Monte Carlo simulations were employed to estimate the P-values of the two-group comparisons. In each simulation, two lists of genes corresponding to the two groups were generated. Each list was constructed by randomly selecting genes, as many as the number of genes in its corresponding group, from the entire collection of genes of its respective microarray gene set. The two random gene lists were then intersected. The P-value (P_c) of the comparison was obtained by generating and intersecting the

two random lists 100 million times, and reported as the fraction of times the random overlap is equal or greater than the observed one. To validate the utility of the various expression cassettes to distinguish HCC tumor from non-tumor liver, the prediction accuracy of each discriminator cassette was assessed on an independent tissue set comprising of 58 liver clinical biopsies from 29 patients using a *k*-Nearest Neighbor (*k*NN) classification algorithm (*k*=3) using Pearson correlation to measure the similarity between expression profiles. The algorithm was trained against the dataset comprising 74 tissue samples from 37 patients before testing against the new tissue set.

Real-time semi-quantitative RT-PCR

Total RNA from individual tissue samples were analyzed for the expression levels of selected genes by real-time semi-quantitative RT-PCR using the LightCycler RNA amplification kit SYBR Green I on the LightCycler (Roche, Basel, Switzerland) according to the manufacturer's instructions. Briefly, one-step RT-PCR reactions consisted of an initial incubation at 55°C for 10 min, followed by a denaturation step at 95°C for 30 s, and amplification for 40 cycles of 1 s at 95°C, 10 s at 57°C, and 13 s at 72°C. For each reaction, 10 ng of total RNA was analyzed. The gene specific primers designed were, for example, as follows: IGFBP3 5'-ATAATCATCATCAAGAAAGGGCAT-3' and 5'-GAAGGGCGACACTGCTT- 3'; EGFR 5'-GCGTCTCTTGCCGGAATG-3' and 5'-GGCTCACCCTCCAGAAGCTT-3'; ERBB2 5'-GGATGTGCGGCTCGTACAC-3' and 5'-TAATTTTGACATGGTTGGGACTCTT- 3'; ERBB3 5'-CGGTTATGTCATGCCAGATACAC-3' and 5'-ACAGAACTGAGACCCACTGAAGAA-3'; PBGD 5'-GAGTGATTCGCGTGGGTACC- 3' and 5'-GGCTCCGATGGTGAAGCC-3'. The relative expression level of each gene of interest in individual tissue sample was normalized against that of the "housekeeping" gene PBGD. Data are presented as the level of gene expression in each HCC tumor relative to its corresponding non-tumor liver specimen.

Assessment Of Global Gene Expression Differences Between HCC Tumors And Non-Tumor Liver Specimens

The gene expression patterns of primary HCC tumors and the corresponding non-tumor liver tissues from 37 patients were examined by cDNA microarray. Amplified RNA prepared from each experimental sample was labeled with Cy5 and hybridized on the array with pooled human 'common reference' amplified RNA labeled with Cy3. Reciprocal dye swap replicate hybridizations were performed to minimize technical noise. Since the overall correlation of reciprocal labeling was good, values obtained from reciprocal labeling experiments were averaged and used in subsequent analyses. Firstly, the overall natural patterns of gene expression in the HCC tumor and non-tumor liver tissues were assessed based on unsupervised hierarchical clustering. Analysis of variance in expression levels for each gene across all the tissues indicated that 500 gene features (containing 493 unique UniGenes) showed the largest variability across both HCC tumor and non-tumor liver tissues (Figure 1). Included in this list are AFP, an often used prognostic marker for HCC, and other genes associated with HCC such as HGF, MYC, and a ras family member RAN. Hierarchical clustering analysis based on these highly variant genes (derived from the 37 pairs of HCC tumor and non-tumor liver samples and using the 500 most variable gene features) separated the tissues into two main clusters, one representing the HCC tumors and the other, the non-tumor liver tissues with only six of 37 HCC tumors misclassified as non-tumors. Thus, the molecular configuration of HCC can be readily distinguished from that of non-tumor liver with minimal data manipulation.

Next, to investigate differential gene expression patterns between HCC tumors and non-tumor livers, the Wilcoxon rank-sum test was used and the top 2.5% candidate genes which displayed the smallest (best) P-value scores ($P<1x10^{-6}$) and at least 1.5-fold change in gene expression were identified, resulting in a list of 218 genes (Table 1). For these 218 genes, false discovery rate analysis indicated a false-positive error of less than 0.4%. Multidimensional scaling analysis based on these outlier genes indicated that the HCC tumors were a more heterogeneous population than the non-tumor liver tissues (Figure 2).

Cancer cell lines derived from the primary tumor have traditionally been used as *in vitro* model systems for investigating the function of genes in the *in vivo* tumor environment. Using the 218 differentially expressed outlier genes identified in the clinical samples, the expression pattern of the same genes in 14 established human liver cancer cell lines was analysed. These cell lines exhibited gene expression profiles that were different from the clinical HCC tumor tissues (Figure 2), suggesting that they may have accumulated additional genetic or epigenetic alterations in culture and are not entirely reflective of the primary tumor biology.

Identification Of Gene Clusters Differentially Expressed In HCC Tumors

Among the statistically significant 218 genes that distinguished HCC tumors from non-tumor liver tissue specimens, more genes were observed to be overexpressed than under-expressed in the malignant tissue specimens relative to the non-tumor tissue specimens. Mapping of the chromosomal location of these 218 unique outlier genes indicated that a disproportionate number of genes was located on chromosome 1 (Figure 3A), particularly in the 1q region, and that majority of these genes were more highly expressed in the tumor tissues. Further characterization of these outlier genes revealed that a substantial proportion of genes was involved in transport (*e.g.*, PEA15), RNA processing (*e.g.*, RDBP), and metabolic processes (*e.g.*, NME1) and showed increased expression in HCC tumor specimens, possibly indicating accelerated rates of metabolism (Figure 3B, Table 1). Several outlier genes (*e.g.*, SMT3H1) are members of the ubiquitin-proteasome pathway, suggesting deregulation of this pathway in HCC. A gene cluster associated with lymphocyte infiltrate that included the expression of genes such as IGKC and IGJ was observed, and transcription factors (*e.g.*, ESR1) and genes involved in controlling growth and differentiation (*e.g.*, GRN) , and signal transduction (*e.g.*, CSTB) formed the other dominant gene groups. Notably, the polycomb group protein BMI1 was consistently expressed at much higher levels in HCC tumor specimens (Figure 4).

Table 1. Genes significantly differentially expressed between HCC tumor and non-tumor liver tissues.

Function	Gene Symbol	Gene Name	UniGene Identifier	Expression change in HCC tumor*	GenBank No.
transcription factors	ILF2	interleukin enhancer binding factor 2, 45kD	Hs.75117	?	AA307289
	BMI1	murine leukemia viral (bmi-1) oncogene homolog	Hs.431	?	AA884913
	TAF9	TAF9 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 32 kD	Hs.60679	?	U21858
	RFX5	regulatory factor X, 5 (influences HLA class II expression)	Hs.166891	?	AL050135
	SSRP1	structure specific recognition protein 1	Hs.79162	?	AI635077
	ZNF146	zinc finger protein 146	Hs.301819	?	X70394
	SREBF2	sterol regulatory element binding transcription factor 2	Hs.108689	?	AA608556
	MAFG	v-maf musculoaponeurotic fibrosarcoma (avian) oncogene family, protein G	Hs.252229	?	AF059195
	CHD4	chromodomain helicase DNA binding protein 4	Hs.74441	?	BE408958
	NR4A1	nuclear receptor subfamily 4, group A, member 1	Hs.1119	?	NM_002135
	ESR1	estrogen receptor 1	Hs.1657	?	AL078582
	ZNF238	zinc finger protein 238	Hs.69997	?	AJ223321
	FOSB	FBJ murine osteosarcoma viral oncogene homolog B	Hs.75678	?	L49169
	ID1	inhibitor of DNA binding 1, dominant negative helix-loop-helix protein	Hs.75424	?	S78825
	FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	Hs.25647	?	V01512
RNA processing	H2AFY	H2A histone family, member Y	Hs.75258	?	AA307460
	SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	Hs.83753	?	BE252108
	RPS7	ribosomal protein S7	Hs.301547	?	AA315872
	MRPS14	mitochondrial ribosomal protein S14	Hs.247324	?	AW973521
	HNRPU	heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A)	Hs.103804	?	X65488
	SNRPD2	small nuclear	Hs.53125	?	AA315774

		ribonucleoprotein D2 polypeptide (16.5kD)			
	NCL	nucleolin	Hs.79110	?	AK000250
	RPS10	ribosomal protein S10	Hs.76230	?	AW245775
	RPL6	ribosomal protein L6	Hs.349961	?	AW675430
	SFPQ	splicing factor proline/glutamine rich (polypyrimidine tract-binding protein-associated)	Hs.180610	?	X70944
	DIM1	similar to S. pombe dim1+	Hs.5074	?	AI814618
	MARS	methionine-tRNA synthetase	Hs.279946	?	BE299937
	SFRS9	splicing factor, arginine/serine-rich 9	Hs.77608	?	AL021546
	RBM3	RNA binding motif protein 3	Hs.301404	?	NM_006743
	U2AF65	U2 small nuclear ribonucleoprotein auxiliary factor (65kD)	Hs.7655	?	AA936430
	SFRS1	splicing factor, arginine/serine-rich 1 (splicing factor 2, alternate splicing factor)	Hs.73737	?	M72709
	SNRPE	small nuclear ribonucleoprotein polypeptide E	Hs.334612	?	X12466
	SF3B4	splicing factor 3b, subunit 4, 49kD	Hs.25797	?	NM_005850
	RDBP	RD RNA-binding protein	Hs.106061	?	X16105
	SNRPF	small nuclear ribonucleoprotein polypeptide F	Hs.105465	?	AA649986
	RRM1	ribonucleotide reductase M1 polypeptide	Hs.2934	?	X59543
	RPL38	ribosomal protein L38	Hs.2017	?	AI832988
	HNRPH1	heterogeneous nuclear ribonucleoprotein H1 (H)	Hs.245710	?	BE296051
	U5-116KD	U5 snRNP-specific protein, 116 kD	Hs.151787	?	D21163
	RPLP1	ribosomal protein, large, P1	Hs.177592	?	AW963733
	OXA1L	oxidase (cytochrome c) assembly 1-like	Hs.151134	?	X80695
DNA replication/ repair	ADPRT	ADP-ribosyltransferase (NAD+; poly (ADP-ribose) polymerase)	Hs.177766	?	M18112
	PRKDC	protein kinase, DNA-activated, catalytic polypeptide	Hs.155637	?	U34994
	SMC4L1	SMC4 (structural maintenance of chromosomes 4, yeast)-like 1	Hs.50758	?	AB019987
	H2AV	histone H2A.F/Z variant	Hs.301005	?	BE409809
	FEN1	flap structure-specific endonuclease 1	Hs.4756	?	BE278623

	MCM2	minichromosome maintenance deficient (S. cerevisiae) 2 (mitotin)	Hs.57101	?	BE250461
	HAT1	histone acetyltransferase 1	Hs.13340	?	AF030424
	RAD50	RAD50 (S. cerevisiae) homolog	Hs.41587	?	Z75311
	CBX1	chromobox homolog 1 (HP1 beta homolog Drosophila)	Hs.77254	?	AL046741
	CSPG6	chondroitin sulfate proteoglycan 6 (bamacan)	Hs.24485	?	NM_005445
	FUS	fusion, derived from t(12;16) malignant liposarcoma	Hs.99969	?	BE396632
	UNG2	uracil-DNA glycosylase 2	Hs.3041	?	AA291356
cell cycle/ growth/ differentia-tion	GPC3	glypican 3	Hs.119651	?	U50410
	CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	Hs.1174	?	AI859822
	MDK	midkine (neurite growth-promoting factor 2)	Hs.82045	?	AA427949
	NTRK1	neurotrophic tyrosine kinase, receptor, type 1	Hs.85844	?	AA075110
	CCNE2	cyclin E2	Hs.30464	?	NM_004702
	HDGF	hepatoma-derived growth factor (high-mobility group protein 1-like)	Hs.89525	?	BE259164
	TP53BP2	tumor protein p53-binding protein, 2	Hs.44585	?	AI123916
	CDC23	CDC23 (cell division cycle 23, yeast, homolog)	Hs.153546	?	AF053977
	GRN	granulin	Hs.180577	?	AI375908
	GHR	growth hormone receptor	Hs.125180	?	X06562
	IGFBP3	insulin-like growth factor binding protein 3	Hs.77326	?	BE336944
	CYR61	cysteine-rich, angiogenic inducer, 61	Hs.8867	?	Y12084
	HGF	hepatocyte growth factor (hepapoietin A; scatter factor)	Hs.809	?	X16323
apoptosis	DAP3	death associated protein 3	Hs.159627	?	AA207194
	PDCD5	programmed cell death 5	Hs.166468	?	AA452724
immune response	PPIA	peptidylprolyl isomerase A (cyclophilin A)	Hs.342389	?	AW732921
	TMPO	thymopoietin	Hs.11355	?	U09087
	PPIB	peptidylprolyl isomerase B (cyclophilin B)	Hs.699	?	BE386706
	CD5L	CD5 antigen-like (scavenger receptor cysteine rich family)	Hs.52002	?	NM_005894
	SCYA14	small inducible cytokine subfamily A (Cys-Cys),	Hs.20144	?	NM_004166

		member 14			
	SDF1	stromal cell-derived factor 1	Hs.237356	?	L36033
	IGHG3	immunoglobulin heavy constant gamma 3 (G3m marker)	Hs.300697	?	D78345
	C7	complement component 7	Hs.78065	?	X86328
	IGJ	immunoglobulin J polypeptide, linker protein for immunoglobulin alpha and mu polypeptides	Hs.76325	?	AW172754
	IGKC	immunoglobulin kappa constant	Hs.156110	?	AW404507
cell adhesion/ cytoskeletal organization	LBR	lamin B receptor	Hs.152931	?	L25931
	ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	Hs.287797	?	W38716
	LAMR1	laminin receptor 1 (67kD, ribosomal protein SA)	Hs.181357	?	AW328280
	CAPZA2	capping protein (actin filament) muscle Z-line, alpha 2	Hs.75546	?	U03851
	ICAP-1A	integrin cytoplasmic domain-associated protein 1	Hs.173274	?	AF012023
	DNCH1	dynein, cytoplasmic, heavy polypeptide 1	Hs.7720	?	AB002323
	ARPC1A	actin related protein 2/3 complex, subunit 1A (41 kD)	Hs.90370	?	Y08999
	DPT	dermatopontin	Hs.80552	?	AW016451
	MMP15	matrix metalloproteinase 15 (membrane-inserted)	Hs.80343	?	D85510
	ARHE	ras homolog gene family, member E	Hs.6838	?	W03441
signal transduction	CAP2	adenylyl cyclase-associated protein 2	Hs.296341	?	AW779995
	CSTB	cystatin B (stefin B)	Hs.695	?	AI831499
	ARMET	arginine-rich, mutated in early stage tumors	Hs.75412	?	AA582041
	EFNA1	ephrin-A1	Hs.1624	?	NM_004428
	PPP2R5A	protein phosphatase 2, regulatory subunit B (B56), alpha isoform	Hs.155079	?	AA234460
	RAN	RAN, member RAS oncogene family	Hs.10842	?	NM_006325
	CALM2	calmodulin 2 (phosphorylase kinase, delta)	Hs.182278	?	D45887
	LASP1	LIM and SH3 protein 1	Hs.334851	?	AI304506
	SHC1	SHC (Src homology 2 domain-containing) transforming protein 1	Hs.81972	?	X68148
	RGS5	regulator of G-protein	Hs.24950	?	AI674877

		signalling 5			
	HAX1	HS1 binding protein	Hs.15318	?	BE260953
	GABRE	gamma-aminobutyric acid (GABA) A receptor, epsilon	Hs.22785	?	NM_004961
	ARFGEF2	ADP-ribosylation factor guanine nucleotide-exchange factor 2 (brefeldin A-inhibited)	Hs.118249	?	AA099582
	MAPK6	mitogen-activated protein kinase 6	Hs.271980	?	NM_002748
	GNB2L1	guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1	Hs.5662	?	BE206815
	ERBB3	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 3	Hs.199067	?	AI565773
	DSCR1	Down syndrome critical region gene 1	Hs.184222	?	U85267
	CRHBP	corticotropin releasing hormone-binding protein	Hs.115617	?	NM_001882
	STK39	serine threonine kinase 39 (STE20/SPS1 homolog, yeast)	Hs.199263	?	F26137
ubiquitin-proteasome pathway	UBD	diubiquitin	Hs.44532	?	NM_006398
	PSMB4	proteasome (prosome, macropain) subunit, beta type, 4	Hs.89545	?	BE336637
	SSA2	Sjogren syndrome antigen A2 (60kD, ribonucleoprotein autoantigen SS-A/Ro)	Hs.554	?	NM_004600
	USP14	ubiquitin specific protease 14 (tRNA-guanine transglycosylase)	Hs.75981	?	NM_005151
	PSMA1	proteasome (prosome, macropain) subunit, alpha type, 1	Hs.82159	?	AI889267
	EIF3S9	eukaryotic translation initiation factor 3, subunit 9 (eta, 116kD)	Hs.57783	?	U62583
	SMT3H1	SMT3 (suppressor of mif two 3, yeast) homolog 1	Hs.85119	?	AA160893
	UBE2D2	ubiquitin-conjugating enzyme E2D 2 (homologous to yeast UBC4/5)	Hs.108332	?	NM_003339
	PSMD11	proteasome (prosome, macropain) 26S subunit, non-ATPase, 11	Hs.90744	?	AB003102
	PSMB3	proteasome (prosome, macropain) subunit, beta type, 3	Hs.82793	?	AI028114
	PSMD4	proteasome (prosome, macropain) 26S subunit, non-ATPase, 4	Hs.148495	?	AA604027
molecular	CCT5	chaperonin containing	Hs.1600	?	D43950

chaperone		TCP1, subunit 5 (epsilon)			
	CCT3	chaperonin containing TCP1, subunit 3 (gamma)	Hs.1708	?	BE302501
	HSPA5	heat shock 70kD protein 5 (glucose-regulated protein, 78kD)	Hs.75410	?	AL043206
	CCT4	chaperonin containing TCP1, subunit 4 (delta)	Hs.79150	?	U38846
	HSPA4	heat shock 70kD protein 4	Hs.90093	?	AB023420
	CCT7	chaperonin containing TCP1, subunit 7 (eta)	Hs.108809	?	AA314436
	HSPA8	heat shock 70kD protein 8	Hs.180414	?	AW249010
	CCT6A	chaperonin containing TCP1, subunit 6A (zeta 1)	Hs.82916	?	L27706
transport	ANXA2	annexin A2	Hs.217493	?	BE293414
	PDZK1	PDZ domain containing 1	Hs.15456	?	AF012281
	SYPL	synaptophysin-like protein	Hs.80919	?	S72481
	TIMM17A	translocase of inner mitochondrial membrane 17 homolog A (yeast)	Hs.20716	?	AW247564
	XPO1	exportin 1 (CRM1, yeast, homolog)	Hs.79090	?	D89729
	HMGN4	high mobility group nucleosomal binding domain 4	Hs.236774	?	U90549
	NUCB2	nucleobindin 2	Hs.3164	?	AW951523
	UGTREL 1	UDP-galactose transporter related	Hs.154073	?	AW192554
	PEA15	phosphoprotein enriched in astrocytes 15	Hs.194673	?	Y13736
	CLTA	clathrin, light polypeptide (Lca)	Hs.104143	?	AW974204
	ATP6IP1	ATPase, H+ transporting, lysosomal interacting protein 1	Hs.6551	?	NM_001183
	SSR2	signal sequence receptor, beta (translocon-associated protein beta)	Hs.74564	?	BE313059
	AP3S1	adaptor-related protein complex 3, sigma 1 subunit	Hs.80917	?	D63643
	VDAC2	voltage-dependent anion channel 2	Hs.78902	?	AI015604
	VPS45A	vacuolar protein sorting 45A (yeast)	Hs.6650	?	AA702845
	VCP	valosin-containing protein	Hs.106357	?	NM_007126
	SACM2L	SAC2 (suppressor of actin mutations 2, yeast, homolog)-like	Hs.169407	?	AK001725
	KPNB1	karyopherin (importin) beta 1	Hs.180446	?	L38951
	SLC21A3	solute carrier family 21 (organic anion transporter), member 3	Hs.46440	?	U21943
	SLC22A1	solute carrier family 22 (organic cation transporter),	Hs.117367	?	X98332

		member 1			
	HSPA5	Heat shock 70kD protein 5 (glucose-regulated protein, 78kD)	Hs. 75410	?	
metabolism	GNPAT	glyceronephosphate O-acyltransferase	Hs.12482	?	AF043937
	NME2	non-metastatic cells 2, protein (NM23B) expressed in	Hs.275163	?	L16785
	NME1	non-metastatic cells 1, protein (NM23A) expressed in	Hs.118638	?	AA147871
	UQCRH	ubiquinol-cytochrome c reductase hinge protein	Hs.73818	?	AI093521
	TALDO1	transaldolase 1	Hs.77290	?	AF010400
	P5CR2	pyrroline 5-carboxylate reductase isoform	Hs.274287	?	AI161110
	GFPT1	glutamine-fructose-6-phosphate transaminase 1	Hs.1674	?	NM_002056
	DPM1	dolichyl-phosphate mannosyltransferase polypeptide 1, catalytic subunit	Hs.5085	?	AW173486
	ACLY	ATP citrate lyase	Hs.174140	?	AW967351
	B4GALT3	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 3	Hs.321231	?	Y12509
	GCN1L1	GCN1 (general control of amino-acid synthesis 1, yeast)-like 1	Hs.75354	?	D86973
	DPAGT1	dolichyl-phosphate (UDP-N-acetylglucosamine) N-acetylglucosaminephosphotransferase 1 (GlcNAc-1-P transferase)	Hs.26433	?	Z82022
	ACAA1	acetyl-Coenzyme A acyltransferase 1 (peroxisomal 3-oxoacyl-Coenzyme A thiolase)	Hs.166160	?	NM_001607
	ALDH8A1	aldehyde dehydrogenase 8 family, member A1	Hs.18443	?	AI051566
	SRD5A2	steroid-5-alpha-reductase, alpha polypeptide 2	Hs.1989	?	M74047
	NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)	Hs.2	?	D90040
	GSTZ1	glutathione transferase zeta 1 (maleylacetoacetate isomerase)	Hs.26403	?	U86529
	ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide	Hs.4	?	M24317
	CYP2C8	cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase), polypeptide 8	Hs.174220	?	M17398

	CYP2E	cytochrome P450, subfamily IIE (ethanol-inducible)	Hs.75183	?	J02843
Unknown		ESTs, Highly similar to H33_HUMAN HISTONE H3.3 [H.sapiens]	Hs.349754	?	AA313375
	ECT2	epithelial cell transforming sequence 2 oncogene	Hs.122579	?	AL137710
	DKFZP564B167	DKFZP564B167 protein	Hs.76285	?	AI032331
	KIAA0016	translocase of outer mitochondrial membrane 20 (yeast) homolog	Hs.75187	?	D13641
	DXS1357E	accessory proteins BAP31/BAP29	Hs.291904	?	Z31696
	KIAA0475	KIAA0475 gene product	Hs.5737	?	AA524523
	C20orf24	chromosome 20 open reading frame 24	Hs.184062	?	AI340141
	FLJ10326	hypothetical protein FLJ10326	Hs.262823	?	AA665998
	KIAA0117	KIAA0117 protein	Hs.322478	?	AL133010
		Unknown		?	NM_002211
	DEK	DEK oncogene (DNA binding)	Hs.110713	?	AI888504
	PODXL	podocalyxin-like	Hs.16426	?	BE395330
	DSS1	Deleted in split-hand/split-foot 1 region	Hs.333495	?	W79057
	PRO1855	hypothetical protein PRO1855	Hs.283558	?	AI379021
		Homo sapiens mRNA; cDNA DKFZp434I052 (from clone DKFZp434I052)	Hs.378917	?	AA425759
	KIAA0470	KIAA0470 gene product	Hs.25132	?	NM_014812
	MYLE	MYLE protein	Hs.11902	?	AA628977
		Homo sapiens cDNA FLJ14232 fis, clone NT2RP4000035	Hs.101810	?	AI675122
	MAGED2	melanoma antigen, family D, 2	Hs.4943	?	Z98046
	FLJ12806	hypothetical protein FLJ12806	Hs.107637	?	BE044582
	YWHAB	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide	Hs.279920	?	AL008725
		Unknown		?	AL031775
	LOC51235	hypothetical protein	Hs.181444	?	AI190653
	KIAA0592	KIAA0592 protein	Hs.13273	?	AL080183
	KIAA0205	KIAA0205 gene product	Hs.3610	?	D86960
	C1orf9	chromosome 1 open reading frame 9	Hs.108636	?	BE466870
	KIAA0788	KIAA0788 protein	Hs.246112	?	AB018331
	MGC1955	hypothetical protein	Hs.334787	?	BE379431

	6	MGC19556			
	KIAA0731	KIAA0731 protein	Hs.6214	?	AB018274
	C7orf14	chromosome 7 open reading frame 14	Hs.84790	?	D86978
	D123	D123 gene product	Hs.82043	?	U27112
	C5orf8	chromosome 5 open reading frame 8	Hs.75864	?	BE254013
		Homo sapiens cDNA: FLJ23020 fis, clone LNG00943	Hs.6127	?	AA054768
	AD24	AD24 protein	Hs.74899	?	AI017605
	WHIP	Werner helicase interacting protein	Hs.236828	?	AA481600
	BC-2	putative breast adenocarcinoma marker (32kD)	Hs.12107	?	AF042384
	DKFZP547E101	DKFZP547E1010 protein	Hs.323817	?	NM_015607
	FLJ22251	hypothetical protein FLJ22251	Hs.289064	?	AA595663
		ESTs	Hs.89267	?	AA284067
	KIAA0187	KIAA0187 gene product	Hs.10848	?	D80009
	MPV17	MpV17 transgene, murine homolog, glomerulosclerosis	Hs.75659	?	NM_002437
	MAWBP	MAWD binding protein	Hs.16341	?	AI866254
		Homo sapiens cDNA FLJ37464 fis, clone BRAWH2011795, weakly similar to LIVER CARBOXYLESTERASE PRECURSOR (EC 3.1.1.1)	Hs.346947	?	N44535
		Homo sapiens SNC73 protein (SNC73) mRNA, complete cds	Hs.293441	?	AA290845
		ESTs, Highly similar to SMHU1B metallothionein 1B [H.sapiens]	Hs.36102	?	R99207
		Homo sapiens unknown mRNA	Hs.367982	?	H72532
	FLJ12666	hypothetical protein FLJ12666	Hs.23767	?	AW952494
	RNAHP	RNA helicase-related protein	Hs.8765	?	AI814448

*gene expression level showing at least 1.5-fold change in HCC tumors relative to non-tumor liver tissues ($P<1x10^{-6}$)

Real-time RT-PCR analysis was performed on a panel of genes, including IGFBP3 and ERBB3 in all the 37 matched HCC tumor and non-tumor liver samples. Expression of a known "housekeeping" gene porphobilinogen deaminase (PBGD) (Fink et al, 1998) was used as normalizing control. The results of real-time RT-PCR analyses of IGFBP3 and ERBB3 indicated that IGFBP3 expression was diminished in

35 of 37 HCC tumors relative to their corresponding non-tumor liver tissues (Figure 5A), while ERBB3 expression was elevated in 34 of 37 tumor samples (Figures 5B). Since ERBB3 is defective in tyrosine kinase activity and requires dimerization with other receptors, possibly another member of the ERBB family (Riese and Stern, 1998), the hypothesis that HCC tumors expressing high levels of ERBB3 were associated with high expression of ERRB2 or EGFR was tested. The expression of ERBB2 was elevated in 12 of 37 tumors (Figure 5C), while high EGFR expression was found in 15 of 37 tumors (Figure 5D). A significant concomitant increase in ERBB2 expression (t-test P-value ~0.0026), but no association with high EGFR expression (t-test P-value ~0.31) was found in the top fifty percentile of high ERBB3-expressing HCC tumors, indicating that the cognate partners of ERRB3 appeared to be present in those tumors expressing high levels of ERBB3. Real-time and semi-quantitative RT-PCR analyses were also conducted on a panel of genes identified as differentially expressed in HCC (Figures 6-21).

Validation Of HCC Tumor Discriminator Expression Cassettes

Changes in gene expression of HCC using microarray technology have been reported (Chen et al, 2002, Okabe et al, 2001; Honda et al, 2001; Shirota et al, 2001; Tackels-Horne et al, 2001; Xu et al, 2001a; Xu et al, 2001b). The intersection of the important gene lists, derived as described herein, with gene lists reported previously was explored, and resulted in the identification of additional gene lists or "expression cassettes" (Tables 2-4) that were capable of distinguishing HCC tumor from non-tumor liver tissues.

In the first gene list, a total of 265 features, containing 245 unique UniGenes from the microarray used herein were observed to overlap (Table 2). Hierarchical clustering analyses based on expression levels of these 265 'overlap' features separated the tissue set into two distinct groups of tumor and non-tumor, with five tissue samples misclassified. Such clustering was significant ($P_a<1x10^{-6}$) based on random permutation testing of sample labels. The likelihood of a randomly chosen set of 265 features producing five or fewer samples misclassified was low ($P_b=1.5x10^{-3}$). Thus, these 265 'overlap' features could distinguish HCC tumor from non-tumor liver with reasonable precision, and the features were unlikely to appear by chance. Among

these genes were smaller subgroups characterized by distinct gene expression signatures involving potential different pathways. A cholesterol biosynthetic pathway was characterized by higher expression in HCC tumors for genes of the enzymes SQLE, ACLY and FDPS. A subgroup involved in growth and differentiation was characterized in the HCC tumor tissues by lower expression of ESR1, IGFBP3 and PDGFRα, and high expression of PPTB1. A subgroup of bZIP transcription factors ATF3, FOS, JUN, and MYBL2 was characterized to be down-regulated in the HCC tumor tissues.

Table 2. Intersection of microarray expression dataset with HCC Genes

UniGene Identifier	Gene	Description	GenBank No.
Hs.101408	BCAT2	branched chain aminotransferase 2, mitochondrial	BE264265, AA436410
Hs.101408	BCAT2	branched chain aminotransferase 2, mitochondrial	NM_001190, AA436410
Hs.102664	VAMP4	vesicle-associated membrane protein 4	AL035296, AA424813
Hs.10319	UGT2B7	UDP glycosyltransferase 2 family, polypeptide B7	J05428, AA746229
Hs.10359		ESTs	AW316760, AA630881
Hs.106061	RDBP	RD RNA-binding protein	X16105, AA056390
Hs.107253	DKFZP761F241	hypothetical protein DKFZp761F241 homo sapiens cDNA: FLJ20925 fis, clone ADSE00963	AW519080, R20416
Hs.108441	HAAO	3-hydroxyanthranilate 3,4-dioxygenase	NM_012205, T80846
Hs.108636	C1orf9	chromosome 1 open reading frame 9 CH1 MEMBRANE PROTENIN CH1	BE466870, N36176
Hs.110613	SMG1	PI-3-kinase-related kinase SMG-1 KIAA0220 KIAA0220 protein	AB007881, R97225
Hs.11314	DKFZP564N136	DKFZP564N1363 protein	AI360105, T87343
Hs.115617	CRHBP	corticotropin releasing hormone-binding protein	NM_001882, AA286752
Hs.118087	KIAA0610	KIAA0610 protein	AB011182, N38860
Hs.118638	NME1	non-metastatic cells 1, protein (NM23A) expressed in	AA147871, AA644092
Hs.118666	PP591	hypothetical protein PP591 human clone 23759 mRNA, partial cds	U79241, AA626336
Hs.119651	GPC3	glypican 3	U50410, AA775872
Hs.12451	EMAPL	echinoderm microtubule-associated protein-like	NM_004434, AA447196
Hs.12482	GNPAT	glyceronephosphate O-acyltransferase	AF043937,

			AA486845
Hs.125180	GHR	growth hormone receptor	X06562, N70358
Hs.125359	THY1	Thy-1 cell surface antigen	N94350, AI346653 AA428836
Hs.1265	BCKDHB	branched chain keto acid dehydrogenase E1, beta polypeptide (maple syrup urine disease)	NM_000056, AA427739
Hs.1279	C1R	complement component 1, r subcomponent	M14058, AA041382 AF045649, T69603
Hs.13999	KIAA0700	KIAA0700 protein	AI018400, N55167
Hs.1430	F11	coagulation factor XI (plasma thromboplastin antecedent)	AF045649, R88990
Hs.14453	ICSBP1	interferon consensus sequence binding protein 1	AW964220, N62269
Hs.14453	ICSBP1	interferon consensus sequence binding protein 1	AA514545, N62269
Hs.144904	NCOR1	nuclear receptor co-repressor 1	AB028970, AA085748
Hs.144904	NCOR1	nuclear receptor co-repressor 1	AA468619, AA085748
Hs.145567	AF038169	hypothetical protein	AI694342, AA406301
Hs.145567	AF038169	hypothetical protein	AI963556, AA406301
Hs.146360	IFITM1	interferon induced transmembrane protein 1 (9-27)	AA428847, AA419251
Hs.14838	FLJ10773	likely ortholog of mouse NPC derived proline rich protein 1	AA044181, R93542, AA401264
Hs.15087	C1orf16	chromosome 1 open reading frame 16 KIAA0250 KIAA0250 gene product	D87437, AA431423
Hs.151518	TARBP1	TAR (HIV) RNA-binding protein 1	NM_005646, N62244
Hs.15154	SRPX	sushi-repeat-containing protein, X chromosome	NM_006307, AA448569
Hs.1531	EHHADH	enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A dehydrogenase	L07077, R02373
Hs.1531	EHHADH	enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A dehydrogenase	AI800553, R02373
Hs.153357	PLOD3	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3	AF046889, AA459305
Hs.154890	FACL2	fatty-acid-Coenzyme A ligase, long-chain 2	D10040, T73556
Hs.155079	PPP2R5A	protein phosphatase 2, regulatory subunit B (B56), alpha isoform	AA234460, R59164
Hs.155560	CANX	calnexin	AA203197, AA126265
Hs.155637	PRKDC	protein kinase, DNA-activated, catalytic polypeptide	U34994, R27615
Hs.155956	NAT1	N-acetyltransferase 1 (arylamine N-acetyltransferase)	R79401, T67128
Hs.157148	MGC13204	hypothetical protein MGC13204 Homo sapiens cDNA FLJ11883 fis, clone	BE262748, N62451

		HEMBA1007178	
Hs.1578	BIRC5	baculoviral IAP repeat-containing 5 (survivin)	AW247335, AA460685
Hs.159301	IL18R1	interleukin 18 receptor 1	U43672, AA482489
Hs.160318	FXYD1	FXYD domain-containing ion transport regulator 1 (phospholemman)	AI125364, H57136
Hs.160786	ASS	argininosuccinate synthetase	BE393272, AA676466
Hs.16341	MAWBP	MAWD binding protein ESTs, Weakly similar to predicted using Genefinder [C. elegans]	AI866254, R54416
Hs.16426	PODXL	podocalyxin-like	BE395330, N64508
Hs.166891	RFX5	regulatory factor X, 5 (influences HLA class II expression)	AL050135, AA418045
Hs.167382	NPR1	natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)	AA598841
Hs.167529	CYP2C9	cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase), polypeptide 9	M61857, R89491
Hs.169517	ALDH1B1	aldehyde dehydrogenase 1 family, member B1	M63967, R93550
Hs.169756	C1S	complement component 1, s subcomponent	NM_001734, AA055520, T62048
Hs.169907	GSTA4	glutathione S-transferase A4	AF025887, AA152346
Hs.170001	EIF2B2	eukaryotic translation initiation factor 2B, subunit 2 (beta, 39kD)	AA678061, R86304
Hs.170001	EIF2B2	eukaryotic translation initiation factor 2B, subunit 2 (beta, 39kD)	AF035280, R86304
Hs.170133	FOXO1A	forkhead box O1A (rhabdomyosarcoma)	AF032885, AA448277
Hs.171955	TROAP	trophinin associated protein (tastin)	U04810, H94949
Hs.172665	MTHFD1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent), methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase	NM_005956, H10778
Hs.173717	PPAP2B	phosphatidic acid phosphatase type 2B	AI458142, T71976
Hs.173880	IL1RAP	interleukin 1 receptor accessory protein	AB006537, R35902, AA256132
Hs.174140	ACLY	ATP citrate lyase	AW967351, H08547
Hs.174220	CYP2C8	cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase), polypeptide 8	M17398, N53136
Hs.177592	RPLP1	ribosomal protein, large, P1	AW963733, AI732304
Hs.17767	KIAA1554	KIAA1554 protein	AI625594, AA857573, H17860
Hs.179718	MYBL2	v-myb avian myeloblastosis viral oncogene	X13293,

		homolog-like 2	AA456878
Hs.180383	DUSP6	dual specificity phosphatase 6	AB013382, AA630374
Hs.180919	ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein	AI950041, H82442
Hs.181345	SAH	SA (rat hypertension-associated) homolog	AI632754, N73827
Hs.182018	IRAK1	interleukin-1 receptor-associated kinase 1	NM_001569, AI202323, AA683550
Hs.18212	DXS9879E	DNA segment on chromosome X (unique) 9879 expressed sequence	W73156, AA479062
Hs.182575	SLC15A2	solute carrier family 15 (H+/peptide transporter), member 2	S78203, AA425352
Hs.1827	NGFR	nerve growth factor receptor (TNFR superfamily, member 16)	NM_002507, R55303
Hs.183858	TIF1	transcriptional intermediary factor 1	AF119042, R38345, AA016972
Hs.18443	ALDH8A1	aldehyde dehydrogenase 8 family, member A1 ESTs	AI051566, N70701
Hs.184697		Homo sapiens clone 23785 mRNA sequence	AF035307, AA041362, AA663440
Hs.18676	SPRY2	sprouty (Drosophila) homolog 2	NM_005842, AA453759
Hs.194660	CLN3	ceroid-lipofuscinosis, neuronal 3, juvenile (Batten, Spielmeyer-Vogt disease)	AW249073, W37752
Hs.194673	PEA15	phosphoprotein enriched in astrocytes 15	Y13736, AA293211
Hs.19554	C1orf2	chromosome 1 open reading frame 2	NM_006589, H11464
Hs.198282	PLSCR1	phospholipid scramblase 1	AB006746, N25945
Hs.19904	CTH	cystathionase (cystathionine gamma-lyase)	S52784, R07167
Hs.20144	SCYA14	small inducible cytokine subfamily A (Cys-Cys), member 14	NM_004166, R96626
Hs.2030	THBD	thrombomodulin	NM_000361, H59861
Hs.20315	IFIT1	interferon-induced protein with tetratricopeptide repeats 1	NM_001548, AA157787
Hs.2128	DUSP5	dual specificity phosphatase 5	NM_004419, W65460
Hs.213289	LDLR	low density lipoprotein receptor (familial hypercholesterolemia)	NM_000527, AA504461
Hs.21413	SLC12A5	solute carrier family 12, (potassium/chloride transporter) member 5	U79245, AA166885
Hs.21635	TUBG1	tubulin, gamma 1	NM_001070, T77732
Hs.2178	H2BFQ	H2B histone family, member Q	BE245642, AA010223
Hs.227656	XPR1	xenotropic and polytropic retrovirus receptor	AL137583, AA453474
Hs.23642	HSU79266	protein predicted by clone 23627	U79266, W95346
Hs.237356	SDF1	stromal cell-derived factor 1	AA442810,

			AA447115
Hs.237356	SDF1	stromal cell-derived factor 1	L36033, AA447115
Hs.23767	FLJ12666	hypothetical protein FLJ12666 Homo sapiens cDNA FLJ1266 fis, clone NT2RM4002256	AW952494, H10192, AA115300, AA131466
Hs.239	FOXM1	forkhead box M1	U83113, AA129552
Hs.239069	FHL1	four and a half LIM domains 1	AA725097, AA455925
Hs.239758	FLJ12389	hypothetical protein FLJ12389 similar to acetoacetyl-CoA synthetase Homo sapiens cDNA FLJ12389 fis, clone MAMMA1002671, weakly similar to Acetyl-coenzyme A synthase (EC 6.2.1.1)	AI697801, R48270
Hs.241561	PRSS2	protease, serine, 2 (trypsin 2)	U66061, AA284528
Hs.2430	TCFL1	transcription factor-like 1	AA705337, AA443950
Hs.24950	RGS5	regulator of G-protein signalling 5	AI674877, N34362, AA668470
Hs.252587	PTTG1	pituitary tumor-transforming 1	AA203476, AA430032
Hs.25313	MCRS1	microspherule protein 1	AF068007, AA488757
Hs.25475	AQP7	aquaporin 7	AW779701, H27752, AI075055
Hs.256583	ILF3	interleukin enhancer binding factor 3, 90kD	AF007140, AA449048
Hs.256583	ILF3	interleukin enhancer binding factor 3, 90kD	NM_012218, AA449048
Hs.25797	SF3B4	splicing factor 3b, subunit 4, 49kD	NM_005850, AA699361
Hs.262958	DKFZP434B044	hypothetical protein DKFZp434B044 ESTs	AA541776, AA460304
Hs.26403	GSTZ1	glutathione transferase zeta 1 (maleylacetoacetate isomerase)	U86529, AA428334
Hs.264330	ASAHL	N-acylsphingosine amidohydrolase (acid ceramidase)-like	BE267007, W47576
Hs.267289	POLA	polymerase (DNA directed), alpha	NM_016937, AA707650
Hs.2699	GPC1	glypican 1	NM_002081, AA455895
Hs.270256		Homo sapiens clone IMAGE:1963178, mRNA sequence ESTs	AI355014, R10140
Hs.270845	KNSL5	kinesin-like 5 (mitotic kinesin-like protein 1)	H63163, AA452513
Hs.279607	CAST	calpastatin	U38525, H78523
Hs.284142	C21orf4	chromosome 21 open reading frame 4	BE256559, W69668
Hs.284142	C21orf4	chromosome 21 open reading frame 4	BE142872, W69668

Hs.28465		Homo sapiens cDNA: FLJ21869 fis, clone HEP02442	AW582012, R63929
Hs.288650	AQP4	aquaporin 4	NM_001650, H09087
Hs.291904	DXS1357E	accessory proteins BAP31/BAP29	Z31696, AA625628
Hs.2934	RRM1	ribonucleotide reductase M1 polypeptide	X59543, AA633549
Hs.293970	ALDH6A1	methylmalonate-semialdehyde dehydrogenase	C00821, H63534, AA196160, H63534
Hs.293970	ALDH6A1	methylmalonate-semialdehyde dehydrogenase	M93405, N62179, AA196160, H63534
Hs.294151	KIAA1917	KIAA1917 protein	BE222511, AA452113
Hs.295923	SIAH1	seven in absentia (Drosophila) homolog 1	AA935716, T71889
Hs.296049	MFAP4	microfibrillar-associated protein 4	L38486, AA442695
Hs.296259	PON3	paraoxonase 3	L48516, R95740, T57069
Hs.296341	CAP2	adenylyl cyclase-associated protein 2	AW779995, AA040613
Hs.296341	CAP2	adenylyl cyclase-associated protein 2	U02390, AA040613
Hs.30151		ESTs, Weakly similar to JC5238 galactosylceramide-like protein, GCP [H.sapiens]	AA926994, N73570
Hs.30340	KIAA1165	hypothetical protein KIAA1165	AB032991, AA449330
Hs.30340	KIAA1165	hypothetical protein KIAA1165	AA770150, AA449330
Hs.3416	ADFP	adipose differentiation-related protein	NM_001122, AA700054, AA142916
Hs.35120	RFC4	replication factor C (activator 1) 4 (37kD)	AA600213, N93924
Hs.3530	FUSIP2	FUS-interacting protein (serine-arginine rich) 2 TLS-associated serine-arginine protein 2	AK001656, H11042
Hs.36102		ESTs, Highly similar to SMHU1B metallothionein 1B [H.sapiens]	R99207, H72722
Hs.37009	ALPI	alkaline phosphatase, intestinal	NM_001631, AA190871
Hs.38163		Homo sapiens, Similar to hypothetical protein, MGC:7035, clone MGC:20737 IMAGE:4563636, mRNA, complete cds ESTs	AW074863, H63116
Hs.3873	PPT1	palmitoyl-protein thioesterase 1 (ceroid-lipofuscinosis, neuronal 1, infantile)	AL037943, AA034250
Hs.388	NUDT1	nudix (nucleoside diphosphate linked moiety X)-type motif 1	AI656937, AA443998
Hs.4	ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide	M24317, N93428
Hs.41726	SERPINB8	serine (or cysteine) proteinase inhibitor,	NM_002640,

		clade B (ovalbumin), member 8	W60100
Hs.4187	LOC55977	hypothetical protein 24636	AI066576, N62562
Hs.42650	ZWINT	ZW10 interactor	AW409765, AA706968
Hs.44532	UBD	diubiquitin	NM_006398, N33920
Hs.460	ATF3	activating transcription factor 3	N39944, H21041
Hs.4742	GPAA1	anchor attachment protein 1 (Gaa1p, yeast) homolog	NM_003801, AA455301
Hs.4756	FEN1	flap structure-specific endonuclease 1	BE278623, AA620553
Hs.4788	NCSTN KIAA0253	Nicastrin nicastrin	D87442, R96527
Hs.4788	NCSTN KIAA0253	Nicastrin nicastrin	BE179772, R96527
Hs.48348	HH114	hypothetical protein HH114 Homo sapiens clone HH114 unknown mRNA	AA428370, AA130117
Hs.4854	CDKN2C	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)	AF041248, N72115
Hs.49265		ESTs	AI141174, AI140241
Hs.49912	PXMP2	peroxisomal membrane protein 2 (22kD)	BE393339, N70714
Hs.50758	SMC4L1	SMC4 (structural maintenance of chromosomes 4, yeast)-like 1 CAP-C chromosome-associated polypeptide C	AB019987, AA452095
Hs.50966	CPS1	carbamoyl-phosphate synthetase 1, mitochondrial	Y15793, N68399
Hs.50966	CPS1	carbamoyl-phosphate synthetase 1, mitochondrial	AA113231, N68399
Hs.5333	KIAA0711	KIAA0711 gene product	NM_014867, AA702544
Hs.5719	CNAP1	chromosome condensation-related SMC-associated protein 1	D63880, AA668256
Hs.5719	CNAP1	chromosome condensation-related SMC-associated protein 1	NM_014865, AA668256
Hs.572	ORM1	orosomucoid 1	X02544, AA700876
Hs.574	FBP1	fructose-1,6-bisphosphatase 1	M19922, AA699427
Hs.5897		Homo sapiens mRNA; cDNA DKFZp586P1622 (from clone DKFZp586P1622)	AI383214, T59658
Hs.61638	MYO10	myosin X	AI198676, AA187977
Hs.6551	ATP6S1	ATPase, H+ transporting, lysosomal (vacuolar proton pump), subunit 1	NM_001183, AA487588
Hs.6566	TRIP13	thyroid hormone receptor interactor 13	BE090548, AA630784
Hs.66	IL1RL1	interleukin 1 receptor-like 1	AB012701, AA125917
Hs.6838	ARHE	ras homolog gene family, member E ESTs	W03441, AA443302

Hs.71465	SQLE	squalene epoxidase	AF098865, R01118
Hs.71622	SMARCD3	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 3 Weakly similar to KIAA0319 [H. sapiens]	U66619, AA136103
Hs.737	ETR101	immediate early protein	AA194084, AA496359
Hs.738	RPL14	ribosomal protein L14 early growth response 1	BE410686, AA486533
Hs.73986	CLK2	CDC-like kinase 2	NM_003993, AA282845
Hs.740	PTK2	PTK2 protein tyrosine kinase 2	NM_005607, AA291486
Hs.74120	APM2	adipose specific 2	AI093004, W94684
Hs.74170	MT1E	metallothionein 1E (functional)	H72532, AA872383
Hs.74561	A2M	alpha-2-macroglobulin	NM_000014, AA775447
Hs.74566	DPYSL3	dihydropyrimidinase-like 3	D78014, AI831083
Hs.74579	KIAA0263	KIAA0263 gene product	D87452, AA634464
Hs.74615	PDGFRA	platelet-derived growth factor receptor, alpha polypeptide	M21574, H23235
Hs.74615	PDGFRA	platelet-derived growth factor receptor, alpha polypeptide	AW887370, H23235
Hs.74711	DNAJC8	DnaJ (Hsp40) homolog, subfamily C, member 8 Splicing factor similar to dnaJ	AA513669, T60163
Hs.748	FGFR1	fibroblast growth factor receptor 1 (fms-related tyrosine kinase 2, Pfeiffer syndrome)	X66945, R54610
Hs.75103	YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide	AA911031, AA609598, H94670, AA485749
Hs.75103	YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide	BE315169 AA609598, H94670, AA485749
Hs.75106	CLU	clusterin (complement lysis inhibitor, SP-40,40, sulfated glycoprotein 2, testosterone-repressed prostate message 2, apolipoprotein J)	M25915, AA292226, AA464163
Hs.75117	ILF2	interleukin enhancer binding factor 2, 45kD	AA307289, AA894687, H95638
Hs.75117	ILF2	interleukin enhancer binding factor 2, 45kD	AA601029, AA894687, H95638
Hs.75196	BAT8	HLA-B associated transcript 8 G9A ankyrin repeat-containing protein	NM_006709, AA434117
Hs.75216	PTPRF	protein tyrosine phosphatase, receptor type, F	F08552, AA598513

Hs.75318	TUBA1	tubulin, alpha 1 (testis specific)	X06956, R36063, AA180742
Hs.75361	PK1.3	gene from NF2/meningioma region of 22q12	AB023200, AA700048
Hs.75438	QDPR	quinoid dihydropteridine reductase	AA159812, R38198
Hs.75545	IL4R	interleukin 4 receptor	X52425, AA292025
Hs.75572	CPB2	carboxypeptidase B2 (plasma, carboxypeptidase U)	NM_001872, H47837
Hs.75618	RAB11A	RAB11A, member RAS oncogene family	BE122870, AA025058
Hs.75658	PYGB	phosphorylase, glycogen; brain	U47025, AA922705
Hs.75678	FOSB	FBJ murine osteosarcoma viral oncogene homolog B	L49169, T61948
Hs.75812	PCK2	phosphoenolpyruvate carboxykinase 2 (mitochondrial)	X92720, AA186901
Hs.76252	EDNRA	endothelin receptor type A	D90348, AA450009
Hs.76325	SLU	ESTs, Highly similar to IGJ_HUMAN IMMUNOGLOBULIN J CHAIN [H.sapiens] step II splicing factor SLU7	AW172754, T70057
Hs.7645	FGB	fibrinogen, B beta polypeptide	AW589878, H91121, T73858
Hs.76461	RBP4	retinol-binding protein 4, plasma	AF074657, T72076
Hs.7647	MAZ	MYC-associated zinc finger protein (purine-binding transcription factor)	BE264373, AA704613
Hs.77256	EZH2	enhancer of zeste (Drosophila) homolog 2	U52965, AA428252
Hs.77326	IGFBP3	insulin-like growth factor binding protein 3	BE336944, AA598601
Hs.77393	FDPS	farnesyl diphosphate synthase (farnesyl pyrophosphate synthetase, dimethylallyltranstransferase, geranyltranstransferase)	D14697, T65790
Hs.77597	PLK	polo (Drosophia)-like kinase	X75932, AA629262
Hs.77667	LY6E	lymphocyte antigen 6 complex, locus E	NM_002346, AA865464
Hs.77854	RGN	regucalcin (senescence marker protein-30)	AB032064, H05140
Hs.78045	ACTG2 TFPI2	actin, gamma 2, smooth muscle, enteric tissue pathway inhibitor 2	NM_001615, AA293402
Hs.78465	JUN	v-jun avian sarcoma virus 17 oncogene homolog	AI885769, W96134
Hs.78524	HTCD37	TcD37 homolog	AI263464, AA022472, AA456635
Hs.78865	TAF6	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 80 kD	NM_005641, R19071
Hs.789	GRO1	GRO1 oncogene (melanoma growth stimulating activity, alpha)	NM_001511, W46900
Hs.78996	PCNA	proliferating cell nuclear antigen	AI624204,

			AA450264
Hs.79078	MAD2L1	MAD2 (mitotic arrest deficient, yeast, homolog)-like 1	NM_002358, AA481076
Hs.79081	PPP1CC	protein phosphatase 1, catalytic subunit, gamma isoform	NM_002710, AA129930
Hs.79088	RCN2	reticulocalbin 2, EF-hand calcium binding domain	AL120373, AA598676
Hs.79334	NFIL3	nuclear factor, interleukin 3 regulated	X64318, AA633811
Hs.79404	D4S234E	neuron-specific protein	AA975473, AA875888
Hs.80248	RBPMS	RNA-binding protein gene with multiple splicing	D84107, T98807
Hs.80658	UCP2	uncoupling protein 2 (mitochondrial, proton carrier)	AW192446, H61242
Hs.81170	PIM1	pim-1 oncogene	M54915, AA447730
Hs.8136	EPAS1	endothelial PAS domain protein 1 Homo sapiens clone 23698mRNA sequence	U51626, R24882
Hs.81687	NME3	non-metastatic cells 3, protein expressed in	U29656, AA398218
Hs.81848	RAD21	RAD21 (S. pombe) homolog	NM_006265, AA683102
Hs.81892	KIAA0101	KIAA0101 gene product	D14657, W68219
Hs.82042	SLC23A1	solute carrier family 23 (nucleobase transporters), member 1	D87075, N23756
Hs.821	BGN	Biglycan Zinc finger protein homologous to Zfp92 in mouse	NM_001711, R77226, N51018
Hs.82112	IL1R1	interleukin 1 receptor, type I	M27492, R56687, AA464525
Hs.82273	FLJ20152	hypothetical protein	AI536745, AA446864
Hs.82503		Homo sapiens cDNA FLJ30550 fis, clone BRAWH2001502 Homo sapiens mRNA for 3'UTR of unknown protein	Y09836, AA670382
Hs.8265	TGM2	transglutaminase 2 (C polypeptide, protein-glutamine-gamma-glutamyltransferase)	M98479, AA156324
Hs.82794	CETN2	centrin, EF-hand protein, 2	NM_004344, N72193
Hs.82906	CDC20	CDC20 (cell division cycle 20, S. cerevisiae, homolog)	BE293657
Hs.8294	KIAA0196	KIAA0196 gene product	NM_014846
Hs.82962	TYMS	thymidylate synthetase	NM_001071
Hs.83164	COL15A1	collagen, type XV, alpha 1	L01697
Hs.83753	SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	BE252108
Hs.86368	CLGN	calmegin	NM_004362
Hs.86724	GCH1	GTP cyclohydrolase 1 (dopa-responsive dystonia)	Z29433
Hs.87409	THBS1	thrombospondin 1	NM_003246
Hs.8765	RNAHP	RNA helicase-related protein	AI127821
Hs.8867	CYR61	cysteine-rich, angiogenic inducer, 61	Y12084
Hs.8889	SHMT1	serine hydroxymethyltransferase 1	AI761724

		(soluble)	
Hs.89538	CETP	cholesteryl ester transfer protein, plasma	M30185
Hs.89691	UGT2B4	UDP glycosyltransferase 2 family, polypeptide B4	AF064200
Hs.89771	GCKR	glucokinase (hexokinase 4) regulatory protein	NM_001486
Hs.91813	BTN2A2	butyrophilin, subfamily 2, member A2	AI636514
Hs.93002	UBE2C	ubiquitin-conjugating enzyme E2C	AI637467
Hs.93194	APOA1	apolipoprotein A-I	X00566
Hs.93210	C8A	complement component 8, alpha polypeptide	M16974
Hs.93597	CDK5R1	cyclin-dependent kinase 5, regulatory subunit 1 (p35)	T04872
Hs.93597	CDK5R1	cyclin-dependent kinase 5, regulatory subunit 1 (p35)	AW088206
Hs.93832	LOC54499	putative membrane protein	AW081809
Hs.94360	MT1L	metallothionein 1L	F26137
Hs.94382	ADK	adenosine kinase	NM_001123
Hs.9568	ZNF261	zinc finger protein 261	X95808
Hs.9568	ZNF261	zinc finger protein 261	NM_005096
Hs.95998	FRDA	Friedreich ataxia	AW409831
Hs.9629	PRCC	papillary renal cell carcinoma (translocation-associated)	BE258195
Hs.9670	FLJ10948	hypothetical protein FLJ10948	AA805411

In the second gene list, a total of 230 features, containing 166 unique UniGenes from the 218 significant gene list (containing 213 unique UniGenes) identified herein (Table 1) were observed to overlap (Table 3). Hierarchical clustering analysis based on the expression levels of these 230 'overlap' features separated the tissue set into distinct tumor and non-tumor groups, with four tissue samples misclassified. Random permutation of sample labels indicated that the clustering was significant ($P_a<1x10^{-8}$) and it was unlikely that a randomly chosen set of 230 features could produce four or fewer samples misclassified ($P_b<1x10^{-4}$). These 230 'overlap' features are therefore able to discern HCC tumor from non-tumor liver.

Table 3. Intersection of Significant Genes Identified Herein with HCC Genes

UniGene Identifier	Gene	Description	GenBank No.
Hs.103804	HNRPU	heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A)	X65488, T97547, AA496741
Hs.104143	CLTA	clathrin, light polypeptide (Lca)	AW974204, AA113872
Hs.105465	SNRPF	small nuclear ribonucleoprotein polypeptide F	AA649986, AA668189

Hs.108332	UBE2D2	ubiquitin-conjugating enzyme E2D 2 (homologous to yeast UBC4/5)	NM_003339, AA159600, AA431868
Hs.10842	RAN	RAN, member RAS oncogene family	NM_006325, AA456636
Hs.10848	KIAA0187	KIAA0187 gene product	D80009, AA121504, AA129555, AA402812
Hs.108636	C1orf9	chromosome 1 open reading frame 9	BE466870, N36176
Hs.108689	SREBF2	sterol regulatory element binding transcription factor 2	AA608556, AA701914, AA608556
Hs.108809	CCT7	chaperonin containing TCP1, subunit 7 (eta)	AA314436, AA676588
Hs.110713	DEK	DEK oncogene (DNA binding)	AI888504, R25377
Hs.1119	NR4A1	nuclear receptor subfamily 4, group A, member 1	NM_002135, N94487
Hs.11355	TMPO	thymopoietin	U09087, H21746, AA676998, T63980
Hs.115617	CRHBP	corticotropin releasing hormone-binding protein	NM_001882, N26546, AA286752
Hs.117367	SLC22A1	solute carrier family 22 (organic cation transporter), member 1	X98332, AA702013
Hs.1174	CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	AI859822, AA877595
Hs.118249	ARFGEF2	ADP-ribosylation factor guanine nucleotide-exchange factor 2 (brefeldin A-inhibited)	AA099582, N34053
Hs.118638	NME1	non-metastatic cells 1, protein (NM23A) expressed in	AA147871, AA644092
Hs.11902	MYLE	MYLE protein	AA628977, T68845
Hs.119651	GPC3	glypican 3	U50410, AA775872
Hs.12107	BC-2	putative breast adenocarcinoma marker (32kD)	AF042384, N25578
Hs.12482	GNPAT	glyceronephosphate O-acyltransferase	AF043937, W72079
Hs.125180	GHR	growth hormone receptor	X06562, N70358, AA775738
Hs.13340	HAT1	histone acetyltransferase 1	AF030424, AA625662
Hs.148495	PSMD4	proteasome (prosome, macropain) 26S subunit, non-ATPase, 4	AA604027, AA450227
Hs.151787	U5-116KD	U5 snRNP-specific protein, 116 kD	D21163, AA779221
Hs.152931	LBR	lamin B receptor	L25931, AA099136
Hs.15318	HAX1	HS1 binding protein	BE260953, R76263
Hs.154073	UGTREL1	UDP-galactose transporter related	AW192554, R41839
Hs.155079	PPP2R5A	protein phosphatase 2, regulatory subunit B (B56), alpha isoform	AA234460, R59164
Hs.155637	PRKDC	protein kinase, DNA-activated, catalytic polypeptide	U34994, R27615
Hs.156110	IGKC	immunoglobulin kappa constant	AW404507, AI732289, AA476918, AA486362
Hs.1600	CCT5	chaperonin containing TCP1, subunit 5	D43950,

		(epsilon)	AA126599, AA629692
Hs.1624	EFNA1	ephrin-A1	NM_004428, AA857015
Hs.16341	MAWBP	MAWD binding protein	AI866254, R54416
Hs.16426	PODXL	podocalyxin-like	BE395330, N64508
Hs.1657	ESR1	estrogen receptor 1	AL078582, AA164585, AA291702
Hs.166468	PDCD5	programmed cell death 5	AA452724, AA156940
Hs.166891	RFX5	regulatory factor X, 5 (influences HLA class II expression)	AL050135, AA418045
Hs.1674	GFPT1	glutamine-fructose-6-phosphate transaminase 1	NM_002056, AA478571
Hs.169407	SACM2L	SAC2 (suppressor of actin mutations 2, yeast, homolog)-like	AK001725, AA454836
Hs.173274	ICAP-1A	integrin cytoplasmic domain-associated protein 1	AF012023, AA456882
Hs.174140	ACLY	ATP citrate lyase	AW967351, H08547, AA126708
Hs.174220	CYP2C8	cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase), polypeptide 8	M17398, N53136
Hs.177592	RPLP1	ribosomal protein, large, P1	AW963733, AI732304
Hs.180414	HSPA8	heat shock 70kD protein 8	AW249010, H64096, AA629567
Hs.180446	KPNB1	karyopherin (importin) beta 1	L38951, AA121732, AA251527, AA425006
Hs.180577	GRN	granulin	AI375908, AI054019, AA496452
Hs.180610	SFPQ	splicing factor proline/glutamine rich (polypyrimidine tract-binding protein-associated)	X70944, R96240, N24024, AA425258
Hs.181357	LAMR1	laminin receptor 1 (67kD, ribosomal protein SA)	AW328280, AA629897
Hs.181444	LOC51235	hypothetical protein	AI190653, AA455565
Hs.184222	DSCR1	Down syndrome critical region gene 1	U85267, AA629707
Hs.18443	ALDH8A1	aldehyde dehydrogenase 8 family, member A1	AI051566, N70701
Hs.194673	PEA15	phosphoprotein enriched in astrocytes 15	Y13736, AA293211
Hs.1989	SRD5A2	steroid-5-alpha-reductase, alpha polypeptide 2	M74047, AI420552
Hs.199067	ERBB3	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 3	AI565773, N24966, AA042878
Hs.199263	MT1L STK39	metallothionein 1L serine threonine kinase 39 (STE20/SPS1 homolog, yeast)	F26137, H84871
Hs.2	NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)	D90040, AI262683
Hs.20144	SCYA14	small inducible cytokine subfamily A (Cys-Cys), member 14	NM_004166, R96626

Hs.20716	TIM17	translocase of inner mitochondrial membrane 17 homolog A (yeast)	AW247564, AA708446
Hs.22785	GABRE	gamma-aminobutyric acid (GABA) A receptor, epsilon	NM_004961, H63532
Hs.236774	HMG17L3	high-mobility group (nonhistone chromosomal) protein 17-like 3	U90549, R17124
Hs.236828	WHIP	Werner helicase interacting protein	AA481600, AA188168
Hs.237356	SDF1	stromal cell-derived factor 1	L36033, AA447115
Hs.23767	FLJ12666	hypothetical protein FLJ12666	AW952494, AA432056
Hs.24485	CSPG6	chondroitin sulfate proteoglycan 6 (bamacan)	NM_005445, W40150, AA463410
Hs.245710	HNRPH1	heterogeneous nuclear ribonucleoprotein H1 (H)	BE296051, R11018, W96058
Hs.247324	MRPS14	mitochondrial ribosomal protein S14	AW973521, T51290, AA460831
Hs.24950	RGS5	regulator of G-protein signalling 5	AI674877, N34362, AA668470
Hs.25132	KIAA0470	KIAA0470 gene product	NM_014812, AI049669, AA167129, AA187982
Hs.252229	MAFG	v-maf musculoaponeurotic fibrosarcoma (avian) oncogene family, protein G	AF059195, N21609, AA045436
Hs.25647	FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	V01512, R12840, N36944, AA485377
Hs.25797	SF3B4	splicing factor 3b, subunit 4, 49kD	NM_005850, AA699361
Hs.26403	GSTZ1	glutathione transferase zeta 1 (maleylacetoacetate isomerase)	U86529, AA428334
Hs.26433	DPAGT1	dolichyl-phosphate (UDP-N-acetylglucosamine) N-acetylglucosaminephosphotransferase 1 (GlcNAc-1-P transferase)	Z82022, R55619, AA452517
Hs.271980	MAPK6	mitogen-activated protein kinase 6	NM_002748, AA603152, H17504
Hs.275163	NME2	non-metastatic cells 2, protein (NM23B) expressed in	L16785, AA422058, AA496512
Hs.287797	ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12) Homo sapiens, clone MGC: 17220	W38716, AA037283, W67173
Hs.291904	DXS1357E	accessory proteins BAP31/BAP29	Z31696, AA625628
Hs.2934	RRM1	ribonucleotide reductase M1 polypeptide	X59543, AA633549
Hs.293441		Homo sapiens SNC73 protein (SNC73) mRNA, complete cds	AA290845, H28469, H73590
Hs.296341	CAP2	adenylyl cyclase-associated protein 2	AW779995, AA040613
Hs.300697	IGHG3	immunoglobulin heavy constant gamma 3 (G3m marker)	D78345, AA740786, N92646, AA465378
Hs.301005	H2AV	histone H2A.F/Z variant	BE409809, H97000
Hs.301404	RBM3	RNA binding motif protein 3	NM_006743,

			AA054287
Hs.301819	ZNF146	zinc finger protein 146	X70394, AA504351
Hs.3041	UNG2	uracil-DNA glycosylase 2	AA291356, AA425900
Hs.3164	NUCB2	nucleobindin 2	AW951523, W93954, AA484939
Hs.321231	B4GALT3	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 3	Y12509, AA424578
Hs.323817	DKFZP547E101	DKFZP547E1010 protein	NM_015607, AA406292, AA418004
Hs.332633	BBS2	Bardet-Biedl syndrome 2	AA425759, AA486738
Hs.333495	DSS1	Deleted in split-hand/split-foot 1 region	W79057, H85464
Hs.334612	SNRPE	small nuclear ribonucleoprotein polypeptide E	X12466, AA678021
Hs.334787	MGC19556	hypothetical protein MGC19556	BE379431, AA609463
Hs.342389	PPIA	peptidylprolyl isomerase A (cyclophilin A)	AW732921, H72674
Hs.349961	RPL6	ribosomal protein L6	AW675430, AA629808
Hs.356525	FLJ12806	ESTs, Weakly similar to CNG1_HUMAN cGMP-gated cation channel alpha 1 (CNG channel alpha 1)	BE044582, T73794
Hs.3610	KIAA0205	KIAA0205 gene product	D86960, R91263
Hs.36102		ESTs, Highly similar to SMHU1B metallothionein 1B [H.sapiens]	R99207, H72722
Hs.4	ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide	M24317, N93428
Hs.41587	RAD50	RAD50 (S. cerevisiae) homolog	Z75311, H99196, AA126482
Hs.431	BMI1	murine leukemia viral (bmi-1) oncogene homolog	AA884913, AA608856, T87514, W90704, AA478036
Hs.44532	UBD	diubiquitin	NM_006398, N33920
Hs.44585	TP53BP2	tumor protein p53-binding protein, 2	AI123916, H69077, N34418
Hs.46440	SLC21A3	solute carrier family 21 (organic anion transporter), member 3	U21943, N62948
Hs.4756	FEN1	flap structure-specific endonuclease 1	BE278623, AA620553
Hs.50758	SMC4L1	SMC4 (structural maintenance of chromosomes 4, yeast)-like 1	AB019987, AA452095
Hs.5085	DPM1	dolichyl-phosphate mannosyltransferase polypeptide 1, catalytic subunit	AW173486, AA004759
Hs.52002	CD5L	CD5 antigen-like (scavenger receptor cysteine rich family)	NM_005894, AA677254
Hs.554	SSA2	Sjogren syndrome antigen A2 (60kD, ribonucleoprotein autoantigen SS-A/Ro)	NM_004600, AA010351
Hs.5662	GNB2L1	guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1	BE206815, AA640657, R96220
Hs.57101	MCM2	minichromosome maintenance deficient (S.	BE250461,

		cerevisiae) 2 (mitotin)	AA454572
Hs.5737	KIAA0475	KIAA0475 gene product	AA524523, N73927
Hs.57783	EIF3S9	eukaryotic translation initiation factor 3, subunit 9 (eta, 116kD)	U62583, AA676471
Hs.6127		Homo sapiens cDNA: FLJ23020 fis, clone LNG00943	AA054768, T67278
Hs.6551	ATP6S1	ATPase, H+ transporting, lysosomal interacting protein 1	NM_001183, AA487588
Hs.6650	VPS45B	vacuolar protein sorting 45B (yeast homolog)	AA702845, AA885433
Hs.6838	ARHE	ras homolog gene family, member E	W03441, W86282, AA443302
Hs.695	CSTB	cystatin B (stefin B)	AI831499, 110374
Hs.699	PPIB	peptidylprolyl isomerase B (cyclophilin B)	BE386706, N45313, AA481464
Hs.69997	ZNF238	zinc finger protein 238	AJ223321, R79722
Hs.74441	CHD4	chromodomain helicase DNA binding protein 4	BE408958, N34372
Hs.75117	ILF2	interleukin enhancer binding factor 2, 45kD	AA307289, AA894687, H95638
Hs.75183	CYP2E	cytochrome P450, subfamily IIE (ethanol-inducible)	J02843, H50500
Hs.75187	KIAA0016	translocase of outer mitochondrial membrane 20 (yeast) homolog	D13641, AA644550
Hs.75258	H2AFY	H2A histone family, member Y	AA307460, AA486003
Hs.75354	GCN1L1	GCN1 (general control of amino-acid synthesis 1, yeast)-like 1	D86973, R55250
Hs.75412	ARMET	arginine-rich, mutated in early stage tumors	AA582041, R91550
Hs.75424	ID1	inhibitor of DNA binding 1, dominant negative helix-loop-helix protein	S78825, AA457158
Hs.75546	CAPZA2	capping protein (actin filament) muscle Z-line, alpha 2	U03851, AA083228
Hs.75659	MPV17	MpV17 transgene, murine homolog, glomerulosclerosis	NM_002437, R55046
Hs.75678	FOSB	FBJ murine osteosarcoma viral oncogene homolog B	L49169, T61948
Hs.75981	USP14	ubiquitin specific protease 14 (tRNA-guanine transglycosylase)	NM_005151, AA039511, T65861
Hs.76230	RPS10	ribosomal protein S10	AW245775, AA828564, AA828819, AI054003
Hs.76285	DKFZP564B 167	DKFZP564B167 protein	AI032331, AA621342
Hs.76325	IGJ	immunoglobulin J polypeptide, linker protein for immunoglobulin alpha and mu polypeptides Homo sapiens, clone MGC: 24130	AW172754, T90492, T70057
Hs.7655	U2AF65	U2 small nuclear ribonucleoprotein auxiliary factor (65kD)	AA936430, AA405748
Hs.7720	DNCH1	dynein, cytoplasmic, heavy polypeptide 1	AB002323, AA010589, W78967
Hs.77254	CBX1	chromobox homolog 1 (HP1 beta homolog Drosophila)	AL046741, AA448667

Hs.77326	IGFBP3	insulin-like growth factor binding protein 3	BE336944, AA598601
Hs.77608	SFRS9	splicing factor, arginine/serine-rich 9	AL021546, N47892, AA490721
Hs.78065	C7	complement component 7	X86328, AA598478
Hs.78902	VDAC2	voltage-dependent anion channel 2	AI015604, AA857093, T66813
Hs.79090	XPO1	exportin 1 (CRM1, yeast, homolog)	D89729, T59055
Hs.79110	NCL	nucleolin	AK000250, AA433818
Hs.79150	CCT4	chaperonin containing TCP1, subunit 4 (delta)	U38846, T98634, AA088226, AA598637
Hs.79162	SSRP1	structure specific recognition protein 1	AI635077, R11356
Hs.80343	MMP15	matrix metalloproteinase 15 (membrane-inserted)	D85510, AA443300
Hs.80552	DPT	dermatopontin	AW016451, R48303
Hs.809	HGF	hepatocyte growth factor (hepapoietin A; scatter factor)	X16323, R52797
Hs.80917	AP3S1	adaptor-related protein complex 3, sigma 1 subunit	D63643, AA996044
Hs.80919	SYPL	synaptophysin-like protein	S72481, AA427447
Hs.81972	SHC1	SHC (Src homology 2 domain-containing) transforming protein 1	X68148, R52960, T50498
Hs.82043	D123	D123 gene product	U27112, AA448289
Hs.82159	PSMA1	proteasome (prosome, macropain) subunit, alpha type, 1	AI889267, R27585
Hs.82793	PSMB3	proteasome (prosome, macropain) subunit, beta type, 3	AI028114, AA620580
Hs.82916	CCT6A	chaperonin containing TCP1, subunit 6A (zeta 1)	L27706, AA872690, H84286
Hs.83753	SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	BE252108, AA599116
Hs.84790	C7orf14	chromosome 7 open reading frame 14	D86978, AA600190
Hs.85119	SMT3H1	SMT3 (suppressor of mif two 3, yeast) homolog 1	AA160893, AA862529, AA872379
Hs.8765	RNAHP	RNA helicase-related protein	AI814448, T56221
Hs.8867	CYR61	cysteine-rich, angiogenic inducer, 61	Y12084, AA777187
Hs.89525	HDGF	hepatoma-derived growth factor (high-mobility group protein 1-like)	BE259164, AA453749
Hs.90093	HSPA4	heat shock 70kD protein 4	AB023420, AA131267, AA433916
Hs.90370	ARPC1A	actin related protein 2/3 complex, subunit 1A (41 kD)	Y08999, AA490209, AA016251, AA151930
Hs.90744	PSMD11	proteasome (prosome, macropain) 26S subunit, non-ATPase, 11	AB003102
Hs.99969	FUS	fusion, derived from t(12;16) malignant liposarcoma	BE396632, 101207

In the third gene list, a total of 68 unique UniGenes from the 218 significant gene list (containing 213 unique UniGenes) identified herein (Table 1) were observed to overlap (Table 4), and the likelihood that the overlap would arise by chance if the two gene lists were totally independent was minuscule ($P_c<1x10^{-8}$).

5

Table 4. Intersection of Significant Genes Identified Herein with HCC Genes

UniGene Identifier	Gene	Description	GenBank No.
Hs.119651	GPC3	glypican 3	U50410, AA775872
Hs.125180	GHR	growth hormone receptor	X06562, N70358
Hs.180577	GRN	granulin	AI375908, AA496452
Hs.44585	TP53BP2	tumor protein p53-binding protein, 2	AI123916, H69077
Hs.77326	IGFBP3	insulin-like growth factor binding protein 3	BE336944, AA598601
Hs.8867	CYR61	cysteine-rich, angiogenic inducer, 61	Y12084, AA777187
Hs.1600	CCT5 HSEC61	chaperonin containing TCP1, subunit 5 (epsilon) sec 61 homolog	D43950, AA629692
Hs.75410	HSPA5	heat shock 70kD protein 5 (glucose-regulated protein, 78kD)	AL043206, AA962446
Hs.152931	LBR	lamin B receptor	L25931, AA099136
Hs.7720	DNCH1	dynein, cytoplasmic, heavy polypeptide 1	AB002323, W78967
Hs.4756	FEN1	flap structure-specific endonuclease 1	BE278623, AA620553
Hs.50758	SMC4L1	SMC4 (structural maintenance of chromosomes 4, yeast)-like 1 CAP-C chromosome associated polypeptide C	AB019987, AA452095
Hs.77254	CBX1	chromobox homolog 1 (HP1 beta homolog Drosophila)	AL046741, AA448667
Hs.2934	RRM1	ribonucleotide reductase M1 polypeptide	X59543, AA633549
Hs.156110	IGKC	immunoglobulin kappa constant	AW404507, AA402920, AA486362
Hs.20144	SCYA14	small inducible cytokine subfamily A (Cys-Cys), member 14	NM_004166, R96626
Hs.237356	SDF1	stromal cell-derived factor 1	L36033, AA447115
Hs.78065	C7	complement component 7	X86328, AA598478
Hs.118638	NME1	non-metastatic cells 1, protein (NM23A) expressed in	AA147871, AA644092
Hs.12482	GNPAT	glyceronephosphate O-acyltransferase	AF043937, AA486845
Hs.174140	ACLY	ATP citrate lyase	AW967351, H08547
Hs.174220	CYP2C8	cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase), polypeptide 8	M17398, N53136
Hs.26403	GSTZ1	glutathione transferase zeta 1 (maleylacetoacetate isomerase)	U86529, AA428334

Hs.4	ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide	M24317, N93428
Hs.177592	RPLP1	ribosomal protein, large, P1	AW963733, AI732304
Hs.25797	SF3B4	splicing factor 3b, subunit 4, 49kD	NM_005850, AA699361
Hs.76230	RPS10	ribosomal protein S10	AW245775, AI054003
Hs.76325	IGJ SLU7	immunoglobulin J polypeptide, linker protein for immunoglobulin alpha and mu polypeptides step II splicing factor SLU7	AW172754, T70057
Hs.83753	SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	BE252108, AA599116
Hs.115617	CRHBP	corticotropin releasing hormone-binding protein	NM_001882, AA286752
Hs.118249	ARFGEF2 BIG2	ADP-ribosylation factor guanine nucleotide-exchange factor 2 (brefeldin A-inhibited) Brefeldin A-inhibited guanine nucleotide-exchange protein	AA099582, N34053
Hs.155079	PPP2R5A	protein phosphatase 2, regulatory subunit B (B56), alpha isoform	AA234460, R59164
Hs.155637	PRKDC	protein kinase, DNA-activated, catalytic polypeptide	U34994, R27615
Hs.1624	EFNA1	ephrin-A1	NM_004428, AA857015
Hs.182278	CALM2	calmodulin 2 (phosphorylase kinase, delta)	D45887, AA043551
Hs.199263	STK39 SPAK	serine threonine kinase 39 (STE20/SPS1 homolog, yeast) ste-20 related kinase	F26137, H84871
Hs.22785	GABRE	gamma-aminobutyric acid (GABA) A receptor, epsilon	NM_004961, H63532
Hs.24950	RGS5	regulator of G-protein signalling 5	AI674877, N34362, AA668470
Hs.296341	CAP2	adenylyl cyclase-associated protein 2	AW779995, AA040613
Hs.81972	SHC1	SHC (Src homology 2 domain-containing) transforming protein 1	X68148, R52960, T50498
Hs.1119	NR4A1	nuclear receptor subfamily 4, group A, member 1	NM_002135, N94487
Hs.1657	ESR1	estrogen receptor 1	AL078582, AA291702
Hs.166891	RFX5	regulatory factor X, 5 (influences HLA class II expression)	AL050135, AA418045
Hs.252229	MAFG	v-maf musculoaponeurotic fibrosarcoma (avian) oncogene family, protein G	AF059195, N21609
Hs.25647	FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	V01512, N36944, AA485377
Hs.431	BMI1	murine leukemia viral (bmi-1) oncogene homolog	AA884913, W90704, AA478036
Hs.75117	ILF2	interleukin enhancer binding factor 2, 45kD	AA307289, H95638, AA894687
Hs.75678	FOSB	FBJ murine osteosarcoma viral oncogene	L49169, T61948

		homolog B	
Hs.6551	ATP6IP1	ATPase, H+ transporting, lysosomal interacting protein 1	NM_001183, AA487588
Hs.79090	XPO1	exportin 1 (CRM1, yeast, homolog)	D89729, T59055
Hs.80917	AP3S1	adaptor-related protein complex 3, sigma 1 subunit	D63643, AA996044
Hs.80919	SYPL	synaptophysin-like protein	S72481, AA427447
Hs.44532	UBD	diubiquitin	NM_006398N3392 0
Hs.106061	RDBP	RD RNA-binding protein	X16105, AA056390
Hs.108636	C1orf9 CH1	chromosome 1 open reading frame 9 membrane protein CH1	BE466870, N36176
Hs.110713	DEK	DEK oncogene (DNA binding)	AI888504, R25377
Hs.11355	TMPO	Thymopoietin ESTs	U09087, T63980
Hs.16341	MAWBP	MAWD binding protein ESTs weakly similar to predicted using genefinder [C. elegans]	AI866254, R54416
Hs.16426	PODXL	podocalyxin-like	BE395330, N64508
Hs.18443	ALDH8A1	aldehyde dehydrogenase 8 family, member A1 ESTs	AI051566, N70701
Hs.194673	PEA15	phosphoprotein enriched in astrocytes 15	Y13736, AA293211
Hs.23767	FLJ12666	hypothetical protein FLJ12666 Homo sapiens cDNA FLJ12666 fis, clone NT2RM4002256	AW952494, H10192, AA115300, AA131466
Hs.291904	DXS1357E	accessory proteins BAP31/BAP29	Z31696, AA625628
Hs.3610	KIAA0205	KIAA0205 gene product	D86960, R91263
Hs.36102		ESTs, Highly similar to SMHU1B metallothionein 1B [H.sapiens] ESTs highly similar to MT1B Human Metallothionein-IB [H.sapiens]	R99207, H72722
Hs.6838	ARHE	ras homolog gene family, member E	W03441, AA443302
Hs.75187	TOMM20-PENDI	translocase of outer mitochondrial membrane 20 (yeast) homolog KIAA0016 translocase of outer mitochondrial membrane 20 (yeast) homolog	D13641, AA644550
Hs.8765	RNAHP	RNA helicase-related protein	AI814448, T56221, N55459

The discriminator cassettes were assessed on an independent tissue set of 58 liver clinical biopsies from 29 patients. Using a kNN prediction algorithm, it was found that all classifier probe cassettes could readily distinguish HCC tumor from non-tumor liver (Table 5), and that the gene discriminators of tumor vs. non-tumor in HCC derived by the intersect analysis of limited tissue sets can be validated in an independent manner.

Table 5. Prediction accuracy of gene classifiers using *k* NN algorithm on 58 liver biopsies from 29 patients.

Gene classifiers	No. of gene classifiers	Misclassification rate	No.of false negative cases*	No. of false positive cases**	Predictive accuracy
Table 1	218	4 of 58	4	-	93%
Table 2	265	3 of 58	3	-	95%
Table 3	166	3 of 58	3	-	95%
Table 4	68	2 of 58	1	1	96%

* False negative cases refer to HCC tumors which were misclassified as non-tumor livers.
**False positive cases refer to non-tumor livers which were misclassified as HCC tumors.

REFERENCES

The following publications are incorporated by reference herein.

Balsara, B. R., Pei, J., de Rienzo, A., Simon, D., Tosolini, A., Lu, Y. Y., Shen, F-M., Fan, X., Lin, W-Y., Buetow, K. H., London, W. T., and Testa, J. R. (2001). Human hepatocellular carcinoma is characterized by a highly consistent pattern of genomic imbalances, including frequent loss of 16q23.1-24.1. Gene Chrom. Cancer 30, 245-253.

Bea, S., Tort, F., Pinyol, M., Puig, X., Hernandez, L., Hernandez, S., Fernandez, P. L., van Lohuizen, M., Colomer, D., and Campo, E. (2001). BMI-1 gene amplification and overexpression in hematological malignancies occur mainly in mantle cell lymphomas. Cancer Res. 61, 2409-2412.

Bosl, G. J., and Head, M. D. (1994). Serum tumor marker half-life during chemotherapy in patients with germ cell tumors. Int. J. Biol. Markers 9, 25-28.

Brown, P. O., and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. Nat. Genet. 21 (suppl.), 33-37.

Carver, R. S., Sliwkowski, M. X., Sitaric, C., and Russell, W. E. (1996). Insulin regulates heregulin binding and ErbB3 expression in rat hepatocytes. J. Biol. Chem. 271, 13491-13496.

Chen, X., Cheung, S. T., So, S., Fan, S. T., Barry, C., Higgins, J., Lai, K-M.,

Ji, J., Dudoit, S., Ng, I. O. L., van de Rijn, M., Botstein, D., and Brown, P. O. (2002). Gene expression patterns in human liver cancers. Mol. Biol. Cell 13, 1929-1939.

Davison, A.C., and Hinkley, D. V. (1997). Bootstrap methods and their application (Cambridge: Cambridge University Press).

Eberwine, J., Yeh, H., Miyashiro K., Cao, Y., Nair, S., Finnell, R., Zettel, M., and Coleman, P. (1992). Analysis of gene expression in single live neurons. Proc. Natl. Acad. Sci. USA 89, 3010-3014.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863-14688.

Filmus, J., and Selleck, S. B. (2001). Glypicans: proteoglycans with a surprise. J. Clin. Invest. 108, 497-501.

Fink, L., Seeger, W., Ermert, L., Hanze, J., Stahl, U., Grimminger, F., Kummer, W., and Bohle, R. M. (1998). Real-time quantitative RT-PCR after laser-assisted cell picking. Nat. Med. 4, 1329-1333.

Gonzalez, F., Delahodde, A., Kodadek, T., and Johnston, S. A. (2002). Recruitment of a 19S proteasome subcomplex to an activated promoter. Science 296, 548-550.

Grimberg, A., and Cohen, P. (2000). Role of insulin-like growth factors and their binding proteins in growth control and carcinogenesis. J. Cell. Physiol. 183,1-9.

Gucev, Z. S., Oh, Y., Kelley K. M., and Rosenfeld, R. G. (1996) Insulin-like growth factor binding protein 3 mediates retinoic acid- and transforming growth factor β2-induced growth inhibition in human breast cancer cells. Cancer Res. 56, 1545-1550.

Honda, M., Kaneko, S., Kawa, H., Shirota, Y., and Kobayashi, K. (2001). Differential gene expression between chronic hepatitis B and C hepatic lesion. Gastroenterology 120, 955-966.

Huynh, H., Chow, P. K. H., Ooi, L. L. P., and Soo K-C. (2002). A possible role for insulin-like growth factor-binding protein-3 autocrine/paracrine loops in controlling hepatocellular carcinoma cell proliferation. Cell Growth Diff. 13, 115-122.

Irniger, S., and Nasmyth, K. (1997). The anaphase-promoting complex is required in G1 arrested yeast cells to inhibit B-tyoe cyclin accumulation and to prevent uncontrolled entry into S-phase. J. Cell Sci. 110, 1523-1531.

Ito, Y., Takeda, T., Sakon, M., Tsujimoto, M., Higashiyama, S., Noda, K., Miyoshi, E., Monden, M., and Matsuura, N. (2001). Expression and clinical significance of erb-B receptor family in hepatocellular carcinoma. Br. J. Cancer 84, 1377-1383.

Johnson, P. J. (2001). The role of serum alpha-fetoprotein estimation in the diagnosis and management of hepatocellular carcinoma. Clin. Liver Dis. 5, 145-159.

Lee, W. M. (1997). Hepatitis B virus infection. New Engl. J. Med. 337, 1733-1745.

Lee, J.-S., and Thorgeirsson, S. S. (2002). Functional and genomic implications of global gene expression profiles in cell lines from human hepatocellular cancer. Hepatology 35, 1134-1143.

Lessard, J., and Sauvageau, G. (2003). Bmi-1 determines the proliferative capacity of normal and leukaemic stem cells. Nature 423, 255-260.

Leung, T. H. Y., Wong, N., Lai, P. B. S., Chan, A., To, K-F., Liew, C. T., Lau, W-Y., and Johnson, P. J. (2002). Identification of four distinct regions of allelic imbalances on chromosome 1 by the combined comparative genomic hybridization and microsatellite analysis on hepatocellular carcinoma. Mod. Pathol. 15, 1213-1220.

Mallery, D. L., Vandenberg, C. J., and Hiom, K. (2002). Activation of the E3 ligase function of the BRCA1/BARD1 complex by polyubiquitin chains. EMBO J. 21, 6755-6762.

Marchio, A., Meddeb, M., Pineau, P., Danglot, G., Tiollais, P., Bernheim, A., and Dejean, A. (1997). Recurrent chromosomal abnormalities in hepatocellular carcinoma detected by comparative genomic hybridization. Genes Chrom. Cancer 18, 59-65.

Midorikawa, Y., Ishikawa, S., Iwanari, H., Imamura, T., Sakamoto, H., Miyazono, K., Kodama, T., Makuuchi, M., and Aburatani, H. (2003). Glypican-3 , overexpressed in hepatocellular carcinoma, modulates FGF2 and BMP-7 signaling. Int. J. Cancer 103, 455-465.

Murthy, S. S., Shen, T., de Rienzo, A., Lee, W. C., Ferriola, P. C., Jhanwar, S. C., Mossman, B. T., Filmus, J., and Testa, J. R. (2000). Expression of GPC3, an X-linked recessive overgrowth gene, is silenced in malignant mesothelioma. Oncogene 19, 410-416.

Nawrocki, S. T., Bruns, C. J., Harbison, M. T., Bold, R. J., Gotsch, B. S., Abbruzzese, J. L., Elliott, P., Adams, J., and McConkey, D. J. (2002). Effects of the proteasome inhibitor PS-341 on apoptosis and angiogenesis in orthotopic human pancreatic tumor xenografts. Mol. Cancer Therapeutics 1, 1243-1253.

Okabe, H., Satoh, S., Kato, T., Kitahara, O., Yanagawa, R., Yamaoka, Y.,

Tsunoda, T., Furukawa, Y., and Nakamura, Y. (2001). Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression. Cancer Res. 61, 2129-2137.

Orlando, V. (2003). Polycomb, epigenomes and control of cell identity. Cell, 112, 599-606.

Park, I-K., Qian, D., Kiel, M., Becker, M. W., Pihalja, M., Weissman, I. L., Morrison, J., and Clarke, M. F. (2003). Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells. Nature 423, 302-305.

Parkin, D. M., Pisani, P., and Ferlay, J. (1999). Global cancer statistics. CA Cancer J. Clin. 49, 33-64.

Pickart, C. M. (2001). Ubiquitin enters the new millennium. Mol. Cell. 8, 499-504.

Raaphorst, F. M., van Kemenade, F. J., Blokzijl, T., Fieret, E., Hamer, K. M., Satijn, D. P. E., Otte, A. P., and Meijer, C. J. L. M. (2000). Coexpression of BMI-1 and EZH2 polycomb group genes in Reed-Sternberg cells of Hodgkin's disease. Am. J. Pathol. 157, 709-715.

Riese, D. J. II, and Stern, D.F. (1998). Specificity within the EGF family/ErbB receptor family signaling network. BioEssays 20, 41-48.

Saikali, Z., and Sinnett, D. (2000). Expression of glypican 3 (GPC3) in embryonal tumors. Int. J. Cancer. (Pred Oncol) 89, 418-422.

Sakamoto, K. M. (2002). Ubiquitin-dependent proteolysis: its role in human diseases and the design of therapeutic strategies. Mol. Genet. Metab. 77, 44-56.

Salomoni, P., and Pnadolfi, P. P. (2002). p53 de-ubiquitination: at the edge

between life and death. Nat. Cell Biol. 4, E152-E153.

Salghetti, S., Caudy, A. A., Chenoweth, J. G., and Tansey, W. P. (2001). Regulation of transcriptional activation domain function by ubiquitin. Science 293, 1651-1653.

Shang, Y., Baumrucker, C. R., and Green, M. H. (1999). Signal relay by retinoic acid receptors á and â in the retinoic acid-induced expression of insulin-like growth factor-binding protein-3 in breast cancer cells. J. Biol. Chem. 274, 18005-18010.

Shirota, Y., Kaneko, S., Honda, M., Kawai, H. F., and Kobayashi, K. (2001). Identification of differentially expressed genes in hepatocellular carcinoma with cDNA microarrays. Hepatology 33, 832-840.

Sotiriou, C., Khanna, C., Jazaeri, A. A., Petersen, D., and Liu, E. T. (2002). Core biopsies can be used to distinguish differences in expression profiling by cDNA microarrays. J. Mol. Diag. 4, 30-36.

Tackels-Horne, D., Goodman, M. D., Williams, A. J., Wilson, D. J., Eskandari, T., Vogt, L. M., Boland, J. F., Scherf, U., and Vockley, J. G. (2001). Identification of differentially expressed genes in hepatocellular carcinoma and metastatic liver tumors by oligonucleotide expression profiling. Cancer 92, 395-405.

The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. Nature Genet. 25, 25-29. (World Wide Web URL: http://www.geneontology.org)

Toretsky, J. A., Zitomersky, N. L., Eskenazi, A. E., Voigt, R. W., Strauch, E. D. Sun, C. C., Huber, R., Meltzer, S. J., and Schlessinger, D. (2001). Glypican-3 expression in Wilms tumor and hepatoblastoma. J. Pediatr. Hematol. Oncol. 23, 496-499.

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kuma-Sinha, C., Sanda, M. G., Ghosh, D., Pienta, K. J., Sewalt, R. G. A. B., Otte, A. P., Rubin, M. A., and Chinnaiyan, A. M. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature 419, 624-629.

Vonlanthen, S., Heighway, J., Altermatt, H. J., Gugger, M., Kappeler, A., Borner, M. M., van Lohuizen, M., and Betticher, D. C. (2001). The bmi-1 oncoprotein is differentially expressed in non-small cell lung cancer and correlates with INK4A-ARF locus expression. Br. J. Cancer 84, 1372-1376.

Wong, N., Lai, P. Lee, S. W., Fan, S., Pang, E., Liew, C. T., Sheng, Z., Lau, J. W., and Johnson, P. J. (1999). Assessment of genetic changes in hepatocellular carcinoma by comparative genomic hybridization analysis: relationship to disease stage, tumor size, and cirrhosis. Am. J. Pathol. 154, 37-43.

Xiang, Y. Y., Ladeda, V., and Filmus, J. (2001). Glypican-3 expression is silenced in human breast cancer. Oncogene 20, 7408-7412.

Xu, L., Hui, L., Wang, S., Gong, J., Jin, Y., Wang, Y., Ji, Y., Wu, X., Han, Z., and Hu, G. (2001a). Expression profiling suggested a regulatory role of liverenriched transcription factors in human hepatocellular carcinoma. Cancer Res. 61, 3176-3781.

Xu, X-R., Huang, J., Xu, Z-G., Qian, B-Z., Zhu, Z-D., Yan, Q., Cai, T., Zhang, X., Xiao, H-S., Qu, J., Liu, F., Huang, Q-H., Cheng, Z-H., Li, N-G., Du. J-J., Hu, W., Shen, K-T., Lu, G., Fu, G., Zhong, M., Xu, S-H., Gu, W-Y., Huang, W., Zhao, X-T., Hu, G-X., Gu, J-R., Chen, Z., and Han, Z-G. (2001b). Insight into hepatocellular carcinogenesis at transcriptome level by comparing gene expression profiles of hepatocellular carcinoma with those of corresponding noncancerous liver. Proc. Natl. Acad. Sci. USA 98, 15089-15094.

Yu, A. S., and Keeffe, E. B. (2003). Management of hepatocellular carcinoma. Rev. Gastroenterol. Disord. 3, 8-24.

Zhang, N., and Deuel, T. F. (1999). Pleitrophin and midkine, a family of mitogenic and angiogenic heparin-binding growth and differentiation factors. Curr. Opin. Hematol. 6, 44-50.

Zimonjic, D. B., Keck, C. L., Thorgeirsson, S. S., and Popescu, N. C. (1999). Novel recurrent genetic imbalances in human hepatocellular carcinoma cell lines identified by comparative genomic hybridization. Hepatology 29, 1208-1214.

Cheung V. G., *et al.*, 1999. *Nature Genetics Supplement*, **21**:15-19.

Lipshutz R. J., *et al.*,1999. *Nature Genetics Supplement*, **21**:20-24.

Bowtell D. D. L., 1999. *Nature Genetics Supplement*, **21**:25-32.

Schweitzer B., *et al.*, 2002. *Nature Biotechnol.*, **20**:359-365.

Halloway A. J., *et al.*, 2002. *Nature Genetics Supplement*, **32**:481-489.

Martzen M. R., McCraith S. M., Spinelli S.L., Torres F. M., Fields S., Grayhack E.J., and Phizicky E. M., 1999. *Science*, **286**:1153.

G. MacBeath and S. L. Schreiber, "Printing Proteins as Microarrays for High-Throughput Function Determination," Science 289(5485):1760-1763, 2000.

ACKNOWLEDGEMENT

Chon Kar Leow contributed to the studies by providing HCC liver tissue samples to the inventors.

OTHER EMBODIMENTS

Although various embodiments of the invention are disclosed herein, many adaptations and modifications may be made within the scope of the invention in accordance with the common general knowledge of those skilled in this art. Such modifications include the substitution of known equivalents for any aspect of the invention in order to achieve the same result in substantially the same way. Accession numbers, as used herein, may refer to Accession numbers from multiple databases, including GenBank, the European Molecular Biology Laboratory (EMBL), the DNA Database of Japan (DDBJ), or the Genome Sequence Data Base (GSDB), for nucleotide sequences, and including the Protein Information Resource (PIR), SWISSPROT, Protein Research Foundation (PRF), and Protein Data Bank (PDB) (sequences from solved structures), as well as from translations from annotated coding regions from nucleotide sequences in GenBank, EMBL, DDBJ, or RefSeq, for polypeptide sequences. Accession numbers, as used herein, may also refer to Accession numbers from databases such as UniGene, OMIM, LocusLink, or HomoloGene. Numeric ranges are inclusive of the numbers defining the range. In the specification, the word "comprising" is used as an open-ended term, substantially equivalent to the phrase "including, but not limited to", and the word "comprises" has a corresponding meaning. Citation of references herein shall not be construed as an admission that such references are prior art to the present invention. All publications are incorporated herein by reference as if each individual publication were specifically and individually indicated to be incorporated by reference herein and as though fully set forth herein. The invention includes all embodiments and variations substantially as hereinbefore described and with reference to the examples and drawings.

WHAT IS CLAIMED IS:

1. A composition comprising an addressable collection of two or more nucleic acid molecules that are differentially expressed in hepatocellular carcinoma, wherein the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4 or complements, fragments, variants, or analogs thereof.

2. The composition of claim 1 wherein the nucleic acid molecules consist essentially of all the nucleic acid molecules set forth in any one or more of Tables 1 through 4.

3. The composition of claim 1 or 2 wherein the nucleic acid molecules are differentially expressed between hepatocellular carcinoma tissue and non-tumor tissue.

4. A composition comprising an addressable collection of two or more polypeptides that are differentially expressed in hepatocellular carcinoma, wherein the polypeptides consist essentially of polypeptides encoded by the nucleic acid molecules set forth in any one or more of Tables 1 through 4 or complements, fragments, variants, or analogs thereof.

5. The composition of claim wherein the polypeptides consist essentially of the polypeptides encoded by all the nucleic acid molecules set forth in any one or more of Tables 1 through 4.

6. The composition of claim 4 or 5 wherein the polypeptides are differentially expressed between hepatocellular carcinoma tissue and non-tumor tissue.

7. The composition of any one of claims 1 through 6 wherein the nucleic acid molecules or the polypeptides are attached to a solid support.

8. Use of the composition of any one of claims 1 through 7 in the preparation of a medicament for diagnosis or therapy of hepatocellular carcinoma.

9. A method of diagnosing hepatocellular carcinoma in a subject comprising:

obtaining a sample from the subject; and

detecting the level of expression of two or more nucleic acid molecules or expression products thereof in the sample, wherein the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4 or complements, fragments, variants, or analogs thereof.

10. The method of claim 9 wherein the sample is liver.

11. The method of claim 9 wherein the sample is serum.

12. The method of any one of claims 9 through 11 wherein the sample is suspected of being cancerous.

13. The method of any one of claims 9 through 11 wherein the sample is non-cancerous.

14. The method of any one of claims 9 through 13 further comprising comparing the level of expression of the nucleic acid molecules or expression products thereof in a non-cancerous sample and in a sample suspected of being cancerous.

15. The method of any one of claims 9 through 14 wherein the subject is suspected of having hepatocellular carcinoma.

16. The method of any one of claims 9 through 15 wherein the nucleic acid molecules consist essentially of all the nucleic acid molecules set forth in any one or more of Tables 1 through 4 or complements, fragments, or analogs thereof.

17. The method of any one of claims 9 through 17 wherein differential expression of the nucleic acid molecules or expression products thereof is indicative of hepatocellular carcinoma.

18. A method of monitoring the progression of hepatocellular carcinoma in a subject comprising:

obtaining a sample from the subject; and

detecting the level of expression of two or more nucleic acid molecules or expression products thereof in the sample, wherein the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4, or complements, fragments, variants, or analogs thereof.

19. The method of claim 18 wherein the sample is obtained at two or more time points.

20. The method of claim 18 or 19 wherein the nucleic acid molecules consist essentially of all the nucleic acid molecules set forth in any one or more of Tables 1 through 4 or complements, fragments, variants, or analogs thereof.

21. The method of any one of claims 18 through 20 wherein the sample is liver.

22. The method of any one of claims 18 through 20 wherein the sample is serum.

23. The method of any one of claims 18 through 22 further comprising comparing the level of expression of the nucleic acid molecules or expression products thereof in a non-cancerous sample and in a sample suspected of being cancerous.

24. The method of any one of claims 18 through 23 wherein differential expression of the nucleic acid molecules or expression products thereof is indicative of progression of hepatocellular carcinoma.

25. The method of any one of claims 19 through 24 further comprising comparing the level of expression of the nucleic acid molecules or expression products in the two or more time points.

26. A method of monitoring the efficacy of a hepatocellular carcinoma therapy in a subject comprising:

administering the therapy to the subject;

obtaining a sample from the subject; and

detecting the level of expression of two or more nucleic acid molecules or expression products thereof in the sample, wherein the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4, or complements, fragments, variants, or analogs thereof.

27. The method of claim 26 wherein the therapy is administered at two or more administration time points.

28. The method of claim 26 or 27 wherein the sample is obtained at two or more sampling time points.

29. The method of any one of claims 26 through 28 wherein the nucleic acid molecules consist essentially of all the nucleic acid molecules set forth in any one or more of Tables 1 through 4 or complements, fragments, variants or analogs thereof.

30. The method of any one of claims 26 through 29 wherein the sample is suspected of being cancerous.

31. The method of any one of claims 26 through 30 wherein the sample is non-cancerous.

32. The method of any one of claims 26 through 31 further comprising comparing the level of expression of the nucleic acid molecules or expression products thereof in a non-cancerous sample and in a sample suspected of being cancerous.

33. The method of any one of claims 26 through 32 wherein differential expression of the nucleic acid molecules or expression products thereof is indicative of the efficacy of the hepatocellular carcinoma therapy.

34. The method of any one of claims 27 through 33 further comprising comparing the level of expression of the nucleic acid molecules or expression products in the two or more administration time points.

35. The method of any one of claims 28 through 34 further comprising comparing the level of expression of the nucleic acid molecules or expression products in the two or more sampling time points.

36. The method of any one of claims 9 through 35 wherein the subject is a human.

37. A method of screening a compound for treating hepatocellular carcinoma comprising:

contacting a sample with a test compound; and

detecting the level of expression of two or more nucleic acid molecules or expression products thereof in the sample, wherein the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4, or complements, fragments, variant or analogs thereof.

38. The method of claim 37 wherein the nucleic acid molecules consist essentially of all the nucleic acid molecules set forth in any one or more of Tables 1 through 4 or complements, fragments, variants or analogs thereof.

39. The method of claim 37 or 38 wherein the samples are selected from the group consisting of cancerous samples and non-cancerous samples.

40. The method of any one of claims 37 through 39 wherein the samples are selected from the group consisting of liver samples and serum samples.

41. The method of any one of claims 37 through 10, further comprising comparing the level of expression of two or more nucleic acid molecules or expression products thereof with a standard.

42. The method of any one of claims 9 through 41, wherein the method is a high throughput method.

43. The method of any one of claims 9 through 42, further comprising preparing a gene expression profile.

44. A solid support comprising two or more nucleic acid molecules that are differentially expressed in hepatocellular carcinoma, wherein the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in Tables 1 through 4 or complements, fragments, variants, or analogs thereof.

45. The solid support of claim 44 wherein the nucleic acid molecules consist essentially of all the nucleic acid molecules set forth in any one or more of Tables 1 through 4.

46. The solid support of claim 44 or 45 wherein the nucleic acid molecules are differentially expressed between hepatocellular carcinoma tissue and non-tumor tissue.

47. A solid support comprising two or more polypeptides that are differentially expressed in hepatocellular carcinoma, wherein the polypeptides consist essentially of polypeptides encoded by the nucleic acid molecules set forth in Tables 1 through 4 or complements, fragments, variants, or analogs thereof.

48. The solid support of claim 47 wherein the polypeptides consist essentially of the polypeptides encoded by all the nucleic acid molecules set forth in any one or more of Tables 1 through 4.

49. The solid support of claim 47 or 48 wherein the polypeptides are differentially expressed between hepatocellular carcinoma tissue and non-tumor tissue.

50. The solid support of any one of claims 47 through 49 wherein the nucleic acid molecules or the polypeptides are covalently attached to the solid support.

51. The solid support of any one of claims 47 through 49 wherein the nucleic acid molecules or the polypeptides are non-covalently attached to the solid support.

52. The solid support of any one of claims 47 through 51 wherein the solid support comprises a microarray.

53. A database comprising information identifying the expression level in liver tissue of two or more nucleic acid molecules or expression products thereof, wherein the nucleic acid molecules consist essentially of the nucleic acid molecules set forth in any one or more of Tables 1 through 4, or complements, fragments, variants, or analogs thereof.

54. The database of claim 53, wherein the liver tissue is selected from the group consisting of cancerous and non-cancerous tissue.