

improving and refining the already established classification of various types of cancer, and generally do not provide new insights into the relationships of the differentially expressed genes, and do not link the findings to treatment strategies in order to improve the clinical outcome of cancer therapy. In addition, cancer treatment and colon cancer clinical trials are still being pursued on the basis of the availability of new active compounds rather than the integrated approach of pharmacogenomics, which utilizes the genetic makeup of the tumor and the genotype of the patient to establish a personalized medication regime.

Although modern molecular biology and biochemistry have revealed more than 100 genes whose activities influence the behavior of tumor cells, state of their differentiation, and their sensitivity or resistance to certain therapeutic drugs, with a few exceptions, the status of these genes has not been exploited for the purpose of routinely making clinical decisions about drug treatments.

SUMMARY OF THE INVENTION

There is a need to identify biomarkers useful for predicting prognosis of patients with colon cancer. The ability to classify patients as high risk (poor prognosis) or low risk (favorable prognosis) would enable selection of appropriate therapies for these patients. For example, high-risk patients are likely to benefit from aggressive therapy, whereas therapy may have no significant advantage for low risk patients. However, in spite of this need, a solution to this problem has not been available.

Therefore, microarray-based prognostic technologies are needed that provide a physician with information on the likelihood of recovery or relapse following administration of a particular treatment regimen, such as resection with or without chemotherapy. Technologies are also needed that can accurately diagnose a colon disease, particularly the diagnosis of a particular stage of colon cancer, or can predict a colon disease patient's response to a particular therapy. Specific knowledge regarding a tumor in a cancer patient would be extremely useful in prolonging remission, increasing the quality of patient life, and reducing healthcare costs. Such technologies may also be used to screen patient candidates for clinical trials for novel therapeutic compounds and methods to facilitate the regulatory approval process.

Disclosed are expression signatures from colon cancer that meet these needs. The disclosed signatures can be used for applications in prognosis of colon cancer,

diagnosis of colon cancer and classifying patient groups. In some
embodimentsexamples, these results permit assessment of genomic evidence of the
efficacy of surgery alone, or in combination with adjuvant chemotherapy for
treatment of colon cancer. The signatures described herein may be significant in, and
5 capable of, discriminating between two diagnoses or prognostic outcomes. An
important aspect of the present disclosure is to use the measured expression of certain
genes in colon cancer tissue to match patients to the most appropriate treatment, and
to provide prognostic information. Thus, disclosed are methods of using such colon
cancer signatures. Therefore, in one aspect, the present invention refers to a method
10 for diagnosing colon cancer in a sample obtained from a subject, comprising:
detecting an expression level of the first 634 colon cancer-related nucleic acid
molecules listed in Table 6 in a sample comprising nucleic acids obtained from a
subject; and comparing the expression level of the first 634 colon cancer-related
nucleic acid molecules, or a decision score derived therefrom, to a control threshold
15 indicative of a diagnosis of colon cancer, wherein the expression level, or a decision
score derived therefrom, on the same side of the threshold of the known diagnosis
control indicates a diagnosis of colon cancer, thereby diagnosing colon cancer in the
sample obtained from the subject. In another aspect, the present invention refers to a
method for classifying a colon cancer sample, comprising: detecting an expression
20 level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a
sample comprising nucleic acids obtained from a subject; and comparing the
expression level of the first 634 colon cancer-related nucleic acid molecules, or a
decision score derived therefrom to a control threshold indicative of known
classification, wherein the expression level, or a decision score derived therefrom, on
25 the same side of the threshold of the known classification control permits
classification of the colon cancer sample. In still another aspect, the present invention
refers to a method for predicting a response to a treatment for colon cancer,
comprising: detecting an expression level of the first 634 colon cancer-related nucleic
acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a
30 subject; and comparing the expression level of the first 634 colon cancer-related
nucleic acid molecules, or a decision score derived therefrom to a control threshold
indicative of a known response to treatment, wherein the expression level, or a
decision score derived therefrom, on the same side of the threshold of the known
response control indicates a similar response to treatment, thereby predicting response

to treatment. In another aspect, the present invention refers to a method for predicting long term survival of a subject with colon cancer, comprising: detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and comparing the expression level of the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of having a history of long term survival, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold of the known long term survival control indicates long term survival of the subject, thereby predicting long term survival of a subject. In yet another aspect, the present invention refers to a method for predicting of recurrence of colon cancer in a subject, comprising: detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and comparing the expression level of the the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of a history of recurrence, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold of the known recurrence control indicates a recurrence in the subject. In another aspect, the present invention refers to a method of preparing a personalized colon cancer genomics profile for a subject, comprising: detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and creating a report summarizing the data obtained by the gene expression analysis. In another aspect, the present invention refers to a method for preparing a gene expression profile indicative of colon cancer prognosis, comprising: detecting the expression level of less than 1000 transcripts in a sample comprising RNA isolated from a colon cancer specimen, wherein the first 634 transcripts listed in Table 6 are detected. In yet another aspect, the present invention refers to a method for prognosing colon cancer, comprising: preparing a gene expression profile for a colon cancer specimen comprising isolated RNA; and classifying the specimen based the expression levels of the first 634 transcripts listed in Table 6, or a decision score derived therefrom, in a low risk or high risk group. ~~The disclosed m~~Other methods disclosed herein include detecting an expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject and comparing the expression level of the at least 2 colon cancer-related nucleic acid

molecules, or a decision score derived therefrom to a control threshold. Depending of the prediction requested, the control threshold can be indicative of a diagnosis of colon cancer, indicative of known classification of colon cancer, indicative of a known response to treatment, indicative of having a history of long term survival,
5 indicative of a history of recurrence and the like.

In various ~~embediment~~examples, RNA is isolated from a colon tissue sample, and used for preparing a gene expression profile. In certain ~~example~~embediments involving prognosis of cancer, the sample is a colorectal tumor specimen, such as a colon cancer sample. In certain ~~example~~embediments, the gene expression profile
10 involves detecting the expression of at least 50 transcripts listed in Table 6, and which may also be listed in Table 1 and/or Table 2. The total number of transcripts detected in the gene expression profile can vary. For example, in some ~~example~~embediments the total number of transcripts detected in the profile is from about 200 to about 1000, or from about 400 to about 800, or in other ~~example~~embediments, the number of
15 transcripts is from about 500 to about 700, or from about 550 to about 650. In various ~~example~~embediments, at least about 50, at least about 100, at least about 200, at least about 300, at least about 400, at least about 500, at least about 600, or all transcripts, listed in Table 6 are detected as part of the total number of transcripts. Where additional transcripts are detected (in addition to those of Table 6), they may be
20 optionally selected from signal or expression level controls, and in some ~~example~~embediments, are transcripts known to be expressed in colon cancer, such as those determined by *Colorectal Cancer DSA*TM. In certain ~~example~~embediments, the additional transcripts may also be indicative of colon cancer prognosis.

The patient's expression profile is scored against an expression signature based on expression levels of the transcripts listed in Table 6 in high risk and low risk patient groups, such as patient with a high or low risk of clinical relapse, and the results may be used to determine a course of treatment. For example, a patient
5 determined to be a high risk patient may be treated with adjuvant chemotherapy after surgery. For a patient deemed to be a low risk patient, adjuvant chemotherapy may be withheld after surgery. Accordingly, the ~~invention-disclosure~~ provides, in certain aspects, a method for preparing a gene expression profile of a colon cancer tumor that is indicative of risk of recurrence.

10 The disclosure further provides a method for prognosing colon cancer. The method according to this aspect comprises preparing a gene expression profile of a colon cancer specimen (*e.g.*, as described herein). The gene expression profile is then classified or scored against a gene expression signature described herein. In various
15 ~~examples~~~~embodiments~~, the gene expression signature is based on the expression level of at least 50 transcripts listed in Table 6, and which may also be listed in Table 1 and/or Table 2. In some ~~examples~~~~embodiments~~, the total number of transcripts on which the signature is based is less than about 800, less than about 700, less than about 600, less than about 500, less than about 400, less than about 300, less than about 200, or less than about 100 transcripts, and which includes transcripts from
20 Table 6. For example, the signature may be based on the expression levels of at least about 400, at least about 500, or at least about 600 transcripts from Table 6. Optionally, the transcripts from Table 6 include the transcripts listed in Table 1.

Also disclosed are methods of preparing a personalized colon cancer genomics profile for a subject. The methods include detecting an expression level of at least 2
25 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject and creating a report summarizing the data obtained by the gene expression analysis.

In some examples, of the disclosed methods, expression levels are determined from nucleic acids obtained from the subject that comprise RNA and/or cDNA
30 transcribed from RNA extracted from a sample of colorectal tissue obtained from the subject, such as colon cancer sample.

Also disclosed are nucleic acid probes and primers (as well as sets of such probes and primers) for detecting a gene expression signature for colon cancer. In

The singular terms “a,” “an,” and “the” include plural referents unless context clearly indicates otherwise. Similarly, the word “or” is intended to include “and” unless the context clearly indicates otherwise. The term “comprises” means “includes.” In case of conflict, the present specification, including explanations of terms, will control.

To facilitate review of the various ~~embodiment~~examples of this disclosure, the following explanations of terms are provided:

Amplifying a nucleic acid molecule: To increase the number of copies of a nucleic acid molecule, such as a gene or fragment of a gene, for example a transcript shown in Table 6. The resulting products are called amplification products.

An example of *in vitro* amplification is the polymerase chain reaction (PCR). Other examples of *in vitro* amplification techniques include quantitative real-time PCR, strand displacement amplification (see U.S. Patent No. 5,744,311); transcription-free isothermal amplification (see U.S. Patent No. 6,033,881); repair chain reaction amplification (see International Patent Publication No. WO 90/01069); ligase chain reaction amplification (see EP-A-320 308); gap filling ligase chain reaction amplification (see U.S. Patent No. 5,427,930); coupled ligase detection and PCR (see U.S. Patent No. 6,027,889); and NASBA™ RNA transcription-free amplification (see U.S. Patent No. 6,025,134).

Array: An arrangement of molecules, such as biological macromolecules (such nucleic acid molecules) or biological samples (such as tissue sections), in addressable locations on or in a substrate. In some examples an array is an array of polynucleotide probes (such as probes that hybridize to the nucleic acids sequences shown in Table 6, or the complement thereof), bound to a solid substrate so as not to be substantially dislodged during a hybridization procedure. A “microarray” is an array that is miniaturized so as to require or be aided by microscopic examination for evaluation or analysis. Arrays are sometimes called DNA chips or biochips.

The array of molecules (“features”) makes it possible to carry out a very large number of analyses on a sample at one time. In certain example arrays, one or more molecules (such as an oligonucleotide probe) will occur on the array a plurality of times (such as twice), for instance to provide internal controls.

In particular examples, an array includes nucleic acid molecules, such as oligonucleotide sequences. The polynucleotides used on an array may be cDNAs (“cDNA arrays”) that are typically about 500 to 5000 bases long, although shorter or

observing a change in light absorption of a solution containing an oligonucleotide (or an analog) and a target nucleic acid at 220 to 300 nm as the temperature is slowly increased. If the oligonucleotide or analog has bound to its target, there is a sudden increase in absorption at a characteristic temperature as the oligonucleotide (or analog) and target disassociate from each other, or melt. In another example, the method involves detecting a signal, such as a detectable label, present on one or both nucleic acid molecules (or antibody or protein as appropriate).

The binding between an oligomer and its target nucleic acid is frequently characterized by the temperature (T_m) at which 50% of the oligomer is melted from its target. A higher (T_m) means a stronger or more stable complex relative to a complex with a lower (T_m).

cDNA (complementary DNA): A piece of DNA lacking internal, non-coding segments (introns) and regulatory sequences which determine transcription. cDNA can be synthesized by reverse transcription from messenger RNA (mRNA) extracted from cells and/or tissue samples, such as colon samples, including colon cancer samples.

Clinical outcome: Refers to the health status of a patient following treatment for a disease or disorder, or in the absence of treatment. Clinical outcomes include, but are not limited to, an increase in the length of time until death, a decrease in the length of time until death, an increase in the chance of survival, an increase in the risk of death, survival, disease-free survival, chronic disease, metastasis, advanced or aggressive disease, disease recurrence, death, and favorable or poor response to therapy.

Colon cancer: Cancer that forms in the tissues of the colon (the longest part of the large intestine). Most colon cancers are adenocarcinomas (cancers that begin in cells that make line internal organs and have gland-like properties). Cancer progression is characterized by stages, or the extent of cancer in the body. Staging is usually based on the size of the tumor, whether lymph nodes contain cancer, and whether the cancer has spread from the original site to other parts of the body. Stages of colon cancer include stage I, stage II, stage III and stage IV. Unless otherwise specified, the term colon cancer refers to colon cancer at Stage 0, Stage I, Stage II (including Stage IIA or IIB), Stage III (including Stage IIIA, IIIB or IIIC), or Stage IV. In some ~~embodiment~~examples herein, the colon cancer is from any stage. In other ~~embodiment~~examples, the colon cancer is a stage II colon cancer.

Chemotherapeutic agents: Any chemical agent with therapeutic usefulness in the treatment of diseases characterized by abnormal cell growth. Such diseases include tumors, neoplasms, and cancer as well as diseases characterized by hyperplastic growth such as psoriasis. In one embodimentexample, a chemotherapeutic agent is an agent of use in treating colon cancer. In one embodimentexample, a chemotherapeutic agent is a radioactive compound. One of skill in the art can readily identify a chemotherapeutic agent of use (see for example, Slapak and Kufe, *Principles of Cancer Therapy*, Chapter 86 in Harrison's Principles of Internal Medicine, 14th edition; Perry *et al.*, *Chemotherapy*, Ch. 17 in Abeloff, *Clinical Oncology 2nd ed.*, 2000 Churchill Livingstone, Inc; Baltzer and Berkery. (eds): *Oncology Pocket Guide to Chemotherapy*, 2nd ed. St. Louis, Mosby-Year Book, 1995; Fischer Knobf, and Durivage (eds): *The Cancer Chemotherapy Handbook*, 4th ed. St. Louis, Mosby-Year Book, 1993). Chemotherapeutic agents used for treating colon cancer include small molecules such as 5-fluorouracil, leucovorin, irinotecan, oxaliplatin, and capecitabine, and antibodies such as bevacizumab and cetuximab. Combination chemotherapy is the administration of more than one agent to treat cancer.

Contacting: Placement in direct physical association; includes both in solid and liquid form. Contacting includes contact between one molecule and another molecule, for example; contacting a sample with a nucleic acid probe, such as a probe for any of the sequences shown in Table 6.

Control: A "control" refers to a sample or standard used for comparison with an experimental sample, such as a tumor sample obtained from a patient with colon cancer. In some embodimentexamples, the control is a sample obtained from a healthy patient or a non-cancerous tissue sample obtained from a patient diagnosed with colon cancer, such as a non-cancerous tissue sample from the same organ in which the tumor resides (*e.g.*, non-cancerous colon tissue can serve as a control for a colon cancer). In some embodimentexamples, the control is a historical control or standard value (*i.e.*, a previously tested control sample or group of samples that represent baseline or normal values).

Controls or standards for comparison to a sample, for the determination of differential expression, include samples believed to be normal (in that they are not altered for the desired characteristic, for example a sample from a subject who does not have colon cancer) as well as laboratory values, even though possibly arbitrarily set. Laboratory standards and values may be set based on a known or determined

least 6, at least 7, at least 9, at least 10 or at least 11 and so on. In some
| embodimentexamples, the profile comprises at least about 200 genes (or “transcripts”)
and up to about 1000 transcripts, such as from about 400 transcripts to about 800
transcripts, or about 500 transcripts to about 700 transcripts. The profile comprises
5 transcripts from Table 6 (e.g., at least 100, at least 200, at least 300, at least 400, at
least 500, or at least 600 transcripts from Table 6), including in some
| embodimentexamples the 636 transcripts listed in Table 6. As used herein, the term
“gene” refers to an expressed transcript, which may be a characterized gene, or may
| be an expressed transcript such as an EST. In some embodimentexamples, the
10 detection platform is a microarray, and each probe is considered as determining the
expression of a separate “gene” or “transcript.”

A gene expression profile (also referred to as a **fingerprint** or **signature**) can
be linked to a tissue or cell type (such as colon tissue), to a particular stage of normal
tissue growth or disease progression (such as colon cancer), or to any other distinct or
15 identifiable condition that influences gene expression in a predictable way. Gene
expression profiles can include relative as well as absolute expression levels of
specific genes, and can be viewed in the context of a test sample compared to a
baseline or control sample profile (such as a sample from a subject who does not have
colon cancer). In one example, a gene expression profile in a subject is read on an
20 array (such as a nucleic acid array).

Hybridization: To form base pairs between complementary regions of two
strands of DNA, RNA, or between DNA and RNA, thereby forming a duplex
molecule, for example a duplex formed between a probe and any of the nucleic acid
sequences shown in Table 6 or the complement thereof. Hybridization conditions
25 resulting in particular degrees of stringency will vary depending upon the nature of
the hybridization method and the composition and length of the hybridizing nucleic
acid sequences. Generally, the temperature of hybridization and the ionic strength
(such as the Na⁺ concentration) of the hybridization buffer will determine the
stringency of hybridization. Calculations regarding hybridization conditions for
30 attaining particular degrees of stringency are discussed in Sambrook *et al.*, (1989) Molecular
Cloning, second edition, Cold Spring Harbor Laboratory, Plainview, NY (chapters 9 and 11).
The following is an exemplary set of hybridization conditions and is not limiting:

Very High Stringency (detects sequences that share at least 90% identity)

Hybridization: 5x SSC at 65°C for 16 hours

point on a nucleic acid sequence. In general, at least one forward and one reverse primer are included in an amplification reaction.

Nucleic acid probes and primers can be readily prepared based on the nucleic acid molecules provided herein, for example, by using computer programs intended for that purpose such as Primer (Version 0.5, © 1991, Whitehead Institute for Biomedical Research, Cambridge, MA) or PRIMER EXPRESS® Software (Applied Biosystems, AB, Foster City, CA).

Further guidelines for PCR primer and probe design may be found in Dieffenbach *et al.*, General Concepts for PCR Primer Design in: PCR Primer, A Laboratory Manual, Cold Spring Harbor Laboratory Press, New York, 1995, pp. 133 155; Innis and Gelfand, Optimization of PCRs in: PCR Protocols, A Guide to Methods and Applications, CRC Press, London, 1994, pp. 5 11; and Plasterer, Primerselect: Primer and probe design. *Methods Mol. Biol.* 70:520 527, 1997.

Prognosis: The likelihood of the clinical outcome for a subject afflicted with a specific disease or disorder. With regard to cancer, the prognosis is a representation of the likelihood (probability) that the subject will survive (such as for one, two, three, four or five years) and/or the likelihood (probability) that the tumor will metastasize. The term "prediction" is used herein to refer to the likelihood that a patient will respond either favorably or unfavorably to a drug or set of drugs, and also the extent of those responses. The predictive methods of the present ~~invention-disclosure~~ can be used clinically to make treatment decisions by choosing the most appropriate treatment modalities for any particular patient. The predictive methods of the present disclosure are valuable tools in predicting if a patient is likely to respond favorably to a treatment regimen, such as surgical intervention, chemotherapy with a given drug or drug combination, and/or radiation therapy.

Purified: The term "purified" does not require absolute purity; rather, it is intended as a relative term. Thus, for example, a purified oligonucleotide preparation is one in which the oligonucleotide is more pure than in an environment including a complex mixture of oligonucleotides.

Sample: A biological specimen containing genomic DNA, RNA (including mRNA and microRNA), protein, or combinations thereof, obtained from a subject. Examples include, but are not limited to, peripheral blood, urine, saliva, tissue biopsy, aspirate, surgical specimen, and autopsy material, and includes fixed and/or paraffin embedded samples. In one example, a sample includes a biopsy of a colon (such as

RNA transcripts of unknown function. In some embodiment examples, a transcript is a nucleic acid sequence shown in Table 6.

Therapeutic: A generic term that includes both diagnosis and treatment.

Treatment: Includes both therapeutic treatment and prophylactic or preventative measures, wherein the object is to prevent or slow down (lessen) the targeted pathologic condition or disorder. Those in need of treatment include those already with the disorder as well as those prone to have the disorder or those in whom the disorder is to be prevented. In tumor (*e.g.* cancer) treatment, a treatment such as surgery, chemotherapy or radiation may directly decrease the pathology of tumor cells, or render the tumor cells more susceptible to further treatment.

Tumor, neoplasia, malignancy or cancer: Neoplastic cell growth and proliferation, whether malignant or benign, and all pre-cancerous and cancerous cells and tissues and the result of abnormal and uncontrolled growth of cells. The terms "cancer" and "cancerous" refer to or describe the physiological condition in mammals that is typically characterized by unregulated cell growth. Neoplasia, malignancy, cancer and tumor are often used interchangeably and refer to abnormal growth of a tissue or cells that results from excessive cell division. The amount of a tumor in an individual is the "**tumor burden**" which can be measured as the number, volume, or weight of the tumor. A tumor that does not metastasize is referred to as "**benign**." A tumor that invades the surrounding tissue and/or can metastasize is referred to as "**malignant**." A "**non-cancerous tissue**" is a tissue from the same organ wherein the malignant neoplasm formed, but does not have the characteristic pathology of the neoplasm. Generally, noncancerous tissue appears histologically normal. A "**normal tissue**" is tissue from an organ, wherein the organ is not affected by cancer or another disease or disorder of that organ. A "**cancer-free**" subject has not been diagnosed with a cancer of that organ and does not have detectable cancer.

The "pathology" of cancer includes all phenomena that compromise the well-being of the patient. This includes, without limitation, abnormal or uncontrollable cell growth, metastasis, interference with the normal functioning of neighboring cells, release of cytokines or other secretory products at abnormal levels, suppression or aggravation of inflammatory or immunological response, neoplasia, premalignancy, malignancy, invasion of surrounding or distant tissues or organs, such as lymph nodes, etc.

samples are correlated with known patient response or clinical outcomes. For example, sensitive methods are also provided to predict patient response to, and prognosis after, treatment for colon cancer, such as surgical resection and/or chemotherapy. Generally, historical patient population data and tissue samples are analyzed to create genetic profiles for patients having a past history of colon cancer. In some ~~embodiment~~examples, the genetic profile of a patient sample is converted to a decision score. The clinical outcomes of each patient are correlated to the genetic profile, or decision score derived mathematically from the genetic profile for each patient's individual cancer.

In some ~~embodiment~~examples, a mathematical algorithm is generated using the known historical patient data and applied to the predictive methods for new patients with colon cancer. In some ~~embodiment~~example, the algorithm creates a threshold that separates two groups of patients depending on selection criteria, for example patient outcome, response to therapy and recurrence, and the like. In some examples, the mathematical algorithm or threshold is validated using further historical patient population data before being used in the predictive methods described herein. The mathematical algorithm or threshold may then be used as a reference, for example as a control, to compare decision scores derived from genetic profiling of patients desirous of predictive methods of colon cancer. In some ~~embodiment~~examples, these results permit assessment of genomic evidence of the efficacy of surgery alone, or in combination with adjuvant chemotherapy for treatment of colon cancer.

The signatures described herein may be significant in, and capable of, discriminating between two diagnoses or prognostic outcomes. An important aspect of the present disclosure is to use the measured expression of certain genes in colon cancer tissue to match patients to the most appropriate treatment, and to provide prognostic information.

In some ~~embodiment~~examples, the signatures are developed using a colorectal cancer-focused microarray research tool. In a specific ~~embodiment~~example, this research tool is a colorectal cancer transcriptome-focused research array developed by Almac Diagnostics, Ltd. (Almac Diagnostics, Ltd., N. Ireland) capable of delivering accurate expression data.

The *Colorectal Cancer DSA*TM research tool contains 61,528 probe sets and encodes 52,306 transcripts confirmed as being expressed in colon cancer and normal tissue. Comparing the *Colorectal Cancer DSA*TM research tool against the National

validation. The weight rank also corresponds to the SEQ ID NO: in the accompanying sequence listing thus the transcript with the greatest weight is SEQ ID NO: 1.

In some embodimentexamples, a signature includes at least 2, such as at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 225, at least 250, at least 275, at least 300, at least 325, at least 350, at least 375, at least 400, at least 425, at least 450, at least 475, at least 500, at least 525, at least 550, at least 575, at least 600, at least 634, or even all 636 of the transcripts in Table 6 that carry the greatest weight, defined as the rank of the average weight in the compound decision score function measured under cross-validation, and still have prognostic value. In some

embodimentexamples, a signature includes the top 10 weighted transcripts, the second top 10 weighted transcripts, the third top 10 weighted transcripts, the fourth top 10 weighted transcripts, the fifth top 10 weighted transcripts, the sixth top 10 weighted transcripts, the seventh top 10 weighted transcripts, the eighth top 10 weighted transcripts, the ninth top 10 weighted transcripts, or the tenth top 10 weighted transcripts listed in Table 6. In yet further embodimentexamples, a signature includes the 636, 634, 620, 610, 600, 590, 580, 570, 560, 550, 540, 530, 520, 510, 500, 490, 480, 470, 460, 450, 440, 430, 420, 410, 400, 390, 380, 370, 360, 350, 340, 330, 320, 310, 300, 290, 280, 270, 260, 250, 240, 230, 220, 210, 200, 190, 180, 170, 160, 150, 140, 130, 120, 110, 100, 90, 80, 70, 60, 50, 40, 30, 20, or 10 transcripts having the greatest weight listed in Table 6. In some embodimentexamples, the signature is based on expression levels of from about 200 to about 1000 transcripts, such as from about 400 to about 800 transcripts, such as from about 500 to about 700 transcripts, or in some embodimentexamples, from about 550 to about 650 transcripts, including those from Table 6 (e.g., at least about 50, at least about 100, at least about 200, at least about 300, at least about 400, at least about 500, or at least about 600, or all transcripts from Table 6) as described above.

In one embodimentexample, a specific signature may be used for the methods disclosed herein that includes transcripts for MUM1 and SIGMAR1. In another embodimentexample, a signature may be used for the methods disclosed herein that includes transcripts for MUM1, SIGMAR1, ARSD, SULT1C2 and PPFIBP1. In yet

ARSD, CXCL9, PCLO, SLC2A3, FCGBP, SLC2A14, SLC2A3, BCL9L and antisense sequences of MUC3A, OLFM4 and RNF39. This signature is represented by Table 1 below.

Table 1: 10 candidate core transcripts within the 636 transcript signature

Gene Name	Weight Rank in 636 transcript signature	Δ AUC (Univariate)	Orientation
ARSD	3	-0.0109	Sense
CXCL9	24	-0.0103	Sense
PCLO	272	-0.0095	Sense
SLC2A3	23	-0.0087	Sense
FCGBP	416	-0.0062	Sense
SLC2A14 /// SLC2A3	55	-0.0061	Sense
BCL9L	175	-0.0059	Sense
MUC3A	112	-0.0084	AntiSense
OLFM4	61	-0.0083	AntiSense
RNF39	14	-0.0064	AntiSense

5

In some embodiment examples, a core set of gene transcripts in colon cancer signature is provided that is identified through a separate study to determine the contribution that each of the 636 probesets makes to the performance of the signature.

10 In this embodiment example, ten probesets from the 636 probeset signature were removed and a new signature was created based on 636 probesets, using the training dataset. The new signature was then used to predict the validation dataset (without threshold) and the AUC was measured. The difference in AUC from the 636 probeset signature was recorded. This process was repeated 0.5 million times and the average difference in AUC that occurred for signatures lacking said probeset was recorded.

15 The probesets with the largest negative Δ AUC are recorded in Table 1. In this embodiment example, this set of 10 transcripts represents a candidate core set of genes whose absence from the signature significantly impairs the predictive performance of the signature.

20 Thus in certain embodiment examples, the transcripts representing the genes in Table 1 are included in a colon cancer signature. In Table 1, the Δ AUC represents the drop in validation AUC if this transcript is omitted from the signature. The orientation describes the

orientation of the transcript expressed in colon tissue. Three transcripts in this signature are expressed as antisense transcripts of MUC3A, OLFM4 and RNF39.

In some embedimentexamples, the signature includes a combination of 626-636 transcripts from Table 6, that include ARSD, CXCL9, PCLO, SLC2A3, FCGBP, SLC2A14, SLC2A3, BCL9L, MUC3A, OLFM4 and RNF39. In yet another embedimentexample, the signature includes transcripts 10-636, 10-50, 50-636, 100-636, listed in Table 6 which includes ARSD, CXCL9, PCLO, SLC2A3, FCGBP, SLC2A14, SLC2A3, BCL9L, MUC3A, OLFM4 and RNF39 where the transcript orientation is noted in Table 6.

Notably, 176 transcripts have been identified as being unrepresented by the leading generic array by probe-level analysis (*i.e.* they are “unique” to the *Colorectal Cancer DSA*TM tool described above). This group of 176 transcripts listed in Table 2 are described herein as transcripts that are unique to the colon gene signatures and methods of use herein. Probe-sequence-level homology searches have identified these transcripts as not being contained on the leading generic array (Affymetrix) (*i.e.* they are “unique” to the *Colorectal Cancer DSA*TM research tool described above). A number of these transcripts are antisense transcripts not previously reported to be expressed. These 176 transcripts are presented in Table 2 below, where the weight rank corresponds to the numbers shown in Table 6. Thus the sequence of these unique transcripts can be found in Table 6.

Table 2: Unique transcripts in 636 transcript signature

Weight Rank in 636 Transcript Signature	Gene Symbol	Orientation	Gene Description
424	AC068491.1 (Clone_based _vega_gene)	Sense	non-protein coding RNA 152 (NCRNA00152), transcript variant 2, non-coding RNA [Source:RefSeq DNA;Acc:NR_024205]
214	AC004968.2 (Clone_based _ensembl_gen e)	Sense	Known long non-coding RNA

			Symbol;Acc:20657]
543	TTC39B	AntiSense	tetratricopeptide repeat domain 39B [Source:HGNC Symbol;Acc:23704]
342	U6 (RFAM)	Sense	U6 spliceosomal RNA [Source:RFAM;Acc:RF00026]
288	WSB1	AntiSense	WD repeat and SOCS box- containing 1 [Source:HGNC Symbol;Acc:19221]
523	YLPM1	AntiSense	YLP motif containing 1 [Source:HGNC Symbol;Acc:17798]
386	YPEL5	AntiSense	yippee-like 5 (Drosophila) [Source:HGNC Symbol;Acc:18329]
612	ZFAND3	Sense	zinc finger, AN1-type domain 3 [Source:HGNC Symbol;Acc:18019]
76	ZHX2	AntiSense	zinc fingers and homeoboxes 2 [Source:HGNC Symbol;Acc:18513]
161	ZNF75A	AntiSense	zinc finger protein 75a [Source:HGNC Symbol;Acc:13146]
409	ZXDC	AntiSense	ZXD family zinc finger C [Source:HGNC Symbol;Acc:28160]

In some embedding examples, a signature includes at least 2, such as at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, or even all 176 of the transcripts listed in Table 2, for example those that carry the greatest weight, defined as the rank of the average weight in the compound decision score function measured under cross-validation, and still have prognostic value. In some embedding examples, a signature includes the top 10 weighted transcripts, the second top 10 weighted transcripts, the third top 10 weighted transcripts, the fourth top 10 weighted transcripts, the fifth top 10 weighted transcripts, the sixth top 10 weighted transcripts, the seventh top 10 weighted transcripts, the eighth top 10 weighted transcripts, the ninth top 10 weighted transcripts, or the tenth top 10

weighted transcripts listed in Table 2. In yet further ~~embodiment~~examples, a signature includes the 176, 160, 150, 140, 130, 120, 110, 100, 90, 80, 70, 60, 50, 40, 30, 20, or 10 transcripts having the greatest weight listed in Table 2.

In some ~~embodiment~~examples, the methods described herein include
5 subjecting RNA isolated from a patient to gene expression profiling. Thus, the gene expression profile may be completed for a set of genes that includes at least two of the transcripts listed in Table 6, which in some examples are normalized as described

below. In particular ~~embodiment~~examples of the methods disclosed herein, the expression level of at least 2, such as at least 3, at least 4, at least 5, at least 6, at least
10 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 225, at least 250, at least 275, at least 300, at least 325, at least 350, at least 375,

15 at least 400, at least 425, at least 450, at least 474, at least 500, at least 525, at least 550, at least 575, at least 600, at least 634, or even all 636 of the transcripts in Table 6 or their expression products, and/or complement is determined, for example the transcripts in Table 6 that carry the greatest weight, defined as the rank of the average weight in the compound decision score function measured under cross-validation, and
20 still have prognostic value. In some ~~embodiment~~examples of this method, the expression level of at least at least 2, such as at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least

25 85, at least 90, at least 95, at least 100, at least 125, at least 150, or even all 176 of the transcripts in Table 2 or their expression products, and/or complement is determined, for example those that carry the greatest weight, defined as the rank of the average weight in the compound decision score function measured under cross-validation, and still have prognostic value. In the methods described herein, the combination of
30 transcripts may be referred to as a signature or expression signature.

The relative expression levels of transcripts in a colon tissue are measured to form a gene expression profile. In one ~~embodiment~~example, the gene expression profile of a set of transcripts from a patient tissue sample is summarized in the form of a compound decision score and compared to a control threshold, such a threshold that is

mathematically derived from a training set of patient data. The threshold separates a patient group based on different characteristics such as, but not limited to, good/poor prognosis, responsiveness/non-responsiveness to treatment, cancer detection/diagnosis and cancer classification. The patient training set data is

5 preferably derived from colon tissue samples having been characterized by prognosis, likelihood of recurrence, or long term survival, diagnosis, cancer classification, personalized genomics profile, clinical outcome, treatment response. Expression profiles, and corresponding decision scores from patient samples may be correlated with the characteristics of patient samples in the training set that are on the same side
10 of the mathematically derived decision threshold. In this embodimentexample, the threshold of the linear classifier compound decision score was optimized to maximize the sum of sensitivity and specificity under cross-validation applied within the training dataset. These methods are also useful for determining prognosis of colon cancer and in a particular embodimentexample a patient with stage II colon cancer. In
15 some examples, the disclosed methods are predictive of poor clinical outcome, which can be measured, for example, in terms of shortened survival or increased risk of cancer recurrence, *e.g.* following surgical removal of the cancer, or following surgical removal of the cancer in combination with adjuvant chemotherapy.

Methods are provided for diagnosing colon cancer in a sample obtained from a
20 subject. Such methods include detecting the expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from the subject and comparing the expression level of the at least 2 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of a diagnosis of colon cancer, wherein the expression
25 level, or a decision score derived therefrom, on the same side of the threshold indicates a diagnosis of colon cancer. In some examples, a control threshold is a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a known colon cancer sample (or samples).

Methods are provided for classifying a colon cancer sample. Such methods
30 include detecting the expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject and comparing the expression level of the at least 2 colon cancer-related nucleic acid molecules, or a decision score derived therefrom, to a control threshold indicative of known classification, wherein the expression level, or a decision score

sample (or samples) obtained from a subject (or subjects) having a history of long term survival.

Also provided are methods for predicting of recurrence of colon cancer in a subject, such as subject diagnosed as having stage II colon cancer. These methods include detecting the expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject and comparing the expression level of the at least 2 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of a history of recurrence, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold indicates a recurrence in the subject. In some examples, a control threshold is a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a colon cancer sample (or samples) having a history of recurrence.

Methods are provided for preparing a personalized colon cancer genomics profile for a subject. The methods include detecting an expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject and creating a report summarizing the data obtained by the gene expression analysis.

In particular ~~embodiment~~examples of the methods disclosed herein, the expression levels for at least 2, such as at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 225, at least 250, at least 275, at least 300, at least 325, at least 350, at least 375, at least 400, at least 425, at least 450, at least 474, at least 500, at least 525, at least 550, at least 575, at least 600, at least 634, or even all 636 of the transcripts in Table 6 or their expression products is determined and compared with a control threshold. In other ~~embodiment~~example of these methods the expression levels for MUM1 and SIGMAR1 or their expression products is determined and compared with a control threshold. In another ~~embodiment~~example, the expression levels for MUM1, SIGMAR1, ARSD, SULT1C2 and PPFIBP1 or their expression products is determined and compared with the control threshold. In additional ~~embodiment~~examples, the expression levels for ARSD, CXCL9, PCLO, SLC2A3, FCGBP, SLC2A14, SLC2A3, BCL9L and antisense

sequences of MUC3A, OLFM4 and RNF39 or their expression products is determined and compared with a control threshold. In still other ~~embodimentexamples~~, expression levels for substantially all the transcripts listed in one of Tables 1, 2, and/or 6 are determined in step and compared with a control threshold.

In some ~~embodimentexamples~~ of the disclosed methods, the RNA levels are corrected for (normalize away) both differences in the amount of RNA assayed and variability in the quality of the RNA used. Control transcripts may be included in assays as positive or negative controls and to normalize readings and ensure reliable measurement data, but are preferably omitted for performing the actual prognosis. The exact identity of the former is typically unimportant and a very broad variety of transcripts could be envisaged for all of the purposes disclosed herein. For the normalization controls, a broad variety of transcripts could be envisaged, although they have to fulfill the basic requirements of approximately constant and stable expression between a broad variety of subjects or conditions for the target tissue of interest, in particular between the prognostic groups under consideration. Similarly the RNA degradation controls have to show intensity behavior, suitable for indicating (overly) degraded RNA. This may or may not include RNA controls, which show a stable intensity regardless of the overall RNA degradation of a sample as positive controls. In relation to these controls the intensity pattern for suitable other RNA controls would be analyzed for which an intensity dependency on the RNA degradation stage is observed. This may or may not include specific analyses depending on varying positions of probe sequences with respect to the 3' end of a transcript.

In some ~~embodimentexamples~~ of the disclosed methods, where a microarray is used for quantifying gene expression, one or more of the following controls can be used:

(a) Alignment controls, which are specific transcripts spiked in labeled form, which bind to specific positions on an array and ensure a proper grid alignment in the image processing of a scanned array.

(b) Amplification controls, which are specific unlabeled transcripts, *e.g.* poly-A control transcripts, spiked in before any amplification is performed, so undergoing the same processing as the sample mRNA to ensure an appropriate performance of the cDNA synthesis and subsequent amplification reactions.

(c) Labeling and hybridization controls, which are specific controls spiked in before the labeling and hybridization to the chip for controlling the efficiency of these two steps separately from the prior amplification reaction.

(d) Background controls, which are probe sequences on the microarray for which no corresponding target sequences should be available in the sample. Thus, in principle no specific target binding should occur. These controls are used to establish background or cross-hybridization intensities. They would potentially be characterized by different GC-contents and a suitable spatial distribution over an entire microarray.

(e) Normalization controls, which are probe sequences detecting specifically chosen target sequences from the sample which are used to correct for varying input mRNA amounts, varying yield of amplification reactions and varying overall sensitivity of the measurement device. They are used to correct the measured intensity values and would thus ensure an increased analytical precision of the overall measurement device including the preparatory laboratory steps.

(f) RNA quality and degradation control, which are probe sequences from various positions with respect to the 3' position of their respective genes designed to indicate the RNA quality and detect RNA degradation. Corresponding probes or probe sets from multiple genes might represent differing RNA degradation behavior from different RNA species.

Whereas controls a) – d) can purely be derived based on sequence considerations and should not be naturally present in the tissue and condition of interest, controls e) and f) can be chosen by suitable analyses of prior patient data. This may or may not be the same training data on which the prognostic gene signature has been derived.

It should be understood that the above controls are only provided as example and that other ~~embodiment~~examples of this disclosure could be envisaged (such as qPCR) in which different controls, with similar functionality would be used.

B. Probes, Primers and Arrays

Disclosed are probes and primers specific for the disclosed colon cancer gene signatures. Also disclosed are arrays, which include probes for the disclosed colon cancer signatures. In some ~~embodiment~~examples, a probe specific for the disclosed colon cancer gene signature includes a nucleic acid sequence that specifically hybridizes

one of SEQ ID NOs: 1-636 or the complement thereof. In some

~~embodiment~~examples, a probe set for a disclosed colon cancer signature includes probes that specifically hybridize to at least 2, such as at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 225, at least 250, at least 275, at least 300, at least 325, at least 350, at least 375, at least 400, at least 425, at least 450, at least 474, at least 500, at least 525, at least 550, at least 575, at least 600, at least 634, or even all 636 of the transcripts in Table 6, that carry the greatest weight, defined as the rank of the average weight in the compound decision score function measured under cross-validation, and still have prognostic value, such as a probe that specifically hybridizes to any one of

SEQ ID NOs: 1-636 or the complement thereof. In some ~~embodiment~~examples, a probe set for a disclosed colon cancer signature includes probes that specifically hybridize to the top 10 weighted transcripts, the second top 10 weighted transcripts, the third top 10 weighted transcripts, the fourth top 10 weighted transcripts, the fifth top 10 weighted transcripts, the sixth top 10 weighted transcripts, the seventh top 10 weighted transcripts, the eighth top 10 weighted transcripts, the ninth top 10 weighted transcripts, or the tenth top 10 weighted transcripts listed in Table 6. In yet further

~~embodiment~~examples, a probe set for a disclosed colon cancer signature includes probes that specifically hybridize to 636, 634, 620, 610, 600, 590, 580, 570, 560, 550, 540, 530, 520, 510, 500, 490, 480, 470, 460, 450, 440, 430, 420, 410, 400, 390, 380, 370, 360, 350, 340, 330, 320, 310, 300, 290, 280, 270, 260, 250, 240, 230, 220, 210, 200, 190, 180, 170, 160, 150, 140, 130, 120, 110, 100, 90, 80, 70, 60, 50, 40, 30, 20, or 10 transcripts having the greatest weight listed in Table 6 or the complement

thereof. In some ~~embodiment~~examples, a probe set for a disclosed cancer signature comprises about 200 to about 1000 probes, such as from about 400 to about 800 probes, such as from about 500 to about 700 probes, such as from about 550 to about 650 probes, where the probes detect transcripts from Table 6. The additional probes may be optionally selected from those that detect transcripts that are expressed in colon cancer, or which function as signal controls or expression level controls. Such optional probes can be selected from those included on the *Colorectal Cancer DSA*TM tool.

In some ~~embodiment~~examples, a probe set for a disclosed colon cancer signature includes probes that specifically hybridize to transcripts for MUM1 and SIGMAR1. In other ~~embodiment~~examples, a probe set for a disclosed colon cancer signature includes probes that specifically hybridize to transcripts for MUM1, SIGMAR1, ARSD, SULT1C2 and PPFIBP1. In yet other ~~embodiment~~examples, a probe set for a disclosed colon cancer signature includes probes that specifically hybridize to transcripts for ARSD, CXCL9, PCLO, SLC2A3, FCGBP, SLC2A14, SLC2A3, BCL9L and antisense sequences of MUC3A, OLFM4 and RNF39. A set of probes or primers can be prepared that is substantially representative of the gene expression signature. "Substantially representative of the gene expression signature" refers to probe sets that specifically hybridize to at least 50%, 60%, 70%, 80%, 90%, 95%, 99%, or 100% of the coding or non-coding transcripts in the gene expression signature, for example at least 50%, 60%, 70%, 80%, 90%, 95%, 99%, or 100% of the coding or non-coding transcripts in the gene expression signatures shown in Table 1, 2, or 6 or the complement thereof.

It is advantageous to use probes which bind to the 3' regions of transcripts in the gene expression signature, specifically where the patient tissue to be analyzed for gene expression is RNA extracted from paraffin embedded tissue. Typically each probe will be capable of hybridizing to a complementary sequence in the respective transcript, which occurs within 1kb, or 500bp, or 300bp, or 200bp, or 100bp of the 3' end of the transcript. In the case of mRNA, the "3' end of the transcript" is defined herein as the polyadenylation site, not including the poly(A) tail.

In one ~~embodiment~~example, a pool of probes making up 30% of the total absolute weight of the signature is used. In alternate ~~embodiment~~examples, a pool of probes making up 40%, 60%, 70%, 80%, 90%, 95% or 100% of the total absolute weight of the signature is used in the methods described herein. The basis for inclusion of markers, as well as the clinical significance of mRNA level variations with respect to the reference set, is indicated below. In some ~~embodiment~~examples, the disclosed probes are part of an array, for example the probes are bound to a solid substrate. Exemplary nucleic acid array and methods of making such arrays are discussed in Section D below.

In some ~~embodiment~~examples, a probe specific for the disclosed colon cancer gene signature is part of a nucleic acid array, such as a microarray. In some examples, such arrays include a nucleic acid sequence that specifically hybridizes one of SEQ ID NOs: 1-636 or the complement thereof. In some ~~embodiment~~examples, a nucleic acid array,

such as a microarray, includes probes that specifically hybridize to at least 2, such as at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 225, at least 250, at least 275, at least 300, at least 325, at least 350, at least 375, at least 400, at least 425, at least 450, at least 474, at least 500, at least 525, at least 550, at least 575, at least 600, at least 634, or even all 636 of the transcripts in Table 6. In some embodimentexamples, a nucleic acid array for a disclosed colon cancer signature includes probes that specifically hybridize to the top 10 weighted transcripts, the second top 10 weighted transcripts, the third top 10 weighted transcripts, the fourth top 10 weighted transcripts, the fifth top 10 weighted transcripts, the sixth top 10 weighted transcripts, the seventh top 10 weighted transcripts, the eighth top 10 weighted transcripts, the ninth top 10 weighted transcripts, or the tenth top 10 weighted transcripts listed in Table 6. In yet further embodimentexamples, a nucleic acid array for a disclosed colon cancer signature includes probes that specifically hybridize to 636, 634, 620, 610, 600, 590, 580, 570, 560, 550, 540, 530, 520, 510, 500, 490, 480, 470, 460, 450, 440, 430, 420, 410, 400, 390, 380, 370, 360, 350, 340, 330, 320, 310, 300, 290, 280, 270, 260, 250, 240, 230, 220, 210, 200, 190, 180, 170, 160, 150, 140, 130, 120, 110, 100, 90, 80, 70, 60, 50, 40, 30, 20, or 10 transcripts having the greatest weight listed in Table 6 or the complement thereof. In some embodimentexamples, a nucleic acid array for a disclosed colon cancer signature comprises about 200 to about 1000 probes, such as from about 400 to about 800 probes, such as from about 500 to about 700 probes, such as from about 550 to about 650 probes, where the probes detect transcripts from Table 6. The additional probes may be optionally selected from those that detect transcripts that are expressed in colon cancer, or which function as signal controls or expression level controls. Such optional probes can be selected from those included on the *Colorectal Cancer DSATM* tool. In some embodimentexamples, a nucleic acid array for a disclosed colon cancer signature comprises more than about 1000 probes.

Also disclosed are primer pairs for the amplification of a gene expression signature for colon cancer nucleic acid. In some examples a primer pair includes a forward primer 15 to 40 nucleotides in length comprising a nucleic acid sequence that specifically hybridizes to any one of the nucleic acid sequences set forth as SEQ ID

NOs: 1-636 or its complement and a reverse primer 15 to 40 nucleotides in length comprising a nucleic acid sequence that specifically hybridizes to any one of the nucleic acid sequences set forth as SEQ ID NOs: 1-636 or its complement, wherein the set of primers is capable of directing the amplification of the nucleic acid.

5 Set of primer pairs for the amplification of a gene expression signature for colon cancer nucleic acids are also disclosed. In some embodimentexamples, a primer set for a disclosed colon cancer signature includes primers that specifically hybridize to and are capable of amplifying at least 2, such as at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 225, at least 250, at least 275, at least 300, at least 325, at least 350, at least 375, at least 400, at least 425, at least 450, at least 474, at least 500, at least 525, at least 550, at least 575, at least 600, at least 634, or even all 636 of the transcripts in Table 6 that carry the greatest weight, defined as the rank of the average weight in the compound decision score function measured under cross-validation, and still have prognostic value such as primers that specifically hybridize to and are capable of amplifying any one of SEQ ID NOs: 1-636 or the complement thereof. In some embodimentexamples, a primer set for a disclosed colon cancer signature includes primers that specifically hybridize to and are capable of amplifying the top 10 weighted transcripts, the second top 10 weighted transcripts, the third top 10 weighted transcripts, the fourth top 10 weighted transcripts, the fifth top 10 weighted transcripts, the sixth top 10 weighted transcripts, the seventh top 10 weighted transcripts, the eighth top 10 weighted transcripts, the ninth top 10 weighted transcripts, or the tenth top 10 weighted transcripts listed in Table 6. In yet further embodimentexamples, a primer set for a disclosed colon cancer signature includes primers that specifically hybridize to and are capable of amplifying 636, 634, 620, 610, 600, 590, 580, 570, 560, 550, 540, 530, 520, 510, 500, 490, 480, 470, 460, 450, 440, 430, 420, 410, 400, 390, 380, 370, 360, 350, 340, 330, 320, 310, 300, 290, 280, 270, 260, 250, 240, 230, 220, 210, 200, 190, 180, 170, 160, 150, 140, 130, 120, 110, 100, 90, 80, 70, 60, 50, 40, 30, 20, or 10 transcripts having the greatest weight listed in Table 6 or the complement thereof.

In some embodimentexamples, a primer set for a disclosed colon cancer signature includes primers that specifically hybridize to and are capable of amplifying transcripts for MUM1 and SIGMAR1. In another embodimentexample, a primer set for a disclosed colon cancer signature includes primers that specifically hybridize to and are capable of amplifying transcripts for MUM1, SIGMAR1, ARSD, SULT1C2 and PPFIBP1. In yet another embodimentexample, a probe set for a disclosed colon cancer signature includes probes that specifically hybridize to transcripts for ARSD, CXCL9, PCLO, SLC2A3, FCGBP, SLC2A14, SLC2A3, BCL9L and antisense sequences of MUC3A, OLFM4 and RNF39. A set of probes or primers can be prepared that is substantially representative of the gene expression signature. “Substantially representative of the gene expression signature” refers to probe sets that specifically hybridize to at least 50%, 60%, 70%, 80%, 90%, 95%, 99%, or 100% of the coding or non-coding transcripts in the gene expression signature, for example at least 50%, 60%, 70%, 80%, 90%, 95%, 99%, or 100% of the coding or non-coding transcripts in the gene expression signatures shown in Table 1, 2, or 6 or the complement thereof.

C. Statistical Determination of Colon Cancer Signatures

The disclosed colon cancer signatures can be evaluated by statistical methods. In some embodimentexamples, the gene expression profile of a patient tissue sample is evaluated by a linear classifier. As used herein, a linear classifier refers to a weighted sum of the individual gene intensities into a compound decision score (“decision function”). The decision score is then compared to a pre-defined cut-off threshold, corresponding to a certain set point in terms of sensitivity and specificity, which indicates if a sample, is above the threshold (decision function positive) or below (decision function negative).

Effectively, this means that the data space, *i.e.* the set of all possible combinations of gene expression values, is split into two mutually exclusive halves corresponding to different clinical classifications or predictions, *e.g.* one corresponding to good prognosis and the other to poor prognosis. In the context of the overall signature, relative over-expression of a certain gene can either increase the decision score (positive weight) or reduce it (negative weight) and thus contribute to an overall decision of, for example, either poor or good prognosis.

The interpretation of this quantity, *i.e.* the cut-off threshold for good versus poor prognosis, is derived in the development phase (“training”) from a set of patients

with known outcome. The corresponding weights and the good/poor prognosis cut-off threshold for the decision score are fixed *a priori* from training data by methods known to those of ordinary skill in the art. In a preferred ~~embodiment~~example of the present method, Partial Least Squares Discriminant Analysis (PLS-DA) is used for determining the weights. (Stähle, *J. Chemom.* 1 185-196, 1987; Nguyen and Rocke, *Bioinformatics* 18 39-50, 2002). Other methods for performing the classification, known to those skilled in the art, may also be with the methods described herein when applied to the transcripts of a colon cancer signature.

Different methods can be used to convert quantitative data measured on these genes or their products into a prognosis or other predictive use. These methods include, but not limited to pattern recognition (Duda *et al.* Pattern Classification, 2nd ed., John Wiley, New York 2001), machine learning (Schölkopf *et al.* Learning with Kernels, MIT Press, Cambridge 2002, Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford 1995), statistics (Hastie *et al.* The Elements of Statistical Learning, Springer, New York 2001), bioinformatics (Dudoit *et al.*, *J. Am. Statist. Assoc.* 97:77-87, 2002; Tibshirani *et al.*, *Proc. Natl. Acad. Sci. USA* 99:6567-6572, 2002) or chemometrics (Vandeginste, *et al.*, Handbook of Chemometrics and Qualimetrics, Part B, Elsevier, Amsterdam 1998).

In some ~~embodiment~~examples, in a training step a set of patient samples for both good and poor prognosis cases are measured and the prediction method is optimised using the inherent information from this training data to optimally predict the training set or a future sample set. In this training step the used method is trained or parameterised to predict from a specific intensity pattern to a specific prognostic call. Suitable transformation or pre-processing steps might be performed with the measured data before it is subjected to the prognostic method or algorithm.

In some ~~embodiment~~examples, a weighted sum of the pre-processed intensity values for each transcript is formed and compared with a threshold value optimised on the training set (Duda *et al.* Pattern Classification, 2nd ed., John Wiley, New York 2001). The weights can be derived by a multitude of linear classification methods, including but not limited to Partial Least Squares (PLS, (Nguyen *et al.*, 2002, *Bioinformatics* 18 (2002) 39-50)) or Support Vector Machines (SVM, (Schölkopf *et al.* Learning with Kernels, MIT Press, Cambridge 2002)).

In some ~~embodiment~~examples, the data is transformed non-linearly before applying a weighted sum, for example as described above. This non-linear transformation might

include increasing the dimensionality of the data. The non-linear transformation and weighted summation might also be performed implicitly, *e.g.* through the use of a kernel function. (Schölkopf *et al.* Learning with Kernels, MIT Press, Cambridge 2002).

5 | In some embedimentexamples, a new data sample is compared with two or more class prototypes, being either real measured training samples or artificially created prototypes. This comparison is performed using suitable similarity measures for example but not limited to Euclidean distance (Duda *et al.* Pattern Classification, 2nd ed., John Wiley, New York 2001), correlation coefficient (van't Veer, *et al.*, 10 | *Nature* 415:530, 2002) etc. A new sample is then assigned to the prognostic group with the closest prototype or the highest number of prototypes in the vicinity.

| In some embedimentexamples, decision trees (Hastie *et al.* The Elements of Statistical Learning, Springer, New York 2001) or random forests (Breiman, 2001Random Forests, Machine Learning 45:5) are used to make a prognostic call 15 | from the measured intensity data for the transcript set or their products.

| In some embedimentexamples, neural networks (Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford 1995) are used to make a prognostic call from the measured intensity data for the transcript set or their products.

| In some embedimentexamples, discriminant analysis (Duda *et al.* Pattern 20 | Classification, 2nd ed., John Wiley, New York 2001), comprising but not limited to linear, diagonal linear, quadratic and logistic discriminant analysis, is used to make a prognostic call from the measured intensity data for the transcript set or their products.

| In some embedimentexamples, Prediction Analysis for Microarrays (PAM, 25 | (Tibshirani *et al.*, *Proc. Natl. Acad. Sci. USA* 99:6567-6572, 2002)) is used to make a prognostic call from the measured intensity data for the transcript set or their products.

| In some embedimentexamples, Soft Independent Modelling of Class Analogy (SIMCA, (Wold, 1976, *Pattern Recogn.* 8:127-139)) is used to make a prognosis from 30 | the measured intensity data for the transcript set or their products.

D. Methods for detection of mRNA

Gene expression can be evaluated by detecting mRNA encoding the gene of interest. Thus, the disclosed methods can include evaluating mRNA. RNA can be isolated from a 35 | sample of a tumor (for example, a colon cancer tumor) from a subject,

signature sequencing (MPSS). In one example, RT-PCR can be used to compare mRNA levels in different samples, to characterize patterns of gene expression, to discriminate between closely related mRNAs, and to analyze RNA structure. In specific examples, the disclosed colon cancer signatures are analyzed by nucleic acid microarray techniques, PCR techniques or combinations thereof.

1. Gene Expression Profiling With Microarray Methods

In some ~~embodiment~~examples, the expression profile of colon cancer-associated genes and/or transcripts, such as those shown in Table 6, can be measured in either fresh or paraffin-embedded tumor tissue, using microarray technology. In this method, polynucleotide sequences of interest, such as polynucleotide sequences that specifically hybridize to the nucleic acid sequences shown in Table 6 or a complement thereof, are plated, or arrayed, on a microchip substrate. The arrayed sequences are then hybridized with nucleic acids from cells or tissues of interest.

Just as in RT-PCR methods (see below), the source of mRNA typically is total RNA isolated from human tumors or tumor cell lines, and corresponding normal tissues or cell lines. Thus RNA can be isolated from a variety of primary tumors or tumor cell lines. If the source of mRNA is a primary tumor, mRNA can be extracted, for example, from frozen or archived paraffin-embedded and/or fixed (*e.g.* formalin-fixed) tissue samples, which are routinely prepared and preserved in everyday clinical practice.

In specific ~~embodiment~~examples of the microarray technique, PCR amplified inserts of cDNA clones or oligonucleotides are applied to a substrate in a dense array. Short oligonucleotides may also be synthesized directly on a substrate using, for example, a combination of semiconductor-based photolithography and solid phase chemical synthesis technologies. (Affymetrix, Inc., Santa Clara, CA). In one ~~embodiment~~example, at least 10,000 nucleotide sequences are present on the substrate. The microarrayed transcripts, immobilized on the substrate are suitable for hybridization under stringent conditions. Fluorescently labeled nucleotide probes may be generated through incorporation of fluorescent nucleotides by reverse transcription of RNA extracted from tissues of interest. Labeled probes applied to the array hybridize with specificity to each nucleotide on the array. After washing to remove non-specifically bound probes, the array is scanned by confocal laser microscopy or

method, such as a CCD camera. Quantitation of hybridization of each arrayed element allows for assessment of corresponding transcript abundance.

With dual color fluorescence, separately labeled nucleotide probes generated from two sources may be hybridized pairwise to the array. The miniaturized scale of the hybridization affords a convenient and rapid evaluation of the expression pattern for large numbers of genes. Such methods have been shown to have the sensitivity required to detect rare transcripts, which are expressed at a few copies per cell, and to reproducibly detect at least approximately two-fold differences in the expression levels (Schena *et al.*, *Proc. Natl. Acad. Sci. USA* 93(2):106 149 (1996)). Microarray analysis can also be performed by commercially available equipment, following manufacturer's protocols, such as by using the Affymetrix GeneChip® technology (Affymetrix, Inc., Santa Clara, CA), or Agilent microarray technology (Agilent Technologies, Inc., Santa Clara, CA).

The development of microarray methods for large-scale analysis of gene expression makes it possible to search systematically for molecular markers of cancer classification and outcome prediction in a variety of tumor types, such as colon cancer tumors.

In particular ~~embodiment~~examples provided herein, arrays can be used to evaluate a colon cancer gene expression profile, for example to prognose or diagnose a patient with colon cancer. When describing an array that consists essentially of probes or primers specific for the genes listed in Table 1, Table 2, and/or the transcripts listed in Table 6, such an array includes probes or primers specific for these colon cancer associated genes, and can further include control probes (for example to confirm the incubation conditions are sufficient). Exemplary control probes include GAPDH, β -actin, and 18S RNA.

i. Array substrates

The solid support of the array can be formed from inorganic material (such as glass) or an organic polymer. Suitable materials for the solid support include, but are not limited to: polypropylene, polyethylene, polybutylene, polyisobutylene, polybutadiene, polyisoprene, polyvinylpyrrolidine, polytetrafluoroethylene, polyvinylidene difluoride, polyfluoroethylene-propylene, polyethylenevinyl alcohol, polymethylpentene, polychlorotrifluoroethylene, polysulfones, hydroxylated biaxially oriented polypropylene, aminated biaxially oriented polypropylene, thiolated

gene expression (Yang *et al.*, Genome Res. 11:1888 1898 (2001)); Competitive PCR and MassARRAY (Oeth *et al.*, 2004, SEQUONOME Application Note); and high coverage expression profiling (HiCEP) analysis (Fukumura *et al.*, Nucl. Acids. Res. 31(16) e94 (2003)).

5 The primers used for the amplification are selected so as to amplify a unique segment of the gene of interest (such as the genes listed in Table 1, Table and Table 6. Primers that can be used to these are commercially available or can be designed and synthesized according to well known methods using the sequences of these genes as available for example in GENBANK®.

10 An alternative quantitative nucleic acid amplification procedure is described in U.S. Pat. No. 5,219,727. In this procedure, the amount of a target sequence in a sample is determined by simultaneously amplifying the target sequence and an internal standard nucleic acid segment. The amount of amplified DNA from each segment is determined and compared to a standard curve to determine the amount of
15 the target nucleic acid segment that was present in the sample prior to amplification.

 In some examples, gene expression is identified or confirmed using the microarray technique. Thus, the expression profile can be measured in either fresh or paraffin-embedded tumor tissue, using microarray technology. In this method, colon cancer signature nucleic acid sequences of interest (including cDNAs and
20 oligonucleotides) are plated, or arrayed, on a microchip substrate. The arrayed sequences are then hybridized with isolated nucleic acids (such as cDNA or mRNA) from cells or tissues of interest. Just as in the RT-PCR method, the source of mRNA typically is total RNA isolated from human tumors, and optionally from corresponding noncancerous tissue and normal tissues or cell lines.

25 In a specific ~~embodiment~~example of the microarray technique, PCR amplified inserts of cDNA clones are applied to a substrate in a dense array. In some examples, the array includes probes specific to at least two of the colon cancer signature genes in Tables 1, 2, and 6. The microarrayed nucleic acids are suitable for hybridization under stringent conditions. Fluorescently labeled cDNA probes may be generated through
30 incorporation of fluorescent nucleotides by reverse transcription of RNA extracted from tissues of interest. Labeled cDNA probes applied to the chip hybridize with specificity to each spot of DNA on the array. After stringent washing to remove non-specifically bound probes, the chip is scanned by confocal laser microscopy or by another detection method, such as a CCD camera. Quantitation of hybridization of

Sample cells or tissues are treated to increase their permeability to allow a probe, such as a cancer survival factor-associated gene-specific probe, to enter the cells. The probe is added to the treated cells, allowed to hybridize at pertinent temperature, and excess probe is washed away. A complementary probe is labeled so that the probe's location and quantity in the tissue can be determined, for example, using autoradiography, fluorescence microscopy or immunoassay. The sample may be any sample as herein described, such as a non-tumor sample or a breast or lung tumor sample. Since the sequences of the cancer survival factor-associated genes of interest are known, probes can be designed accordingly such that the probes specifically bind the gene of interest.

In situ PCR is the PCR-based amplification of the target nucleic acid sequences prior to ISH. For detection of RNA, an intracellular reverse transcription step is introduced to generate complementary DNA from RNA templates prior to *in situ* PCR. This enables detection of low copy RNA sequences.

Prior to *in situ* PCR, cells or tissue samples are fixed and permeabilized to preserve morphology and permit access of the PCR reagents to the intracellular sequences to be amplified. PCR amplification of target sequences is next performed either in intact cells held in suspension or directly in cytocentrifuge preparations or tissue sections on glass slides. In the former approach, fixed cells suspended in the PCR reaction mixture are thermally cycled using conventional thermal cyclers. After PCR, the cells are cytocentrifuged onto glass slides with visualization of intracellular PCR products by ISH or immunohistochemistry. *In situ* PCR on glass slides is performed by overlaying the samples with the PCR mixture under a coverslip, which is then sealed to prevent evaporation of the reaction mixture. Thermal cycling is achieved by placing the glass slides either directly on top of the heating block of a conventional or specially designed thermal cycler or by using thermal cycling ovens.

Detection of intracellular PCR products is generally achieved by one of two different techniques, indirect *in situ* PCR by ISH with PCR-product specific probes, or direct *in situ* PCR without ISH through direct detection of labeled nucleotides (such as digoxigenin-11-dUTP, fluorescein-dUTP, ³H-CTP or biotin-16-dUTP), which have been incorporated into the PCR products during thermal cycling.

In some ~~embedment~~ examples of the detection methods, the expression of one or more "housekeeping" genes or "internal controls" can also be evaluated. These terms include any constitutively or globally expressed gene (or protein, as discussed below)

recurrence whereas patients with a predicted good prognosis have a 13% probability of recurrence within 5 years. One of the major advantages of the current approach is that it is based on expression profiling from FFPE tissue which is the preferred method of storage for the majority of available tissue banks. (Abramovitz *Proteome Sci.* 4:5, 2006). RNA extracted from FFPE tissue samples tends to have a shorter median length due to degradation and formalin-induced modification, which makes it difficult for generic arrays to detect. When defining the colon cancer transcriptome, a 3'-based sequencing approach was employed facilitating design of probesets to the 3' extremity of each transcript. This approach ensures much higher detection rate and is thus optimally designed to detect RNA transcripts from both fresh frozen and FFPE tissue samples. The results from the current study showed that the Almac Diagnostics *Colorectal Cancer DSA*TM research tool is capable of producing biologically meaningful and reproducible data from FFPE derived tissue.

Example 2

Prognosis of Cancer

This example describes particular methods that can be used to prognose a subject diagnosed with colon cancer. However, one skilled in the art will appreciate that methods that deviate from these specific methods can also be used to successfully provide the prognosis of a subject with colon cancer.

A tumor sample and adjacent non-tumor sample is obtained from the subject. Approximately 1-100 µg of tissue is obtained for each sample type, for example using a fine needle aspirate. RNA and/or protein is isolated from the tumor and non-tumor tissues using routine methods (for example using a commercial kit).

In one example, the prognosis of a colon cancer tumor is determined by detecting expression levels of 2 or more of the transcript in Tables 1, 2, and/or 6 in a tumor sample obtained from a subject by microarray analysis or real-time quantitative PCR. For example, the disclosed gene signature can be utilized. The relative expression level of in the tumor sample is compared to the control (*e.g.*, RNA isolated from adjacent non-tumor tissue from the subject). In other cases, the control is a reference value, such as the relative amount of such molecules present in non-tumor samples obtained from a group of healthy subjects or cancer subjects.

~~In view of the many possible embodiments to which the principles of the disclosure may be applied, it should be recognized that the illustrated embodiments~~

We claim:

1. A method for diagnosing colon cancer in a sample obtained from a subject, comprising:

5 detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and
comparing the expression level of the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom, to a control threshold indicative of a diagnosis of colon cancer, wherein the expression level, or a decision score derived
10 therefrom, on the same side of the threshold of the known diagnosis control indicates a diagnosis of colon cancer, thereby diagnosing colon cancer in the sample obtained from the subject.

2. The method of claim 1, wherein the control threshold comprises:

15 a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a known colon cancer sample (or samples), wherein the expression level, or a decision score derived therefrom, on the same side of the threshold as a known colon cancer group indicates a diagnosis of colon cancer.

20 3. A method for classifying a colon cancer sample, comprising:

detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and
comparing the expression level of the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of known
25 classification, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold of the known classification control permits classification of the colon cancer sample.

4. The method of claim 3, wherein the control threshold comprises:

30 a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a colon cancer sample (or samples) of known classification, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold as a colon cancer sample (or samples) of known classification permits classification of the colon cancer sample.

5. The method of any one of claims 3 or 4, wherein the colon cancer sample is classified as stage I, stage II, stage III and stage IV.

5 6. The method of any one of claims 3-5, further comprising choosing a treatment plan that will be effective for the classified colon cancer.

7. The method of claim 6, wherein the treatment is surgical resection, chemotherapy, radiation or any combination thereof.

10 8. A method for predicting a response to a treatment for colon cancer, comprising:

detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and

15 comparing the expression level of the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of a known response to treatment, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold of the known response control indicates a similar response to treatment, thereby predicting response to treatment.

20 9. The method of claim 8, wherein the control threshold comprises:
a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a colon cancer sample (or samples) having a known response to treatment, wherein the expression level, or a decision score derived therefrom, on
25 the same side of the threshold as a colon cancer sample (or samples) having a known response to treatment indicates a similar response to treatment, thereby predicting response to treatment.

30 10. The method of any one of claims 8 or 9, wherein the treatment is surgical resection.

11. The method of any one of claims 8-10, wherein the treatment is chemotherapy and/or radiation.

12. A method for predicting long term survival of a subject with colon cancer, comprising:

detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and

5 comparing the expression level of the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of having a history of long term survival, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold of the known long term survival control indicates long term survival of the subject, thereby predicting long term survival of a subject.

10 13. The method of claim 12, wherein the control threshold comprises:

a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a colon cancer sample (or samples) obtained from a subject (or subjects) having a history of long term survival, wherein the expression level, or a
15 decision score derived therefrom, on the same side of the threshold as a colon cancer sample (or samples) obtained from a subject (or subjects) having a history of long term survival indicates long term survival of the subject, thereby predicting long term survival of a subject.

14. The method of claim 13, wherein long term survival comprises at least 5 year
20 survival.

15. A method for predicting of recurrence of colon cancer in a subject, comprising:

25 detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and

comparing the expression level of the the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of a history of recurrence, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold of the known recurrence control indicates a recurrence in the
30 subject.

16. The method of claim 15, wherein the control threshold comprises:

a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a colon cancer sample (or samples) having a history of

recurrence, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold as a colon cancer sample (or samples) of known history of recurrence, indicates a recurrence in the subject.

5 17. A method of preparing a personalized colon cancer genomics profile for a subject, comprising:

detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and

10 creating a report summarizing the data obtained by the gene expression analysis.

18. The method of any one of claims 1-17, wherein the nucleic acids obtained from the subject comprise RNA and/or cDNA transcribed from RNA extracted from a sample of colorectal tissue obtained from the subject.

15 19. The method of claim 18, wherein the sample is a biopsy sample.

20 20. The method of any one of claims 18 or 19, wherein the sample is a fixed and/or paraffin embedded sample.

21. The method of any one of claims 1-20, wherein the expression level is normalized against a control gene or genes.

25 22. The method of any one of claims 1-21, wherein the level of expression is determined with PCR and/or or microarray-based methods.

30 23. The method of any one of claims 1-22, wherein detecting the expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 comprises detecting the expression levels for the transcripts listed in Table 1.

24. The method of any one of claims 1-23, wherein detecting the expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 comprises detecting the expression levels for the transcripts listed in Table 2.

25. A method for preparing a gene expression profile indicative of colon cancer prognosis, comprising:

detecting the expression level of less than 1000 transcripts in a sample comprising
5 RNA isolated from a colon cancer specimen, wherein the first 634 transcripts listed in Table
6 are detected.

26. The method of claim 25, wherein from 400 to 800 transcript expression levels
are detected.

27. The method of claim 25 or 26, wherein from 500 to 700 transcript expression
levels are detected.

28. The method of any one of claim 25-27, wherein the transcripts listed in Table
15 6 comprise the transcripts listed in Table 1.

29. The method of any one of claims 25-28, wherein the colon cancer specimen is
a formalin fixed paraffin-embedded tissue sample.

30. The method of any one of claims 25-29, further comprising, scoring the
20 expression level of the transcripts listed in Table 6, or a decision score derived therefrom,
against corresponding levels or scores for high risk and low risk patient populations.

31. The method of claim 30, further comprising, selecting adjuvant chemotherapy
25 where the patient is determined to be in the high risk group.

32. A method for prognosing colon cancer, comprising:
preparing a gene expression profile for a colon cancer specimen comprising isolated
RNA;

30 and classifying the specimen based the expression levels of the first 634 transcripts
listed in Table 6, or a decision score derived therefrom, in a low risk or high risk group.

33. The method of claim 32, wherein the level of less than 800 transcripts are

detected in the gene expression profile.

34. The method of any one of claims 32-33, wherein the level of less than 700 transcripts are detected in the gene expression profile.

5

35. The method of any one of claims 32-34, wherein the transcripts listed in Table 6 comprise the transcripts listed in Table 1.

36. The method of any one of claims 32-35, wherein the colon cancer specimen is a formalin fixed paraffin-embedded tissue sample.

10

37. The method of any one of claims 32-36, further comprising, selecting adjuvant chemotherapy where the patient is determined to be in the high risk group.

We claim:

1. A method for diagnosing colon cancer in a sample obtained from a subject, comprising:

5 detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and comparing the expression level of the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom, to a control threshold indicative of a diagnosis of colon cancer, wherein the expression level, or a decision score derived
10 therefrom, on the same side of the threshold indicates a diagnosis of colon cancer, thereby diagnosing colon cancer in the sample obtained from the subject.

2. The method of claim 1, wherein the control threshold comprises:
a threshold derived from corresponding transcripts from colon cancer-related nucleic
15 acid molecules listed in Table 6 in a known colon cancer sample (or samples), wherein the expression level, or a decision score derived therefrom, on the same side of the threshold as a known colon cancer group indicates a diagnosis of colon cancer.

3. A method for classifying a colon cancer sample, comprising:
20 detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and comparing the expression level of the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of known classification, wherein the expression level, or a decision score derived therefrom, on the
25 same side of the threshold permits classification of the colon cancer sample.

4. The method of claim 3, wherein the control threshold comprises:
a threshold derived from corresponding transcripts from colon cancer-related nucleic
acid molecules listed in Table 6 in a colon cancer sample (or samples) of known
30 classification, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold as a colon cancer sample (or samples) of known classification permits classification of the colon cancer sample.

5. The method of any one of claims 3 or 4, wherein the colon cancer sample is

classified as stage I, stage II, stage III and stage IV.

6. The method of any one of claims 3-5, further comprising choosing a treatment plan that will be effective for the classified colon cancer.

5

7. The method of claim 6, wherein the treatment is surgical resection, chemotherapy, radiation or any combination thereof.

8. A method for predicting a response to a treatment for colon cancer,
10 comprising:

detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and

15 comparing the expression level of the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of a known response to treatment, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold indicates a similar response to treatment, thereby predicting response to treatment.

9. The method of claim 8, wherein the control threshold comprises:
20 a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a colon cancer sample (or samples) having a known response to treatment, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold as a colon cancer sample (or samples) having a known response to treatment indicates a similar response to treatment, thereby predicting response to
25 treatment.

10. The method of any one of claims 8 or 9, wherein the treatment is surgical resection.

30 11. The method of any one of claims 8-10, wherein the treatment is chemotherapy and/or radiation.

12. A method for predicting long term survival of a subject with colon cancer, comprising:

detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and
comparing the expression level of the the first 634 colon cancer-related nucleic acid molecules, or a decision score derived therefrom to a control threshold indicative of having a
5 history of long term survival, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold indicates long term survival of the subject, thereby predicting long term survival of a subject.

13. The method of claim 12, wherein the control threshold comprises:
10 a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a colon cancer sample (or samples) obtained from a subject (or subjects) having a history of long term survival, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold as a colon cancer sample (or samples) obtained from a subject (or subjects) having a history of long term survival
15 indicates long term survival of the subject, thereby predicting long term survival of a subject.

14. The method of claim 13, wherein long term survival comprises at least 5 year survival.

20 15. A method for predicting of recurrence of colon cancer in a subject, comprising:

detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and
comparing the expression level of the the first 634 colon cancer-related nucleic acid
25 molecules, or a decision score derived therefrom to a control threshold indicative of a history of recurrence, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold indicates a recurrence in the subject.

16. The method of claim 15, wherein the control threshold comprises:
30 a threshold derived from corresponding transcripts from colon cancer-related nucleic acid molecules listed in Table 6 in a colon cancer sample (or samples) having a history of recurrence, wherein the expression level, or a decision score derived therefrom, on the same side of the threshold as a colon cancer sample (or samples) of known history of recurrence, indicates a recurrence in the subject.

17. A method of preparing a personalized colon cancer genomics profile for a subject, comprising:

5 detecting an expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 in a sample comprising nucleic acids obtained from a subject; and
creating a report summarizing the data obtained by the gene expression analysis.

10 18. The method of any one of claims 1-17, wherein the nucleic acids obtained from the subject comprise RNA and/or cDNA transcribed from RNA extracted from a sample of colorectal tissue obtained from the subject.

19. The method of claim 18, wherein the sample is a biopsy sample.

15 20. The method of any one of claims 18 or 19, wherein the sample is a fixed and/or paraffin embedded sample.

20 21. The method of any one of claims 1-20, wherein the expression level is normalized against a control gene or genes.

22. The method of any one of claims 1-21, wherein the level of expression is determined with PCR and/or or microarray-based methods.

25 23. The method of any one of claims 1-22, wherein detecting the expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 comprises detecting the expression levels for the transcripts listed in Table 1.

30 24. The method of any one of claims 1-23, wherein detecting the expression level of the first 634 colon cancer-related nucleic acid molecules listed in Table 6 comprises detecting the expression levels for the transcripts listed in Table 2.

25. A nucleic acid probe for detecting a gene expression signature for colon cancer, comprising or consisting substantially of a nucleic acid molecule between 20 and 40 nucleotides in length, capable of specifically hybridizing to one of the nucleic acid sequences

set forth as SEQ ID NOs: 1-634 or its complement.

26. The nucleic acid probe according to claim 25, wherein the probe is labeled.

5 27. The nucleic acid probe according to claim 26, wherein the probe is radiolabeled, fluorescently-labeled, biotin-labeled, enzymatically-labeled, or chemically-labeled.

28. A set of probes, for detecting a gene expression signature for colon cancer,
10 comprising 2 or more probes, wherein each probe comprises or consists substantially of a nucleic acid molecule between 20 and 40 nucleotides in length capable of specifically hybridizing to one of the nucleic acid sequences set forth as SEQ ID NOs: 1-634 or its complement.

15 29. The set of probes of claim 28, wherein the probes are labeled.

30. The set of probes of claim 29, wherein the probes are radiolabeled, fluorescently-labeled, biotin-labeled, enzymatically-labeled, or chemically-labeled.

20 31. The set of probes of any one of claims 28-29, wherein the set contains at least one probe complementary to each of the transcripts listed in Table 6.

32. The set of probes of claim 31, comprising at least one probe complementary to each transcript in a subset of the transcripts listed in Table 6, wherein the subset comprises at
25 least 1%, 5%, 10%, 25%, 50%, 75% or 95% of the transcripts in Table 6.

33. A device for detecting a gene expression signature for colon cancer, comprising a nucleic acid array comprising the set of probes of any one of claims 28-32.

30 34. A pair of primers for the amplification of a gene expression signature for colon cancer nucleic acid, comprising:

a forward primer 15 to 40 nucleotides in length comprising a nucleic acid sequence that specifically hybridizes to any one of the nucleic acid sequences set forth as SEQ ID NOs: 1-634 or its complement; and

a reverse primer 15 to 40 nucleotides in length comprising a nucleic acid sequence that specifically hybridizes to any one of the nucleic acid sequences set forth as SEQ ID NOs: 1-634 or its complement, wherein the set of primers is capable of directing the amplification of the nucleic acid.

5

35. A set of primer pairs for the amplification of a gene expression signature for colon cancer nucleic acid, comprising at least 2 primer pairs suitable for use in amplification of any one of the nucleic acid sequences set forth as SEQ ID NOs: 1-634.

10

36. The set of primer pairs of claim 35, wherein the set of primer pairs comprises primer pairs suitable for use in amplification of a subset of nucleic acid sequences set forth as SEQ ID NOs: 1-634 wherein the subset comprises at least 1%, 5%, 10%, 25%, 50%, 75% or 95% of the nucleic acid sequences set forth as SEQ ID NOs: 1-636.

15

37. A method for preparing a gene expression profile indicative of colon cancer prognosis, comprising:

detecting the expression level of less than 1000 transcripts in a sample comprising RNA isolated from a colon cancer specimen, wherein the first 634 transcripts listed in Table 6 are detected.

20

38. The method of claim 37, wherein from 400 to 800 transcript expression levels are detected.

25

39. The method of claim 37 or 38, wherein from 500 to 700 transcript expression levels are detected.

40. The method of any one of claim 37-39, wherein the transcripts listed in Table 6 comprise the transcripts listed in Table 1.

30

41. The method of any one of claims 37-40, wherein the colon cancer specimen is a formalin fixed paraffin-embedded tissue sample.

42. The method of any one of claims 37-41, further comprising, scoring the expression level of the transcripts listed in Table 6, or a decision score derived therefrom,

against corresponding levels or scores for high risk and low risk patient populations.

43. The method of claim 42, further comprising, selecting adjuvant chemotherapy where the patient is determined to be in the high risk group.

5

44. A method for prognosing colon cancer, comprising:
preparing a gene expression profile for a colon cancer specimen comprising isolated RNA;

10 and classifying the specimen based the expression levels of the first 634 transcripts listed in Table 6, or a decision score derived therefrom, in a low risk or high risk group.

45. The method of claim 44, wherein the level of less than 800 transcripts are detected in the gene expression profile.

15 46. The method of any one of claims 44-45, wherein the level of less than 700 transcripts are detected in the gene expression profile.

47. The method of any one of claims 44-46, wherein the transcripts listed in Table 6 comprise the transcripts listed in Table 1.

20

48. The method of any one of claims 44-47, wherein the colon cancer specimen is a formalin fixed paraffin-embedded tissue sample.

25 49. The method of any one of claims 44-48, further comprising, selecting adjuvant chemotherapy where the patient is determined to be in the high risk group.