

COLON CANCER GENE EXPRESSION SIGNATURES AND METHODS OF USE

CROSS REFERENCE TO RELATED APPLICATIONS

5 This application claims the benefit of U.S. Provisional Application No. 61/435,922, filed on January 25, 2011, which is incorporated herein by reference in its entirety.

FIELD OF THE INVENTION

10 The present disclosure relates to gene expression profiling in colon tissues, such as colon cancer tissues. In particular, the present disclosure concerns sensitive methods to measure mRNA levels in biopsied colon tumor tissues, including archived paraffin-embedded biopsy material. In addition, the disclosure provides sets of expressed transcripts forming gene expression signatures for the prognosis, diagnosis
15 and treatment of colon cancer.

BACKGROUND OF THE INVENTION

 Approximately 30% of all colon cancer patients are diagnosed with stage II disease. (Jemal *et al.*, *CA Cancer J. Clin.*, 2004). The 5-year survival for patients with
20 stage II colon cancer treated by surgery is approximately 75-80%, demonstrating that the majority of patients are cured by surgery alone. (Benson, *The Oncologist*, 2006; Nauta *et al.*, *Arch. Surg.*, 1989.) Nevertheless, approximately 20-25% of these patients will develop recurrent disease within their lifetime. (Benson, *The Oncologist*, 2006; Gill *et al.*, *J. Clin. Oncol.*, 2004). In theory, these patients should benefit from
25 adjuvant chemotherapy. However, only around 3-4% of patients have an absolute improvement in survival at 5-years with the use of adjuvant chemotherapy in stage II colon cancer. (Benson, *The Oncologist*, 2006; André *et al.*, *Annals of Surgical Oncology* 2006). As a consequence, the American Society of Clinical Oncology guidelines recommend that these patients should not be routinely treated with
30 adjuvant chemotherapy. (Benson *et al.*, *J. Clin. Oncol.*, 2004). Despite this, it is clear that approximately 20% of stage II colon cancer patients, at higher risk of relapse, may be candidates for adjuvant treatment. (Benson, *The Oncologist*, 2006; Nauta *et al.*, *Arch. Surg.*, 1989; Gill *et al.*, *J. Clin. Oncol.*, 2004; André *et al.*, *Annals of Surgical Oncology* 2006.)

In diseases such as colon cancer, the first treatment is often the most important and offers the greatest chance of success, so there exists a need to use the treatment most effective for a patient's particular stage of colon cancer as the first treatment. This has traditionally been impossible because no method was available for predicting which drug treatment would be the most effective for a particular individual's physiology. Many times patients would needlessly undergo toxic drug therapy. For example, in Stage II tumor node metastasis (TNM) colon cancer, there has been no method of determining which patients will respond to adjuvant chemotherapy after surgery. Only one third of the 20% of stage II patients at risk for relapse after surgery derive any benefit from chemotherapy. This means that prescribing adjuvant chemotherapy exposes some patients to treatment that is unnecessary. Alternatively, a decision to withholding adjuvant chemotherapy at this stage will expose some patients to a higher risk of cancer relapse.

Currently, diagnostic tests used in clinical practice are based on a single analyte test, and therefore do not capture the potential value of knowing relationships between dozens of different markers. Moreover, diagnostic tests are frequently not quantitative, relying on immunohistochemistry. This method often yields different results in different laboratories, in part because the reagents are not standardized, and in part because the interpretations can be subjective and may not be easily quantified. RNA-based tests have not often been used because of the problem of RNA degradation over time and the fact that it is difficult to obtain fresh tissue samples from patients for analysis. Fixed paraffin-embedded tissue is more readily available and methods have been established to detect RNA in fixed tissue. However, these methods typically do not allow for the study of large numbers of genes (DNA or RNA) from small amounts of material. Thus, traditionally fixed tissue has been rarely used other than for immunohistochemical detection of proteins.

Recently, several groups have published studies concerning the classification of various cancer types by microarray gene expression analysis (see, e.g. Golub *et al.*, *Science* 286:531-537 (1999); Bhattacharjee *et al.*, *Proc. Natl. Acad. Sci. USA* 98:13790-13795 (2001); Chen-Hsiang *et al.*, *Bioinformatics* 17 (Suppl. 1):S316-S322 (2001); Ramaswamy *et al.*, *Proc. Natl. Acad. Sci. USA* 98:15149-15154 (2001), Salazar *et al.*, *Journal of Clinical Oncology* 29: 17-24 (2010), O'Connell *et al.*, *Journal of Clinical Oncology* 28: 3937-3944 (2010) and Kerr *et al.*, *Journal of Clinical Oncology* 27 (suppl) 15s (2009)). However, these studies mostly focus on

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 embodiments, 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

5 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. The disclosed methods include detecting an expression level of at least 2 colon cancer-

10 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

15 cancer, indicative of a known response to treatment, indicative of having a history of long term survival, indicative of a history of recurrence and the like.

In various embodiments, RNA is isolated from a colon tissue sample, and used for preparing a gene expression profile. In certain embodiments involving prognosis of cancer, the sample is a colorectal tumor specimen, such as a colon cancer sample.

20 In certain embodiments, the gene expression profile 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 embodiments 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

25 other embodiments, the number of transcripts is from about 500 to about 700, or from about 550 to about 650. In various embodiments, 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

30 6), they may be optionally selected from signal or expression level controls, and in some embodiments, are transcripts known to be expressed in colon cancer, such as those determined by *Colorectal Cancer DSA*TM. In certain embodiments, 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 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 embodiments, the gene expression signature is based on the expression level of at
15 least 50 transcripts listed in Table 6, and which may also be listed in Table 1 and/or Table 2. In some 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 Table 6. For example, the
20 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

some examples the probes are part of an array for use in the detection of a colon cancer signature.

5 The foregoing and other features and advantages of the disclosure will become more apparent from the following detailed description of several embodiments, which proceeds with reference to the accompanying figures.

BRIEF DESCRIPTION OF THE DRAWINGS

10 FIG. 1 provides a flow chart showing an exemplary procedure used to derive a colon cancer transcript expression signature.

FIG. 2 provides a flow chart showing an exemplary outline of the stage II colon cancer prognostic signature generation and validation, using the *Colorectal Cancer DSA*[™].

15 FIG. 3A provides a graph of the receiver operating characteristic (ROC) curve of the 636 transcript prognostic signature in the training set.

FIG. 3B provides a Kaplan-Meier plot of recurrence from training data from the candidate model.

FIG. 4A provides a graph of the receiver operating characteristic (ROC) curve of the 636 transcript prognostic signature in the validation set.

20 FIG. 4B provides a Kaplan-Meier plot of recurrence from validation data from the candidate model.

FIG. 5 provides a Kaplan-Meier plot of overall survival from validation data from the candidate model.

FIG. 6 is Table 3 as described below.

25 FIG. 7 is Table 6 as described below.

BRIEF DESCRIPTION OF THE TABLES

30 Table 1 provides a list of 10 candidate transcripts included in a core colon signature. These transcripts have been identified as having the highest impact on the classification of samples into poor and good prognosis groups

Table 2 provides a list 178 unique transcripts included in the colon signature. This table includes the weight rank of the transcript in the 636 transcript signature as well as the orientation of the transcript expressed in colon tissue.

Table 3 provides key patient and tumor characteristics in the study to identify the 636 transcript signature.

Table 4 provides performance metrics for the cross-validated training set and validation set used to identify the transcript signature.

5 Table 5 provides results of the statistical analysis showing Hazards Ratio for patient age, patient gender, pT-stage, tumor grade, tumor location and mucinous/non-mucinous subtype status.

Table 6 provides a list of the transcripts included in the 636-transcript colon signature.

10

SEQUENCE LISTING

The nucleic and amino acid sequences listed in the accompanying sequence listing are shown using standard letter abbreviations for nucleotide bases, as defined in 37 C.F.R. §1.822. Only one strand of each nucleic acid sequence is shown, but the
15 complementary strand is understood as included by any reference to the displayed strand. In the accompanying sequence listing:

SEQ ID NOs: 1-636 are oligonucleotide transcripts from human colon cancer.

The Sequence Listing is submitted as an ASCII text file in the form of the file named ADL-0311_Sequence_Listing.txt, which was created on January 25, 2012, and
20 is 232,154 bytes, which is incorporated by reference herein.

DETAILED DESCRIPTION

I. Summary of Terms

Unless defined otherwise, technical and scientific terms used herein have the
25 same meaning as commonly understood by one of ordinary skill in the art to which this disclosure belongs. Definitions of common terms in molecular biology may be found in Benjamin Lewin, *Genes IX*, published by Jones and Bartlet, 2008 (ISBN 0763752223); Kendrew *et al.* (eds.), *The Encyclopedia of Molecular Biology*, published by Blackwell Science Ltd., 1994 (ISBN 0632021829); Robert A. Meyers
30 (ed.), *Molecular Biology and Biotechnology: a Comprehensive Desk Reference*, published by VCH Publishers, Inc., 1995 (ISBN 9780471185710); Singleton *et al.*, *Dictionary of Microbiology and Molecular Biology* 2nd ed., J. Wiley & Sons (New York, N.Y. 1994), and March, *Advanced Organic Chemistry Reactions, Mechanisms and Structure* 4th ed., John Wiley & Sons (New York, N.Y. 1992).

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
5 terms, will control.

To facilitate review of the various embodiments 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
10 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
15 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
25 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
30 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

longer cDNAs can also be used. Alternatively, the polynucleotides can be oligonucleotides, which are typically about 20 to 80 bases long, although shorter and longer oligonucleotides are also suitable. In one example, the molecule includes oligonucleotides attached to the array via their 5'- or 3'-end.

5 Within an array, each arrayed sample is addressable, in that its location can be reliably and consistently determined within the at least two dimensions of the array. The number of addressable locations on the array can vary, for example from at least four, to at least 9, at least 10, at least 14, at least 15, at least 20, at least 30, at least 50, at least 75, at least 100, at least 150, at least 200, at least 300, at least 500, least 550,
10 at least 600, at least 800, at least 1000, at least 10,000, or more. The feature application location on an array can assume different shapes. For example, the array can be regular (such as arranged in uniform rows and columns) or irregular. Thus, in ordered arrays the location of each sample is assigned to the sample at the time when it is applied to the array, and a key may be provided in order to correlate each location
15 with the appropriate target or feature position. Often, ordered arrays are arranged in a symmetrical grid pattern, but samples could be arranged in other patterns (such as in radially distributed lines, spiral lines, or ordered clusters). Addressable arrays usually are computer readable, in that a computer can be programmed to correlate a particular address on the array with information about the sample at that position (such as
20 hybridization or binding data, including for instance signal intensity). In some examples of computer readable formats, the individual features in the array are arranged regularly, for instance in a Cartesian grid pattern, which can be correlated to address information by a computer.

Binding or stable binding: An association between two substances or
25 molecules, such as the association of a nucleic acid to another nucleic acid (such as the binding of a probe to a transcript shown in Table 6 or its complement), or the association of a protein with another protein or nucleic acid molecule. Binding can be detected by any procedure known to one skilled in the art, for example in the case of a nucleic acid, such as by physical or functional properties of the target:oligonucleotide
30 complex.

Physical methods of detecting the binding of complementary strands of nucleic acid molecules, include but are not limited to, such methods as DNase I or chemical footprinting, gel shift and affinity cleavage assays, Northern blotting, dot blotting and light absorption detection procedures. For example, one method involves

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
5 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
10 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
15 cells and/or tissue samples, such a 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
20 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
25 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
30 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 embodiments herein, the colon cancer is from any stage. In other embodiments, 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 embodiment, a chemotherapeutic agent is an agent of use in treating colon cancer. In one embodiment, 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 bevacuzimab 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 embodiments, 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 embodiments, 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

population value and can be supplied in the format of a graph or table that permits comparison of measured, experimentally determined values.

Detecting expression: Determining of a level expression in either a qualitative or quantitative manner can detect nucleic acid. Exemplary methods include microarray analysis, RT-PCR, and Northern blot. In some examples, detecting expression includes detecting the expression of one or more of the transcripts in Table 6.

Differential expression or altered expression: A difference, such as an increase or decrease, in the conversion of the information encoded in a gene (such as any of the genes from Table 1, 2, and/or nucleic acid transcripts in Table 6) into messenger RNA, the conversion of mRNA to a protein, or both. In some examples, the difference is relative to a control or reference value, such as an amount of expression of a nucleic acid transcript in tissue not affected by a disease, such as colon cancer, from the same subject, or an amount expected in a different subject who does not have colon cancer. The difference can also be in a non-cancerous tissue from a subject (that has the cancer in the same organ) as compared to tissue from a different subject not afflicted with colon cancer. Detecting differential expression can include measuring a change in gene or protein expression, such as a change in expression of one or more of the genes listed in Table 1, 2, and/or the expression one or more transcripts shown in Table 6.

Downregulated or decreased: When used in reference to the expression of a nucleic acid molecule, refers to any process that results in a decrease in production of the nucleic acid. A gene product can be RNA (such as mRNA, rRNA, tRNA, and structural RNA) or protein. Therefore, gene downregulation or deactivation includes processes that decrease transcription of a gene or translation of mRNA.

Gene downregulation includes any detectable decrease in the production of a gene product. In certain examples, production of a gene product decreases by at least 1.2 fold, such as at least 2-fold, at least 3-fold or at least 4-fold, as compared to a control (such an amount of gene expression, such as a normalized gene expression in a normal cell). In several examples, a control is a relative amount of gene expression or protein expression in one or more subjects who do not have colon cancer, such as the relative amount of gene expression or protein expression in “cancer-free” subjects who do not have any known cancer.

Exon: In theory, a segment of an interrupted gene that is represented in the messenger RNA product. In theory the term "intron" refers to any segment of DNA that is transcribed but removed from within the transcript by splicing together the exons on either side of it. Operationally, exon sequences occur in the mRNA
5 sequence of a gene as defined by Ref. Seq ID numbers. Operationally, intron sequences are the intervening sequences within the genomic DNA of a gene, bracketed by exon sequences and having GT and AG splice consensus sequences at their 5' and 3' boundaries.

Expression: The process by which the coded information of a gene is
10 converted into an operational, non-operational, or structural part of a cell, such as the synthesis of nucleic acid or a protein. Gene expression can be influenced by external signals. For instance, exposure of a cell to a hormone may stimulate expression of a hormone-induced gene. Different types of cells can respond differently to an identical signal. Expression of a gene also can be regulated anywhere in the pathway from
15 DNA to RNA to protein. Regulation can include controls on transcription, translation, RNA transport and processing, degradation of intermediary molecules such as mRNA, or through activation, inactivation, compartmentalization or degradation of specific protein molecules after they are produced.

The expression of a nucleic acid molecule can be altered, for example relative
20 to expression in a normal (*e.g.*, non-cancerous) sample. An alteration in gene expression, such as differential expression, includes but is not limited to: (1) overexpression; (2) underexpression; or (3) suppression of expression. Alternations in the expression of a nucleic acid molecule can be associated with, and in fact cause, a change in expression of the corresponding protein. **"Expression" and/or "relative
25 expression"** can be considered the expression value after normalization of a specific transcript with respect to a threshold value, which is defined in the context of the expression of all other transcripts in an expression signature, such as a colon cancer expression signature. The overall expression data for a given sample is normalized using methods known to those skilled in the art in order to correct for differing
30 amounts of starting material, varying efficiencies of the extraction and amplification reactions etc. Using a linear classifier on the normalized data to make a diagnostic or prognostic call (*e.g.* good or poor prognosis) effectively means to split the data space, *i.e.* all possible combinations of expression values for all genes in the signature, into two disjoint halves by means of a separating hyperplane. This split is empirically

derived on a large set of training examples, for example from patients with good and poor prognosis. Without loss of generality, one can assume a certain fixed set of values for all but one genes, which would automatically define a threshold value for this remaining gene where the decision would change from, for example, good to poor prognosis. Expression values above this dynamic threshold would then either indicate good (for a gene with a negative weight) or poor prognosis (for a gene with a positive weight). The precise value of this threshold depends on the actual measured expression profile of all other genes within the signature, but the general indication of certain genes remains fixed, *i.e.* high values or "relative over-expression" always contributes to either a poor prognosis decision (genes with a positive weight) or good prognosis decision (genes with a negative weights). Therefore, in the context of the overall gene expression signature relative expression can indicate if either up- or down-regulation of a certain transcript is indicative of good or poor prognosis.

Gene amplification: A process by which multiple copies of a gene or gene fragment are formed in a particular cell or cell line. The duplicated region (a stretch of amplified DNA) is often referred to as an "amplicon." Usually, the amount of the messenger RNA (mRNA) produced, *i.e.*, the level of gene expression, also increases in the proportion of the number of copies made of the particular gene expressed.

Expression profile (or fingerprint or signature): A pattern of gene expression, which is characteristic of, or correlated with, a specific disease stage or a specific prognostic outcome. The gene expression signature may be represented by a set of informative genes, or transcripts thereof, coding or non-coding or both. The expression levels of the transcripts within the signatures can be evaluated to make a prognostic determination with, but not limited to, the methods provided herein. Gene expression levels may be used to distinguish between two clinical conditions or outcomes such as normal and diseased tissue for diagnosis, or responsiveness compared to non-responsiveness for prognostic methods and recurring compared to non-recurring for predictive methods. Differential or altered gene expression can be detected by changes in the detectable amount of gene expression (such as cDNA or mRNA) or by changes in the detectable amount of proteins expressed by those genes. A distinct or identifiable pattern of gene expression, for instance a pattern of high and low expression of a defined set of genes or gene-indicative nucleic acids such as ESTs; in some examples, as few as one or two genes provides a profile, but more genes can be used in a profile, for example at least 2, at least 3, at least 4, at least 5, at

least 6, at least 7, at least 9, at least 10 or at least 11 and so on. In some embodiments, 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 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 embodiments 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 embodiments, the 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 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 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 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 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

Wash twice: 2x SSC at room temperature (RT) for 15 minutes each

Wash twice: 0.5x SSC at 65°C for 20 minutes each

High Stringency (detects sequences that share at least 80% identity)

Hybridization: 5x-6x SSC at 65°C-70°C for 16-20 hours

5 Wash twice: 2x SSC at RT for 5-20 minutes each

Wash twice: 1x SSC at 55°C-70°C for 30 minutes each

Low Stringency (detects sequences that share at least 60% identity)

Hybridization: 6x SSC at RT to 55°C for 16-20 hours

Wash at least twice: 2x-3x SSC at RT to 55°C for 20-30 minutes each

10 **Isolated:** An “isolated” biological component (such as a nucleic acid molecule, protein, or cell) has been substantially separated or purified away from other biological components in the cell of the organism, or the organism itself, in which the component naturally occurs, such as other chromosomal and extra-chromosomal DNA and RNA, proteins and cells. The term also embraces nucleic acid
15 molecules prepared by recombinant expression in a host cell as well as chemically synthesized nucleic acid molecules. For example, an isolated cell, such as a colon cancer cell, is one that is substantially separated from other types of cells.

Label: An agent capable of detection, for example by ELISA, spectrophotometry, flow cytometry, or microscopy or other visual techniques. For
20 example, a label can be attached to a nucleic acid molecule or protein, thereby permitting detection of the nucleic acid molecule or protein. For example a nucleic acid molecule or an antibody that specifically binds to a target molecule, such as a target nucleic acid molecule. Examples of labels include, but are not limited to, radioactive isotopes, enzyme substrates, co-factors, ligands, chemiluminescent agents,
25 fluorophores, haptens, enzymes, and combinations thereof. Methods for labeling and guidance in the choice of labels appropriate for various purposes are discussed for example in Sambrook *et al.* (Molecular Cloning: A Laboratory Manual, Cold Spring Harbor, New York, 1989) and Ausubel *et al.* (In Current Protocols in Molecular Biology, John Wiley & Sons, New York, 1998).

30 **Long term survival:** Disease-free survival for at least 3 years, more preferably for at least 5 years, even more preferably for at least 8 years following surgery or other treatment (*e.g.*, chemotherapy) for colon cancer.

More aggressive: As used herein, a “more aggressive” form of a colon cancer is a colon cancer with a relatively increased risk of metastasis or recurrence (such as

following surgical removal of the tumor). A “more aggressive” colon cancer can also refer to a colon cancer that confers an increased likelihood of death, or a decrease in the time until death, upon a subject with the colon cancer. A subject having a “more aggressive” form of a colon cancer is considered high risk (poor prognosis).

5 **Nucleic acid molecules representing genes:** Any nucleic acid, for example DNA (intron or exon or both), cDNA, or RNA (such as mRNA), of any length suitable for use as a probe or other indicator molecule, and that is informative about the corresponding gene, such as those listed in Tables 1, or 2, for example the transcripts listed in Table 6.

10 **Oligonucleotide:** A relatively short polynucleotide, including, without limitation, single-stranded deoxyribonucleotides, single- or double-stranded ribonucleotides, RNA:DNA hybrids and double-stranded DNAs. Oligonucleotides, such as single-stranded DNA probe oligonucleotides, are often synthesized by chemical methods, for example using automated oligonucleotide synthesizers that are
15 commercially available. However, oligonucleotides can be made by a variety of other methods, including *in vitro* recombinant DNA-mediated techniques and by expression of DNAs in cells and organisms.

Patient: As used herein, the term “patient” includes human and non-human animals. The preferred patient for treatment is a human. “Patient” and “subject” are
20 used interchangeably herein.

Patient response: can be assessed using any endpoint indicating a benefit to the patient, including, without limitation, (1) inhibition, to some extent, of tumor growth, including slowing down and complete growth arrest; (2) reduction in the number of tumor cells; (3) reduction in tumor size; (4) inhibition (*i.e.*, reduction,
25 slowing down or complete stopping) of tumor cell infiltration into adjacent peripheral organs and/or tissues; (5) inhibition (*i.e.* reduction, slowing down or complete stopping) of metastasis; (6) enhancement of anti-tumor immune response, which may, but does not have to, result in the regression or rejection of the tumor; (7) relief, to some extent, of one or more symptoms associated with the tumor; (8) increase in the
30 length of survival following treatment; and/or (9) decreased mortality at a given point of time following treatment.

Polynucleotide: When used in singular or plural, generally refers to any polyribonucleotide or polydeoxribonucleotide, which may be unmodified RNA or DNA or modified RNA or DNA, or even combinations thereof. Thus, for instance,

polynucleotides as defined herein include, without limitation, single- and double-stranded DNA, DNA including single- and double-stranded regions, single- and double-stranded RNA, and RNA including single- and double-stranded regions, hybrid molecules comprising DNA and RNA that may be single-stranded or, more typically, double-stranded or include single- and double-stranded regions. The term "polynucleotide" also includes DNAs and RNAs that contain one or more modified bases. Thus, DNAs or RNAs with backbones modified for stability or for other reasons are "polynucleotides" as that term is intended herein. Moreover, DNAs or RNAs comprising unusual bases, such as inosine, or modified bases, such as tritiated bases, are included within the term "polynucleotides" as defined herein. In general, the term "polynucleotide" embraces all chemically, enzymatically and/or metabolically modified forms of unmodified polynucleotides, as well as the chemical forms of DNA and RNA characteristic of viruses and cells, including simple and complex cells.

Probes and primers: A probe comprises an isolated nucleic acid capable of hybridizing to a target nucleic acid (such one of the nucleic acid sequences shown in Table 6 or the complement thereof). A detectable label or reporter molecule can be attached to a probe. Typical labels include radioactive isotopes, enzyme substrates, co-factors, ligands, chemiluminescent or fluorescent agents, haptens, and enzymes. Methods for preparing and using nucleic acid probes and primers are described, for example, in Sambrook *et al.* (In *Molecular Cloning: A Laboratory Manual*, CSHL, New York, 1989), Ausubel *et al.* (ed.) (In *Current Protocols in Molecular Biology*, John Wiley & Sons, New York, 1998), and Innis *et al.* (*PCR Protocols, A Guide to Methods and Applications*, Academic Press, Inc., San Diego, CA, 1990). Methods for labeling and guidance in the choice of labels appropriate for various purposes are discussed, for example in Sambrook *et al.* (In *Molecular Cloning: A Laboratory Manual*, CSHL, New York, 1989) and Ausubel *et al.* (In *Current Protocols in Molecular Biology*, John Wiley & Sons, New York, 1998).

Probes are generally at least 12 nucleotides in length, such as at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 45, at least 50, or more contiguous nucleotides complementary to the target nucleic acid molecule, such as a primer of 15-

50 nucleotides, 20-50 nucleotides, or 15-30 nucleotides. In some examples, a probe is even longer, such as a cDNA probe, which can be from about 500 to more than 5000 nucleotides in length.

Primers are short nucleic acid molecules, for instance DNA oligonucleotides 10 nucleotides or more in length, which can be annealed to a complementary target nucleic acid molecule by nucleic acid hybridization to form a hybrid between the primer and the target nucleic acid strand. A primer can be extended along the target nucleic acid molecule by a polymerase enzyme. Therefore, primers can be used to amplify a target nucleic acid molecule (such as a nucleic acid sequence shown in Table 6).

10 The specificity of a primer and/or a probe increases with its length. Thus, for example, a primer that includes 30 consecutive nucleotides will anneal to a target sequence with a higher specificity than a corresponding primer of only 15 nucleotides. Thus, to obtain greater specificity, probes and primers can be selected that include at least 15, 20, 25, 30, 35, 40, 45, 50 or more consecutive nucleotides. In particular
15 examples, a primer is at least 15 nucleotides in length, such as at least 15 contiguous nucleotides complementary to a target nucleic acid molecule. Particular lengths of primers that can be used to practice the methods of the present disclosure include primers having at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least
20 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 45, at least 50, or more contiguous nucleotides complementary to the target nucleic acid molecule to be amplified, such as a primer of 15-50 nucleotides, 20-50 nucleotides, or 15-30 nucleotides. One of most important factors considered in PCR primer design include
25 primer length, melting temperature (T_m), and GC content, specificity, complementary primer sequences, and 3'-end sequence. In general, optimal PCR primers are generally 17-30 bases in length, and contain about 20-80%, such as, for example, about 50-60% G+C bases. T_m 's between 50°C and 80°C, e.g. about 50°C to 70° C are typically preferred.

30 Primer pairs can be used for amplification of a nucleic acid sequence, for example, by PCR, real-time PCR, or other nucleic-acid amplification methods known in the art. An “upstream” or “forward” primer is a primer 5' to a reference point on a nucleic acid sequence. A “downstream” or “reverse” primer is a primer 3' to a reference

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 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

colon cancer tumor), a sample of noncancerous tissue, or a sample of normal tissue (from a subject not afflicted with a known disease or disorder, such as a cancer-free subject).

Sequence identity/similarity: The identity/similarity between two or more nucleic acid sequences, or two or more amino acid sequences, is expressed in terms of the identity or similarity between the sequences. Sequence identity can be measured in terms of percentage identity; the higher the percentage, the more identical the sequences are. Sequence similarity can be measured in terms of percentage similarity (which takes into account conservative amino acid substitutions); the higher the percentage, the more similar the sequences are.

Methods of alignment of sequences for comparison are well known in the art. Various programs and alignment algorithms are described in: Smith & Waterman, *Adv. Appl. Math.* 2:482, 1981; Needleman & Wunsch, *J. Mol. Biol.* 48:443, 1970; Pearson & Lipman, *Proc. Natl. Acad. Sci. USA* 85:2444, 1988; Higgins & Sharp, *Gene*, 73:237-44, 1988; Higgins & Sharp, *CABIOS* 5:151-3, 1989; Corpet *et al.*, *Nuc. Acids Res.* 16:10881-90, 1988; Huang *et al.* *Computer Appls. in the Biosciences* 8, 155-65, 1992; and Pearson *et al.*, *Meth. Mol. Bio.* 24:307-31, 1994. Altschul *et al.*, *J. Mol. Biol.* 215:403-10, 1990, presents a detailed consideration of sequence alignment methods and homology calculations.

The NCBI Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, *J. Mol. Biol.* 215:403-10, 1990) is available from several sources, including the National Center for Biotechnology (NCBI, National Library of Medicine, Building 38A, Room 8N805, Bethesda, MD 20894) and on the Internet, for use in connection with the sequence analysis programs blastp, blastn, blastx, tblastn and tblastx. Additional information can be found at the NCBI web site.

BLASTN is used to compare nucleic acid sequences, while BLASTP is used to compare amino acid sequences. If the two compared sequences share homology, then the designated output file will present those regions of homology as aligned sequences. If the two compared sequences do not share homology, then the designated output file will not present aligned sequences.

Once aligned, the number of matches is determined by counting the number of positions where an identical nucleotide or amino acid residue is presented in both sequences. The percent sequence identity is determined by dividing the number of matches either by the length of the sequence set forth in the identified sequence, or by

an articulated length (such as 100 consecutive nucleotides or amino acid residues from a sequence set forth in an identified sequence), followed by multiplying the resulting value by 100. For example, a nucleic acid sequence that has 1166 matches when aligned with a test sequence having 1554 nucleotides is 75.0 percent identical to the test sequence ($1166 \div 1554 * 100 = 75.0$). The percent sequence identity value is rounded to the nearest tenth. For example, 75.11, 75.12, 75.13, and 75.14 are rounded down to 75.1, while 75.15, 75.16, 75.17, 75.18, and 75.19 are rounded up to 75.2. The length value will always be an integer. In another example, a target sequence containing a 20-nucleotide region that aligns with 20 consecutive nucleotides from an identified sequence as follows contains a region that shares 75 percent sequence identity to that identified sequence (that is, $15 \div 20 * 100 = 75$).

One indication that two nucleic acid molecules are closely related is that the two molecules hybridize to each other under stringent conditions, as described above. Nucleic acid sequences that do not show a high degree of identity may nevertheless encode identical or similar (conserved) amino acid sequences, due to the degeneracy of the genetic code. Changes in a nucleic acid sequence can be made using this degeneracy to produce multiple nucleic acid molecules that all encode substantially the same protein. Such homologous nucleic acid sequences can, for example, possess at least about 60%, 70%, 80%, 90%, 95%, 98%, or 99% sequence identity to a molecule listed in Table 6 determined by this method.

One of skill in the art will appreciate that the particular sequence identity ranges are provided for guidance only; it is possible that strongly significant homologs could be obtained that fall outside the ranges provided.

Splicing or RNA splicing: An RNA processing that removes introns and joins exons to produce mature mRNA with continuous coding sequence that moves into the cytoplasm of a eukaryotic cell.

Transcript or gene product: An RNA molecule that is generated or derived through the process of transcription from its corresponding DNA or a cDNA template. Transcripts include coding and non-coding RNA molecules such as, but not limited to, messenger RNAs (mRNA), alternatively spliced mRNAs, ribosomal RNA (rRNA), transfer RNAs (tRNAs) in addition to a large range of other transcripts, which are not translated into protein such as small nuclear RNAs (snRNAs), antisense molecules such as short interfering RNA (siRNA) and microRNA (miRNA) and other

RNA transcripts of unknown function. In some embodiments, 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
 5 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
 10 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
 15 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
 20 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
 25 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,
 30 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.

Tumor-Node-Metastasis (TNM): The TNM classification of malignant tumors is a cancer staging system for describing the extent of cancer in a patient's body. T describes the size of the primary tumor and whether it has invaded nearby tissue; N describes any lymph nodes that are involved; and M describes metastasis.

- 5 TNM is developed and maintained by the International Union Against Cancer to achieve consensus on one globally recognized standard for classifying the extent of spread of cancer.

Upregulated or activation: When used in reference to the expression of a nucleic acid molecule, refers to any process that results in an increase in production of
10 a gene product. A gene product can be RNA (such as mRNA, rRNA, tRNA, and structural RNA) or protein. Therefore, gene upregulation or activation includes processes that increase transcription of a gene or translation of mRNA, such as an inflammatory gene.

Examples of processes that increase transcription include those that facilitate
15 formation of a transcription initiation complex, those that increase transcription initiation rate, those that increase transcription elongation rate, those that increase processivity of transcription and those that relieve transcriptional repression (for example by blocking the binding of a transcriptional repressor). Gene upregulation can include inhibition of repression as well as stimulation of expression above an
20 existing level. Examples of processes that increase translation include those that increase translational initiation, those that increase translational elongation and those that increase mRNA stability.

Gene upregulation includes any detectable increase in the production of a gene product, such as an inflammatory gene. In certain examples, production of a gene
25 product increases by at least 1.2 fold, such as at least 2-fold, at least 3-fold, at least 4-fold, at least 5-fold, at least 8-fold, at least 10-fold, or at least 15-fold, as compared to a control (such an amount of gene expression and/or normalized gene expression in a normal cell).

Weight: With reference to the gene signatures disclosed herein, refers to the
30 relative importance of an item in a statistical calculation, for example the relative importance of a Transcript in Table 6. The weight of each transcript in a gene expression signature may be determined on a data set of patient samples using analytical methods known in the art. Exemplary procedures are described below.

Suitable methods and materials for the practice or testing of this disclosure are described below. Such methods and materials are illustrative only and are not intended to be limiting. Other methods and materials similar or equivalent to those described herein can be used. For example, conventional methods well known in the art to which this disclosure pertains are described in various general and more specific references, including, for example, Sambrook *et al.*, Molecular Cloning: A Laboratory Manual, 2d ed., Cold Spring Harbor Laboratory Press, 1989; Sambrook *et al.*, Molecular Cloning: A Laboratory Manual, 3d ed., Cold Spring Harbor Press, 2001; Ausubel *et al.*, Current Protocols in Molecular Biology, Greene Publishing Associates, 1992 (and Supplements to 2000); Ausubel *et al.*, Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology, 4th ed., Wiley & Sons, 1999; Harlow and Lane, Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory Press, 1990; Harlow and Lane, Using Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory Press, 1999; Oligonucleotide Synthesis, (M. J. Gait, ed., 1984); Animal Cell Culture, Freshney, ed., 1987; Methods in Enzymology, Academic Press, Inc.); Handbook of Experimental Immunology, 4th ed., D. M. Weir & C. C. Blackwell, eds., Blackwell Science Inc., 1987; Gene Transfer Vectors for Mammalian Cells, J. M. Miller & M. P. Calos, eds., 1987); and PCR: The Polymerase Chain Reaction, Mullis *et al.*, eds., 1994. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

II. Description of Several Embodiments

A. Colon Cancer Expression Signature and Methods of Use

Disclosed herein are expression signatures from colon cancer. The disclosed signatures can be used for applications in prognosis of colon cancer, diagnosis of colon cancer and classifying patient groups. In some embodiments, a sample obtained from a subject, such as a patient, is processed into a set of polynucleotide binding targets that represent transcripts expressed in the tissue sample. The polynucleotide binding targets are probed with complementary polynucleotide probes representing, or corresponding to the signatures described herein in order to obtain information on expression levels of the transcripts. A decision score is optionally calculated that represents the expression levels of the transcripts in the signature. The decision score is then compared to a control, such as a patient population, and genetically similar

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 embodiments, 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 embodiments, 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, 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 embodiments, 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 embodiments, the signatures are developed using a colorectal cancer-focused microarray research tool. In a specific embodiment, 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

Center for Biotechnology Information (NCBI) human Reference Sequence (RefSeq) RNA database (available on the world wide web at ncbi.nlm.nih.gov/RefSeq/) using BLAST analysis, 21,968 (42%) transcripts are present and 26,676 (51%) of transcripts are absent from the human RefSeq database. Furthermore 7% of the content represents expressed antisense transcripts to annotated genes. (Johnston *et al.*, *J. Clin. Oncol.* 24: 3519, 2006; Pruitt *et al.*, *Nucleic Acids Research* 33: D501-D504, 2005). In addition, probe-level analysis of the *Colorectal Cancer DSA*TM compared with leading generic arrays, highlighted that approximately 20,000 (40%) transcripts are not contained on the leading generic microarray platform (Affymetrix) and are unique to the *Colorectal Cancer DSA*TM. Thus, the *Colorectal Cancer DSA*TM research tool includes transcripts that have not been available in hitherto performed gene expression studies.

In some embodiments, the expression of a transcript in a gene expression signature is considered informative if expression levels are increased or decreased between the conditions of interest. Increases or decreases in gene expression can be assessed by methods known to those skilled in the art that include, but are not limited to, using fold changes, t-tests, F-tests, Wilcoxon rank-sum tests, ANOVA (Cui *et al.*, *Genome Biology* 4:210, 2003)) or dedicated methods for detecting differential expression such as Significance Analysis of Microarrays (Tusher *et al.*, *Proc. Natl. Acad. Sci. USA* 98:5116-21, 2001)) or LIMMA (Smyth, *Stat. Appl. Genet. Mol. Biol.*, 3:Art.3, 2004)).

In some embodiments, the transcripts in the signature are used to form a weighted sum of their signals, where individual weights can be positive or negative. The resulting sum ("decisive function") is compared with a pre-determined reference point. The comparison with the reference point may be used to diagnose, or predict a clinical condition or outcome.

One of ordinary skill in the art will appreciate that the transcripts included in the signature provided in Table 1, 2, and/or 6 will carry unequal weights in a signature for diagnosis or prognosis of colon cancer. Therefore, while as few as 1 sequence may be used to diagnose or predict an outcome, the specificity and sensitivity or diagnosis or prediction accuracy may increase using more sequences. Table 6 ranks the transcripts in order of decreasing weight in the signature, defined as the rank of the average weight in the compound decision score function measured under cross-

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

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	DAUC (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 embodiments, 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. In this embodiment, 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. The probesets with the largest negative Δ AUC are recorded in Table 1. In this embodiment, 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. Thus in certain embodiments, the transcripts representing the genes in Table 1 are included in a colon cancer signature. In Table 1, the DAUC represents the drop in validation AUC if this transcript is omitted from the signature. The orientation describes the

20

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 embodiments, 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 embodiment, 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_gene)	Sense	Known long non-coding RNA

50	AC010522.1 (Clone-based (Ensembl) /// ZNF418 /// ZNF814	AntiSense	cDNA FLJ52732, moderately similar to Zinc finger protein 418 [Source:UniProtKB/TrEMBL;Acc:B 4DR41] /// zinc finger protein 418 [Source:HGNC Symbol;Acc:20647] /// zinc finger protein 814 [Source:HGNC Symbol;Acc:33258]
13	AC018359.1 (Clone_based _vega_gene) /// AC123023.1 (Clone_based _vega_gene)	AntiSense	Novel processed transcript /// Putative processed transcript.
559	AC069513.3 (Clone_based _vega_gene)	AntiSense	Novel processed transcript.
242	AC130352.2 (Clone_based _ensembl_gen e)	AntiSense	Novel miRNA.
593	AC138128.1 (Clone_based _ensembl_gen e)	Sense	Novel long non-coding RNA.
488	AC138128.1 (Clone_based _ensembl_gen e)	Sense	Novel long non-coding RNA.
177	ACTN4	AntiSense	actinin, alpha 4 [Source:HGNC Symbol;Acc:166]
427	AL354822.1 (Clone_based _ensembl_gen	Sense	Putative uncharacterized protein ENSP00000383640 [Source:UniProtKB/TrEMBL;Acc:B

	e) /// AC145212.2 (Clone_based _ensembl_gene)		7WNX9] /// Known protein coding.
290	AL604028.2 (Clone_based _ensembl_gene)	Sense	Known protein coding.
498	AMAC1L1	Sense	acyl-malonyl condensing enzyme 1-like 1 [Source:HGNC Symbol;Acc:31043]
93	ANGPTL6	Sense	angiopoietin-like 6 [Source:HGNC Symbol;Acc:23140]
73	APBB2	AntiSense	amyloid beta (A4) precursor protein-binding, family B, member 2 [Source:HGNC Symbol;Acc:582]
250	ARHGAP26	AntiSense	Rho GTPase activating protein 26 [Source:HGNC Symbol;Acc:17073]
81	ARHGEF1	AntiSense	Rho guanine nucleotide exchange factor (GEF) 1 [Source:HGNC Symbol;Acc:681]
326	ARHGEF2/// RP11- 336K24.6 (Clone_based _vega_gene)	Sense	Rho/Rac guanine nucleotide exchange factor (GEF) 2 [Source:HGNC Symbol;Acc:682] /// Known nncoding transcript with no ORF.
391	ASPH	AntiSense	aspartate beta-hydroxylase [Source:HGNC Symbol;Acc:757]
435	ATP2B4	Sense	ATPase, Ca++ transporting, plasma membrane 4 [Source:HGNC Symbol;Acc:817]
108	AXIN2	AntiSense	axin 2 [Source:HGNC

			Symbol;Acc:904]
217	BIRC6	Sense	baculoviral IAP repeat-containing 6 [Source:HGNC Symbol;Acc:13516]
224	BLCAP /// RP11- 425M5.5 (Clone_based _vega_gene)	AntiSense	bladder cancer associated protein [Source:HGNC Symbol;Acc:1055] /// Putative processed transcript.
373	BMPR1A	AntiSense	bone morphogenetic protein receptor, type IA [Source:HGNC Symbol;Acc:1076]
384	BMPR1A	AntiSense	bone morphogenetic protein receptor, type IA [Source:HGNC Symbol;Acc:1076]
552	C2orf89	Sense	UPF0632 protein C2orf89 Precursor [Source:UniProtKB/Swiss- Prot;Acc:Q86V40]
486	C6orf203	Sense	Uncharacterized protein C6orf203 [Source:UniProtKB/Swiss- Prot;Acc:Q9P0P8]
436	C8orf38	Sense	UPF0551 protein C8orf38, mitochondrial Precursor (Putative phytoene synthase) [Source:UniProtKB/Swiss- Prot;Acc:Q330K2]
256	CAMK1D	Sense	calcium/calmodulin-dependent protein kinase ID [Source:HGNC Symbol;Acc:19341]
400	CAPN12	Sense	calpain 12 [Source:HGNC Symbol;Acc:13249]
87	CCND2	AntiSense	cyclin D2 [Source:HGNC Symbol;Acc:1583]
82	CCND2	AntiSense	cyclin D2 [Source:HGNC

			Symbol;Acc:1583]
289	CD200	Sense	CD200 molecule [Source:HGNC Symbol;Acc:7203]
281	CDC42SE2	AntiSense	CDC42 small effector 2 [Source:HGNC Symbol;Acc:18547]
510	CEACAM5	AntiSense	carcinoembryonic antigen-related cell adhesion molecule 5 [Source:HGNC Symbol;Acc:1817]
159	CHD2	AntiSense	chromodomain helicase DNA binding protein 2 [Source:HGNC Symbol;Acc:1917]
309	CHD2	Sense	chromodomain helicase DNA binding protein 2 [Source:HGNC Symbol;Acc:1917]
531	COMMD10	AntiSense	COMM domain containing 10 [Source:HGNC Symbol;Acc:30201]
429	CPEB2	Sense	cytoplasmic polyadenylation element binding protein 2 [Source:HGNC Symbol;Acc:21745]
505	CSNK1A1	Sense	casein kinase 1, alpha 1 [Source:HGNC Symbol;Acc:2451]
466	CTBP2	AntiSense	C-terminal binding protein 2 [Source:HGNC Symbol;Acc:2495]
522	DDX17	AntiSense	DEAD (Asp-Glu-Ala-Asp) box polypeptide 17 [Source:HGNC Symbol;Acc:2740]
63	DEDD	Sense	death effector domain containing [Source:HGNC Symbol;Acc:2755]
407	DHRS11	AntiSense	dehydrogenase/reductase (SDR family) member 11 [Source:HGNC Symbol;Acc:28639]
300	DLG5	Sense	discs, large homolog 5 (Drosophila) [Source:HGNC Symbol;Acc:2904]

344	DSP	AntiSense	desmoplakin [Source:HGNC Symbol;Acc:3052]
378	ECE1	Sense	endothelin converting enzyme 1 [Source:HGNC Symbol;Acc:3146]
151	EEF2K	AntiSense	eukaryotic elongation factor-2 kinase [Source:HGNC Symbol;Acc:24615]
587	EGFR	AntiSense	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian) [Source:HGNC Symbol;Acc:3236]
274	EPHB4	AntiSense	EPH receptor B4 [Source:HGNC Symbol;Acc:3395]
588	FAM190A /// SEPP1 /// UBTD1 /// UTY	AntiSense	family with sequence similarity 190, member A [Source:HGNC Symbol;Acc:29349] /// selenoprotein P, plasma, 1 [Source:HGNC Symbol;Acc:10751] /// ubiquitin domain containing 1 [Source:HGNC Symbol;Acc:25683] /// ubiquitously transcribed tetratricopeptide repe
192	FAM60A	AntiSense	family with sequence similarity 60, member A [Source:HGNC Symbol;Acc:30702]
229	FANCD2	Sense	Fanconi anemia, complementation group D2 [Source:HGNC Symbol;Acc:3585]
625	FAT1	AntiSense	FAT tumor suppressor homolog 1 (Drosophila) [Source:HGNC Symbol;Acc:3595]
12	FNDC3B	Sense	fibronectin type III domain containing 3B [Source:HGNC Symbol;Acc:24670]
446	FNDC3B	Sense	fibronectin type III domain

			containing 3B [Source:HGNC Symbol;Acc:24670]
183	GCC2	Sense	GRIP and coiled-coil domain containing 2 [Source:HGNC Symbol;Acc:23218]
395	GFPT1	AntiSense	glutamine--fructose-6-phosphate transaminase 1 [Source:HGNC Symbol;Acc:4241]
623	GLB1	AntiSense	galactosidase, beta 1 [Source:HGNC Symbol;Acc:4298]
332	GMDS	Sense	GDP-mannose 4,6-dehydratase [Source:HGNC Symbol;Acc:4369]
363	GNL1	Sense	guanine nucleotide binding protein-like 1 [Source:HGNC Symbol;Acc:4413] /// Guanine nucleotide-binding protein-like 1 (GTP-binding protein HSR1) [Source:UniProtKB/Swiss-Prot;Acc:P36915]
512	GPRC5A	AntiSense	G protein-coupled receptor, family C, group 5, member A [Source:HGNC Symbol;Acc:9836]
206	GPT2	AntiSense	glutamic pyruvate transaminase (alanine aminotransferase) 2 [Source:HGNC Symbol;Acc:18062]
341	GRB7	Sense	growth factor receptor-bound protein 7 [Source:HGNC Symbol;Acc:4567]
62	GRHL2	Sense	grainyhead-like 2 (Drosophila) [Source:HGNC Symbol;Acc:2799]
49	GSTO2	Sense	glutathione S-transferase omega 2 [Source:HGNC Symbol;Acc:23064]
56	GSTO2	Sense	glutathione S-transferase omega 2 [Source:HGNC Symbol;Acc:23064]

533	HELZ	Sense	helicase with zinc finger [Source:HGNC Symbol;Acc:16878]
412	HNRNPL	AntiSense	heterogeneous nuclear ribonucleoprotein L [Source:HGNC Symbol;Acc:5045]
198	HSPD1	Sense	heat shock 60kDa protein 1 (chaperonin) [Source:HGNC Symbol;Acc:5261]
114	IGLL5	Sense	immunoglobulin lambda-like polypeptide 5 [Source:HGNC Symbol;Acc:38476]
495	IL32	AntiSense	interleukin 32 [Source:HGNC Symbol;Acc:16830]
394	INPP4B	Sense	inositol polyphosphate-4- phosphatase, type II, 105kDa [Source:HGNC Symbol;Acc:6075]
165	ITGA6	AntiSense	integrin, alpha 6 [Source:HGNC Symbol;Acc:6142]
166	ITGA6	AntiSense	integrin, alpha 6 [Source:HGNC Symbol;Acc:6142]
287	KANK1	AntiSense	KN motif and ankyrin repeat domains 1 [Source:HGNC Symbol;Acc:19309]
226	KANK1	AntiSense	KN motif and ankyrin repeat domains 1 [Source:HGNC Symbol;Acc:19309]
179	KCNK1	AntiSense	potassium channel, subfamily K, member 1 [Source:HGNC Symbol;Acc:6272]
513	KIAA0319L	Sense	KIAA0319-like [Source:HGNC Symbol;Acc:30071]
126	KIF24	Sense	kinesin family member 24 [Source:HGNC Symbol;Acc:19916]

278	KLRAQ1	AntiSense	KLRAQ motif containing 1 [Source:HGNC Symbol;Acc:30595]
237	LRRC37B	AntiSense	leucine rich repeat containing 37B [Source:HGNC Symbol;Acc:29070]
519	MACC1	Sense	metastasis associated in colon cancer 1 [Source:HGNC Symbol;Acc:30215]
301	MACF1	AntiSense	microtubule-actin crosslinking factor 1 [Source:HGNC Symbol;Acc:13664]
238	MAVS	AntiSense	mitochondrial antiviral signaling protein [Source:HGNC Symbol;Acc:29233]
336	MEF2A	Sense	myocyte enhancer factor 2A [Source:HGNC Symbol;Acc:6993]
567	MIR612	AntiSense	hsa-mir-612 [Source:miRBase;Acc:MI0003625]
431	MMP1	AntiSense	matrix metalloproteinase 1 (interstitial collagenase) [Source:HGNC Symbol;Acc:7155]
397	MMP25	Sense	matrix metalloproteinase 25 [Source:HGNC Symbol;Acc:14246]
72	MORC3	Sense	MORC family CW-type zinc finger 3 [Source:HGNC Symbol;Acc:23572]
506	MORC3	Sense	MORC family CW-type zinc finger 3 [Source:HGNC Symbol;Acc:23572]
634	MUC2	AntiSense	mucin 2, oligomeric mucus/gel- forming [Source:HGNC Symbol;Acc:7512]
110	MUC6	Sense	mucin 6, oligomeric mucus/gel- forming [Source:HGNC Symbol;Acc:7517]
133	MUC6	Sense	mucin 6, oligomeric mucus/gel-

			forming [Source:HGNC Symbol;Acc:7517]
1	MUM1	Sense	melanoma associated antigen (mutated) 1 [Source:HGNC Symbol;Acc:29641]
385	MYO10	Sense	myosin X [Source:HGNC Symbol;Acc:7593]
599	MYO1E	AntiSense	myosin IE [Source:HGNC Symbol;Acc:7599]
448	N/A	No Transcript match	N/A
57	N/A	No Transcript match	N/A
628	N/A	No Transcript match	N/A
370	N/A	Sense	non-protein coding RNA 152 (NCRNA00152), transcript variant 2, non-coding RNA [Source:RefSeq DNA;Acc:NR_024205]
6	N/A	No Transcript match	N/A
66	N/A	No Genome match	N/A
610	N/A	No Transcript match	N/A
308	N/A	No Transcript match	N/A
439	N/A	No Genome match	N/A
131	N/A	No Genome match	N/A
359	NAA50	AntiSense	N(alpha)-acetyltransferase 50, NatE catalytic subunit [Source:HGNC

			Symbol;Acc:29533]
333	NAA50	AntiSense	N(alpha)-acetyltransferase 50, NatE catalytic subunit [Source:HGNC Symbol;Acc:29533]
356	NBN	AntiSense	nibrin [Source:HGNC Symbol;Acc:7652]
137	NCAPD2	Sense	non-SMC condensin I complex, subunit D2 [Source:HGNC Symbol;Acc:24305]
348	NCRNA00188	AntiSense	small nucleolar RNA, C/D box 65 [Source:HGNC Symbol;Acc:32726]
297	NCRNA00262	No Transcript match	non-protein coding RNA 262[Source:HGNC Symbol;Acc:26785]
606	NDUFA13 /// YJEFN3	AntiSense	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 13 [Source:HGNC Symbol;Acc:17194] /// YjeF N-terminal domain containing 3 [Source:HGNC Symbol;Acc:24785]
502	NR6A1	AntiSense	nuclear receptor subfamily 6, group A, member 1 [Source:HGNC Symbol;Acc:7985]
420	NR6A1	AntiSense	nuclear receptor subfamily 6, group A, member 1 [Source:HGNC Symbol;Acc:7985]
61	OLFM4	AntiSense	olfactomedin 4 [Source:HGNC Symbol;Acc:17190]
129	PARP14	AntiSense	poly (ADP-ribose) polymerase family, member 14 [Source:HGNC Symbol;Acc:29232]
515	PML	AntiSense	promyelocytic leukemia [Source:HGNC Symbol;Acc:9113]

323	POSTN	AntiSense	periostin, osteoblast specific factor [Source:HGNC Symbol;Acc:16953]
554	PPDPF	Sense	pancreatic progenitor cell differentiation and proliferation factor homolog (zebrafish) [Source:HGNC Symbol;Acc:16142]
381	PPFIBP1	Sense	PTPRF interacting protein, binding protein 1 (liprin beta 1) [Source:HGNC Symbol;Acc:9249]
504	PPP3CA	Sense	protein phosphatase 3, catalytic subunit, alpha isozyme [Source:HGNC Symbol;Acc:9314]
382	PRKDC	AntiSense	protein kinase, DNA-activated, catalytic polypeptide [Source:HGNC Symbol;Acc:9413]
450	PRPF40A	AntiSense	PRP40 pre-mRNA processing factor 40 homolog A (S. cerevisiae) [Source:HGNC Symbol;Acc:16463]
525	PTK2	AntiSense	PTK2 protein tyrosine kinase 2 [Source:HGNC Symbol;Acc:9611]
215	PTP4A1	AntiSense	protein tyrosine phosphatase type IVA, member 1 [Source:HGNC Symbol;Acc:9634]
298	RABGAP1	AntiSense	RAB GTPase activating protein 1 [Source:HGNC Symbol;Acc:17155]
194	RBM47	Sense	RNA binding motif protein 47 [Source:HGNC Symbol;Acc:30358]
461	RERE	AntiSense	arginine-glutamic acid dipeptide (RE) repeats [Source:HGNC Symbol;Acc:9965]
355	RHBDD1	Sense	rhomboid domain containing 1 [Source:HGNC Symbol;Acc:23081]
454	RNF145	AntiSense	ring finger protein 145

			[Source:HGNC Symbol;Acc:20853]
171	RNF43	Sense	ring finger protein 43 [Source:HGNC Symbol;Acc:18505]
496	RP11-357H14.7 (Clone_based_vega_gene)	Sense	Novel processed transcript.
573	RP11-460N11.2 (Clone_based_vega_gene)	AntiSense	Known pseudogene.
172	RP11-460N11.2 (Clone_based_vega_gene)	AntiSense	Known pseudogene.
155	RP11-460N11.2 (Clone_based_vega_gene)	AntiSense	Known pseudogene.
247	RP11-706O15.1 (Clone_based_vega_gene)	AntiSense	HCG1981372, isoform CRA_cNovel protein ; [Source:UniProtKB/TrEMBL;Acc:B1B108]
251	RP11-761E20.1 (Clone_based_vega_gene)	Sense	Novel processed transcript.
307	RP11-86H7.1 (Clone_based_vega_gene)	Sense	Novel processed transcript.
575	RP4-717I23.3 (Clone_based_vega_gene)	AntiSense	Novel processed transcript.

209	RUNX1	AntiSense	runt-related transcription factor 1 [Source:HGNC Symbol;Acc:10471]
95	SAMD4B	AntiSense	sterile alpha motif domain containing 4B [Source:HGNC Symbol;Acc:25492]
17	SATB2	AntiSense	SATB homeobox 2 [Source:HGNC Symbol;Acc:21637]
264	SH3D19	AntiSense	SH3 domain containing 19 [Source:HGNC Symbol;Acc:30418]
235	SH3GLB1	AntiSense	SH3-domain GRB2-like endophilin B1 [Source:HGNC Symbol;Acc:10833]
388	SIPA1L3	Sense	signal-induced proliferation- associated 1 like 3 [Source:HGNC Symbol;Acc:23801]
157	SLC6A6	Sense	solute carrier family 6 (neurotransmitter transporter, taurine), member 6 [Source:HGNC Symbol;Acc:11052]
259	SLC6A6	Sense	solute carrier family 6 (neurotransmitter transporter, taurine), member 6 [Source:HGNC Symbol;Acc:11052]
462	SMURF2	Sense	SMAD specific E3 ubiquitin protein ligase 2 [Source:HGNC Symbol;Acc:16809]
377	SND1	Sense	staphylococcal nuclease and tudor domain containing 1 [Source:HGNC Symbol;Acc:30646]
335	SNTB2	AntiSense	syntrophin, beta 2 (dystrophin- associated protein A1, 59kDa, basic component 2) [Source:HGNC Symbol;Acc:11169]

329	SOD2	AntiSense	superoxide dismutase 2, mitochondrial [Source:HGNC Symbol;Acc:11180]
263	SP100	Sense	SP100 nuclear antigen [Source:HGNC Symbol;Acc:11206]
243	SPDYE2	AntiSense	speedy homolog E2 (Xenopus laevis) [Source:HGNC Symbol;Acc:33841]
594	SPDYE2	AntiSense	speedy homolog E2 (Xenopus laevis) [Source:HGNC Symbol;Acc:33841]
636	SRSF1	AntiSense	serine/arginine-rich splicing factor 1 [Source: HGNC Symbol; Acc:10780]
561	SSFA2	Sense	sperm specific antigen 2 [Source:HGNC Symbol;Acc:11319]
369	TBL1XR1	AntiSense	transducin (beta)-like 1 X-linked receptor 1 [Source:HGNC Symbol;Acc:29529]
605	TEX10	AntiSense	testis expressed 10 [Source:HGNC Symbol;Acc:25988]
453	TFAM	AntiSense	transcription factor A, mitochondrial [Source:HGNC Symbol;Acc:11741]
629	TLCD2	Antisense	TLC domain containing 2 [Source: HGNC Symbol; Acc:33522]
470	TLCD2	Antisense	TLC domain containing 2 [Source: HGNC Symbol; Acc:33522]
35	TMEM87A	Sense	transmembrane protein 87A [Source:HGNC Symbol;Acc:24522]
624	TMPRSS4	AntiSense	transmembrane protease, serine 4 [Source:HGNC Symbol;Acc:11878]
102	TRIM5	Sense	tripartite motif-containing 5 [Source:HGNC Symbol;Acc:16276]
44	TRPS1	AntiSense	trichorhinophalangeal syndrome I [Source:HGNC Symbol;Acc:12340]
221	TSPAN1	Sense	tetraspanin 1 [Source:HGNC

			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 embodiments, 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 embodiments, 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 embodiments, 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 embodiments, the methods described herein include 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 embodiments 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 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, 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 still have prognostic value. In some embodiments 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 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 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, 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 embodiment, 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 embodiment a patient with stage II colon cancer. In some examples, the
15 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

derived therefrom, on the same side of the threshold permits classification of the colon cancer sample. 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) of known classification. In
5 some examples, the colon cancer sample is classified as stage I, stage II, stage III and stage IV. In some examples the method further includes choosing a treatment plan that will be effective for the classified colon cancer, for example surgical resection, chemotherapy, radiation or any combination thereof.

Methods are provided for predicting a response to a treatment for colon
10 cancer, such as a subject with stage II colon cancer. 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 a subject and comparing the expression level of the at least 2 colon cancer-related nucleic acid molecules, or a
15 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. 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
20 known response to treatment. In some embodiments, the method is a method of predicting response from surgical resection, chemotherapy, radiation or any combination thereof.

Methods are provided for predicting long term survival of a subject with colon cancer, such as a subject diagnosed with stage II colon cancer. These methods include
25 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 having a history of long term survival, wherein the expression level, or a decision
30 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. In some examples, the 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) 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 embodiments 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 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, the expression levels for MUM1, SIGMAR1, ARSD, SULT1C2 and PPFIBP1 or their expression products is determined and compared with the control threshold. In additional embodiments, 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 embodiments, 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.

5 In some embodiments 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.

10 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
15 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
20 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 embodiments of the disclosed methods, where a microarray is used
25 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.*
30 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 embodiments 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 embodiments, 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 embodiments, 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 embodiments, 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 embodiments, 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 embodiments, 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 DSATM* tool.

In some embodiments, a probe set for a disclosed colon cancer signature includes probes that specifically hybridize to transcripts for MUM1 and SIGMAR1. In other embodiments, 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 embodiments, 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, a pool of probes making up 30% of the total absolute weight of the signature is used. In alternate embodiments, 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 embodiments, 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 embodiments, 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 embodiments, 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 embodiments, 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 embodiments, 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 embodiments, 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 embodiments, 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 embodiments, 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
- 20 embodiments, 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
- 25 embodiments, 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.
- 30

In some embodiments, 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 embodiment, 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 embodiment, 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 embodiments, 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 of the present method, Partial Least Squares Discriminant Analysis (PLS-DA) is used for
5 determining the weights. (Ståhle, *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
10 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
15 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 embodiments, in a training step a set of patient samples for both good
20 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
25 before it is subjected to the prognostic method or algorithm.

In some embodiments, 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
30 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 embodiments, 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 embodiments, 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.*, *Nature* 10 415:530, 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 embodiments, decision trees (Hastie *et al.* The Elements of Statistical Learning, Springer, New York 2001) or random forests (Breiman, 2001 Random Forests, Machine Learning 45:5) are used to make a prognostic call from the 15 measured intensity data for the transcript set or their products.

 In some embodiments, 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 embodiments, 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 embodiments, Prediction Analysis for Microarrays (PAM, (Tibshirani 25 *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 embodiments, Soft Independent Modelling of Class Analogy (SIMCA, (Wold, 1976, *Pattern Recogn.* 8:127-139)) is used to make a prognosis from the measured intensity data for the transcript set or their products.

30

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 sample of a tumor (for example, a colon cancer tumor) from a subject,

a sample of adjacent non-tumor tissue from the subject, a sample of tumor-free tissue from a normal (healthy) subject, or combinations thereof, using methods well known to one of ordinary skill in the art, including commercially available kits.

General methods for mRNA extraction are well known in the art and are disclosed in standard textbooks of molecular biology, including Ausubel *et al.*, Current Protocols of Molecular Biology, John Wiley and Sons (1997). Methods for RNA extraction from paraffin embedded tissues are disclosed, for example, in Rupp and Locker, *Biotechniques* 6:56-60, 1988, and De Andres *et al.*, *Biotechniques* 18:42-44, 1995. In one example, RNA isolation can be performed using purification kit, buffer set and protease from commercial manufacturers, such as QIAGEN® (Valencia, CA), according to the manufacturer's instructions. For example, total RNA from cells in culture (such as those obtained from a subject) can be isolated using QIAGEN® RNeasy® mini-columns. Other commercially available RNA isolation kits include MASTERPURE® Complete DNA and RNA Purification Kit (EPICENTRE® Madison, Wis.), and Paraffin Block RNA Isolation Kit (Ambion, Inc.). Total RNA from tissue samples can be isolated using RNA Stat-60 (Tel-Test). RNA prepared from tumor or other biological sample can be isolated, for example, by cesium chloride density gradient centrifugation.

The present signatures and methods described herein accommodate the use of archived paraffin-embedded biopsy material for assay of all markers in the set, and therefore are compatible with the most widely available type of biopsy material. The expression level of transcripts in a colon tissue sample may be determined using RNA obtained from a formalin-fixed, paraffin-embedded tissue sample, fresh frozen tissue or fresh tissue that has been stored in solutions such as RNAlater®. The isolation of RNA can, for example, be carried out following any of the procedures described above or throughout the application, or by any other method known in the art. While all techniques of gene expression profiling, as well as proteomics techniques, are suitable for use in performing the methods described herein, the gene expression levels are often determined by DNA microarray technology.

If the source of the tissue is a formalin-fixed, paraffin embedded tissue sample, the RNA may be fragmented, resulting in loss of information. The signatures provided herein are derived from pools of transcripts sequenced from their 3' end thereby providing an accurate representation of the transcriptome of the tissue. Thus

the signatures provided herein are useful for both fresh frozen and fixed paraffin-embedded tissues.

In some embodiments, RNA samples used in the methods described herein may be prepared from a fixed, wax-embedded colon tissue specimen, by using one or more of the following steps, such as all of the following steps:

- (a) deparaffinizing using conventional methods and with multiple wash steps in organic solvent;
- (b) air drying and treating with protease to break inter- and intracellular bonds, resulting the release of RNA from the tissue;
- (c) removing contaminating genomic DNA;
- (d) washing in organic solvent; and eluting in a suitable RNase-free elution buffer.

The RNA-extraction methods may also include incubation of the tissue in a highly denaturing lysis buffer, which has the additional function of reversing much of the formalin crosslinking that occurs in tissues preserved this way to improve RNA yield and quality for performance in downstream assays.

Following RNA recovery, the RNA may optionally be further purified resulting in RNA that is substantially free from contaminating DNA or proteins. Further RNA purification may be accomplished by any of the aforementioned techniques for RNA recovery or with the use of commercially available RNA cleanup kits, such as RNeasy® MinElute® Cleanup Kit (QIAGEN®). The tissue specimen may, for example, be obtained from a tumor, and the RNA may be obtained from a microdissected portion of the tissue specimen enriched for tumor cells.

Methods of gene expression profiling include methods based on hybridization analysis of polynucleotides and methods based on sequencing of polynucleotides. In some examples, mRNA expression in a sample is quantified using Northern blotting or *in situ* hybridization (Parker & Barnes, *Methods in Molecular Biology* 106:247-283, 1999); RNase protection assays (Hod, *Biotechniques* 13:852-4, 1992); and PCR-based methods, such as reverse transcription polymerase chain reaction (RT-PCR) (Weis *et al.*, *Trends in Genetics* 8:263-4, 1992). Alternatively, antibodies can be employed that can recognize specific duplexes, including DNA duplexes, RNA duplexes, and DNA-RNA hybrid duplexes or DNA-protein duplexes. Representative methods for sequencing-based gene expression analysis include Serial Analysis of Gene Expression (SAGE), and gene expression analysis by massively parallel

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
5 microarray techniques, PCR techniques or combinations thereof.

1. Gene Expression Profiling With Microarray Methods

In some embodiments, 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
10 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.

15 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-
20 fixed) tissue samples, which are routinely prepared and preserved in everyday clinical practice.

In specific embodiments 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
25 combination of semiconductor-based photolithography and solid phase chemical synthesis technologies. (Affymetrix, Inc., Santa Clara, CA). In one embodiment, 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
30 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 by another detection

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 embodiments 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

biaxially oriented polypropylene, ethyleneacrylic acid, thylene methacrylic acid, and blends of copolymers thereof (see U.S. Patent No. 5,985,567).

In general, suitable characteristics of the material that can be used to form the solid support surface include: being amenable to surface activation such that upon activation, the surface of the support is capable of covalently attaching a biomolecule such as an oligonucleotide thereto; amenability to “*in situ*” synthesis of biomolecules; being chemically inert such that at the areas on the support not occupied by the oligonucleotides are not amenable to non-specific binding, or when non-specific binding occurs, such materials can be readily removed from the surface without removing the oligonucleotides.

In another example, a surface activated organic polymer is used as the solid support surface. One example of a surface activated organic polymer is a polypropylene material aminated via radio frequency plasma discharge. Other reactive groups can also be used, such as carboxylated, hydroxylated, thiolated, or active ester groups.

ii. Array formats

A wide variety of array formats can be employed in accordance with the present disclosure. One example includes a linear array of oligonucleotide bands, generally referred to in the art as a dipstick. Another suitable format includes a two-dimensional pattern of discrete cells (such as 4096 squares in a 64 by 64 array). As is appreciated by those skilled in the art, other array formats including, but not limited to slot (rectangular) and circular arrays are equally suitable for use (see U.S. Patent No. 5,981,185). In some examples, the array is a multi-well plate. In one example, the array is formed on a polymer medium, which is a thread, membrane or film. An example of an organic polymer medium is a polypropylene sheet having a thickness on the order of about 1 mil (0.001 inch) to about 20 mil, although the thickness of the film is not critical and can be varied over a fairly broad range. The array can include biaxially oriented polypropylene (BOPP) films, which in addition to their durability, exhibit low background fluorescence.

The array formats of the present disclosure can be included in a variety of different types of formats. A “format” includes any format to which the solid support can be affixed, such as microtiter plates (*e.g.*, multi-well plates), test tubes, inorganic sheets, dipsticks, and the like. For example, when the solid support is a polypropylene

thread, one or more polypropylene threads can be affixed to a plastic dipstick-type device; polypropylene membranes can be affixed to glass slides. The particular format is, in and of itself, unimportant. All that is necessary is that the solid support can be affixed thereto without affecting the functional behavior of the solid support or any
5 biopolymer absorbed thereon, and that the format (such as the dipstick or slide) is stable to any materials into which the device is introduced (such as clinical samples and hybridization solutions).

The arrays of the present disclosure can be prepared by a variety of approaches. In one example, oligonucleotide or protein sequences are synthesized
10 separately and then attached to a solid support (see U.S. Patent No. 6,013,789). In another example, sequences are synthesized directly onto the support to provide the desired array (see U.S. Patent No. 5,554,501). Suitable methods for covalently coupling oligonucleotides and proteins to a solid support and for directly synthesizing the oligonucleotides or proteins onto the support are known to those working in the
15 field; a summary of suitable methods can be found in Matson *et al.*, *Anal. Biochem.* 217:306-10, 1994. In one example, the oligonucleotides are synthesized onto the support using conventional chemical techniques for preparing oligonucleotides on solid supports (such as PCT applications WO 85/01051 and WO 89/10977, or U.S. Patent No. 5,554,501).

20 A suitable array can be produced using automated means to synthesize oligonucleotides in the cells of the array by laying down the precursors for the four bases in a predetermined pattern. Briefly, a multiple-channel automated chemical delivery system is employed to create oligonucleotide probe populations in parallel rows (corresponding in number to the number of channels in the delivery system)
25 across the substrate. Following completion of oligonucleotide synthesis in a first direction, the substrate can then be rotated by 90° to permit synthesis to proceed within a second set of rows that are now perpendicular to the first set. This process creates a multiple-channel array whose intersection generates a plurality of discrete cells.

30 The oligonucleotides can be bound to the polypropylene support by either the 3' end of the oligonucleotide or by the 5' end of the oligonucleotide. In one example, the oligonucleotides are bound to the solid support by the 3' end. However, one of skill in the art can determine whether the use of the 3' end or the 5' end of the oligonucleotide is suitable for bonding to the solid support. In general, the internal

complementarity of an oligonucleotide probe in the region of the 3' end and the 5' end determines binding to the support.

In particular examples, the oligonucleotide probes on the array include one or more labels that permit detection of oligonucleotide probe:target sequence

5 hybridization complexes.

2. Gene Expression Profiling With Microarray Methods

One of the most sensitive and most flexible quantitative methods is RT-PCR, which can be used to compare mRNA levels in different sample populations, in
10 normal and tumor tissues, with or without drug treatment, to characterize patterns of gene expression, to discriminate between closely related mRNAs, and to analyze RNA structure.

The first step is the isolation of RNA from a target sample such as human tumors or tumor cell lines, and corresponding normal tissues or cell lines,
15 respectively. If the source of RNA is a primary tumor, RNA can be extracted, for example, from frozen or archived paraffin-embedded and/or fixed (*e.g.* formalin-fixed) tissue samples.

A variation of RT-PCR is real time quantitative RT-PCR, which measures PCR product accumulation through a dual-labeled fluorogenic probe (*e.g.*, TaqMan®
20 probe). Real time PCR is compatible both with quantitative competitive PCR, where internal competitor for each target sequence is used for normalization, and with quantitative comparative PCR using a normalization gene contained within the sample, or a housekeeping gene for RT-PCR (see Heid *et al.*, *Genome Research* 6:986-994, 1996). Quantitative PCR is also described in U.S. Pat. No. 5,538,848.
25 Related probes and quantitative amplification procedures are described in U.S. Pat. No. 5,716,784 and U.S. Pat. No. 5,723,591. Instruments for carrying out quantitative PCR in microtiter plates are available from PE Applied Biosystems (Foster City, CA).

In other examples, mRNA levels are measured using TaqMan® RT-PCR technology. TaqMan® RT-PCR can be performed using commercially available
30 equipment. The system can include a thermocycler, laser, charge-coupled device (CCD) camera, and computer. In some examples, the system amplifies samples in a 96-well format on a thermocycler. During amplification, laser-induced fluorescent signal is collected in real-time through fiber optics cables for all 96 wells, and

detected at the CCD. The system includes software for running the instrument and for analyzing the data.

To minimize errors and the effect of sample-to-sample variation, RT-PCR can be performed using an internal standard. The ideal internal standard is expressed at a
5 constant level among different tissues, and is unaffected by an experimental treatment. RNAs commonly used to normalize patterns of gene expression are mRNAs for the housekeeping genes GAPDH, β -actin, and 18S ribosomal RNA.

The steps of a representative protocol for quantitating gene expression using fixed, paraffin-embedded tissues as the RNA source, including mRNA isolation,
10 purification, primer extension and amplification are given in various published journal articles (see Godfrey *et al.*, *J. Mol. Diag.* 2:84 91, 2000; Specht *et al.*, *Am. J. Pathol.* 158:419-29, 2001). Briefly, a representative process starts with cutting about 10 μ m thick sections of paraffin-embedded tumor tissue samples. The RNA is then extracted, and protein and DNA are removed. Alternatively, RNA is isolated directly from a
15 tumor sample or other tissue sample. After analysis of the RNA concentration, RNA repair and/or amplification steps can be included, if necessary, and RNA is reverse transcribed using gene specific promoters followed by RT-PCR and/or hybridization to a nucleic acid array.

In alternate embodiments, commonly used methods known in the art for the
20 quantification of mRNA expression in a sample may be used with the colon signatures provided herein. Such methods include, but are not limited to, northern blotting and *in situ* hybridization (Parker & Barnes, *Methods in Molecular Biology* 106:247 283 (1999)); RNase protection assays (Hod, *Biotechniques* 13:852 854 (1992)). Alternatively, antibodies may be employed that can recognize specific
25 duplexes, including DNA duplexes, RNA duplexes, and DNA-RNA hybrid duplexes or DNA-protein duplexes.

Further PCR-based techniques include, for example, differential display (Liang and Pardee, *Science* 257:967 971 (1992)); amplified fragment length polymorphism (iAFLP) (Kawamoto *et al.*, *Genome Res.* 12:1305 1312 (1999));
30 BeadArray™ technology (Illumina, San Diego, Calif.; Oliphant *et al.*, *Discovery of Markers for Disease* (Supplement to *Biotechniques*), June 2002; Ferguson *et al.*, *Analytical Chemistry* 72:5618 (2000)); BeadsArray for Detection of Gene Expression (BADGE), using the commercially available Luminex100 LabMAP system and multiple color-coded microspheres (Luminex Corp., Austin, Tex.) in a rapid assay for

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 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

each arrayed element allows for assessment of corresponding mRNA abundance. With dual color fluorescence, separately labeled cDNA probes generated from two sources of RNA are hybridized pairwise to the array. The relative abundance of the transcripts from the two sources corresponding to each specified gene is thus
5 determined simultaneously. The miniaturized scale of the hybridization affords a convenient and rapid evaluation of the expression pattern for colon cancer signature genes in Tables 1, 2, and 6. Microarray analysis can be performed by commercially available equipment, following manufacturer's protocols, such as are supplied with Affymetrix GeneChip® technology (Affymetrix, Santa Clara, CA), or Agilent's
10 microarray technology (Agilent Technologies, Santa Clara, CA).

3. *Additional Methods of Gene Expression Analysis*

Serial analysis of gene expression (SAGE) is another method that allows the simultaneous and quantitative analysis of a large number of gene transcripts, without
15 the need of providing an individual hybridization probe for each transcript. First, a short sequence tag (about 10-14 base pairs) is generated that contains sufficient information to uniquely identify a transcript, provided that the tag is obtained from a unique position within each transcript. Then, many transcripts are linked together to form long serial molecules that can be sequenced, revealing the identity of the
20 multiple tags simultaneously. The expression pattern of any population of transcripts can be quantitatively evaluated by determining the abundance of individual tags, and identifying the gene corresponding to each tag (see, for example, Velculescu *et al.*, *Science* 270:484-7, 1995; and Velculescu *et al.*, *Cell* 88:243-51, 1997).

In situ hybridization (ISH) is another method for detecting and comparing
25 expression of genes of interest. ISH applies and extrapolates the technology of nucleic acid hybridization to the single cell level, and, in combination with the art of cytochemistry, immunocytochemistry and immunohistochemistry, permits the maintenance of morphology and the identification of cellular markers to be maintained and identified, and allows the localization of sequences to specific cells
30 within populations, such as tissues and blood samples. ISH is a type of hybridization that uses a complementary nucleic acid to localize one or more specific nucleic acid sequences in a portion or section of tissue (*in situ*), or, if the tissue is small enough, in the entire tissue (whole mount ISH). RNA ISH can be used to assay expression patterns in a tissue, such as the expression of cancer survival factor-associated genes.

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 embodiments 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)

whose presence enables an assessment of cancer survival factor-associated gene (or protein) levels. Such an assessment includes a determination of the overall constitutive level of gene transcription and a control for variations in RNA (or protein) recovery.

5

The disclosure is further illustrated by the following non-limiting Examples.

EXAMPLES

Example 1

10 This example describes the generation and validation of an exemplary predictive tool for the categorization of colon cancer samples using the methods and reagents disclosed herein. This example includes materials published by inventors of the subject technology in Kennedy *et al.*, *J. Clin. Oncol.*, 29(35) 4620-4626, 2011, which is specifically incorporated herein by reference in its entirety.

15 A colorectal cancer transcriptome focused research array was developed (*Colorectal Cancer DSA*TM (Almac Diagnostics, N. Ireland; which can be found on the world wide web at almac-diagnostics.com)) capable of delivering accurate expression data from FFPE derived RNA (Johnston *et al.*, *J Clin. Oncol.* 24: 3519, 2006).

20 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 Center for Biotechnology Information (NCBI) human Reference Sequence (RefSeq) RNA database (which can be found on the world wide web at
25 ncbi.nlm.nih.gov/RefSeq/) using BLAST analysis, 21,968 (42%) transcripts are present and 26,676 (51%) of transcripts are absent from the human RefSeq database. Furthermore 7% of the content represents expressed antisense transcripts to annotated genes. (Johnston *et al.*, *J Clin. Oncol.* 24: 3519, 2006; Pruitt *et al.*, *Nucleic Acids Research* 33: D501-D504, 2005). In addition, probe-level analysis of the *Colorectal*
30 *Cancer DSA*TM compared with leading generic arrays, highlighted that approximately 20,000 (40%) transcripts are not contained on the leading generic microarray platform (Affymetrix) and are unique to the *Colorectal Cancer DSA*TM. Thus, the *Colorectal Cancer DSA*TM research tool includes transcripts that have not been available in hitherto performed gene expression studies. Finally, because the transcript

information used to design the *Colorectal Cancer DSA*TM was generated in part by a high throughput sequencing approach, it has been possible to generate probes closer to the 3' end of the transcripts than are contained on other generic microarrays. The combination of relevant disease specific content and 3' based probe design has yielded a unique product capable of robust profiling from FFPE derived RNA.

The aim of this study was to assess the use of the *Colorectal Cancer DSA*TM research array, using FFPE derived tumor material to generate and independently validate a prognostic gene signature capable of accurately classifying stage II colon cancer patients as being at low or high risk of relapse, post surgery. Stage II colon cancer as used in this example is AJCC T3 or T4 node negative (NO) non metastatic (MO) colon cancer.

METHODS

Sample Selection.

Samples were collected retrospectively with the following eligibility criteria: stage II colon adenocarcinoma only, with no evidence of residual disease; patient age 45 years or older at time of primary surgery; six or more regional lymph nodes assessed; a minimum of 50% tumor cells present in the tissue section; no family history of colon cancer; no preoperative or postoperative cancer therapy within 1 year of surgery (although therapy given after recurrence was acceptable); and minimum patient follow-up of 5 years for low-risk patients. Low-risk patients were defined as those with no cancer recurrence within 5 years of primary surgery. High-risk patients were defined as those with metastatic cancer recurrence within 5 years of primary surgery. Patients with local disease recurrence were excluded because this recurrence may have been a result of local residual disease after surgery rather than metastatic tumor. Samples were collected from 12 centers. All samples underwent independent histopathologic review by a pathologist. The data set was compared with the Surveillance, Epidemiology, and End Results database to ensure it represented a general population with stage II colon cancer. Key patient and tumor characteristics are given in Table 3 (see FIG. 6).

Gene Expression Profiling From FFPE Tissue.

Total RNA was extracted from FFPE tumor samples using the Roche High Pure RNA Paraffin Kit (Roche, Basel, Switzerland). Amplified cDNA targets were prepared using the Nugen WT-Ovation® FFPE System v2 in combination with the Nugen FL-Ovation® cDNA Biotin Module v2 and were performed in accordance

with manufacturer's instructions. Hybridization, washing, staining and scanning of fragmented, labelled cDNA was carried out according to standard Affymetrix protocols. Between 3.0 and 3.5µg fragmented, labelled cDNA was hybridized to the *Colorectal Cancer DSA™* microarray (Almac, Craigavon, United Kingdom) on the

5 Affymetrix 7G scanner (Affymetrix, Santa Clara, CA). A sample profile scheduling strategy was used that involved the stratification of samples into batches that were randomized against targeted clinical and sample property factors in addition to operators, reagent, and material lots. Quality control criteria were applied, and biologic and technical factors were balanced between low- and high-risk samples.

10 This is performed in order to minimise systematic bias and diffuse any residual technical bias into technical variation.

Classifier Model Identification.

Model development started with 5,014 probe sets identified as stable and/or having comparable longitudinal stability under FFPE fixation to avoid the issue of

15 differential degradation of probe sets. Signature generation was subsequently performed using the partial least squares classification method with selection of important features based on recursive feature elimination (RFE) during 10 repeats of five-fold cross validation. All aspects of the model development were appropriately nested within the cross validation, including an initial filtering to remove 50% of the

20 probe sets with the lowest variance and intensity, reference-based robust multichip averaging (RefRMA) normalization and summarization, and RFE discarding the least important 10% of probe sets at each iteration. The total number of features to include in the final model was determined by the feature length with the highest average area under the receiver operating characteristics curve (AUC) under cross validation. The

25 threshold for dichotomization of the predictions from each model was selected based on the maximum of the sum of sensitivity and specificity (minimum of the Youden J statistic (Youden, *Cancer* 3:32-35, 1950) from cross-validated training data. In the case of multiple thresholds with largely identical performance, the hazard ratio (HR) from Cox proportional hazards regression was used as a tiebreaker to favor higher HR

30 values.

The precision of the predictions was evaluated by predicting technical replicates of a colorectal cancer cell line (HCT116) embedded in FFPE, which was profiled concurrently with the clinical samples. The repeated technical measurements of this sample were not included in model development but were predicted by all 50

cross-validation training subsets as an independent test set with a view to select models with high repeatability and reproducibility. Additionally, a permutation test was performed where the true class labels were reshuffled randomly 100 times followed by complete model development. This was done to assess what

5 classification performance one can expect by chance from a data set with these characteristics and to reveal any bias in the signature generation procedure.

The independence of the final model in the context of known clinical factors was evaluated using univariate and multivariate Cox proportional hazards regression. The input used was the predicted dichotomized class labels together with tumor stage,

10 patient tumor grade, tumor location, patient age, patient sex, mucinous/nonmucinous subtype, and number of lymph nodes retrieved. Microsatellite instability was not included as a factor because this information was not available for the majority of the samples. Gene Ontology annotation and enrichment of Gene Ontology biologic processes and molecular functions were performed using an internally developed tool

15 based on the genes in the final signature. The hypergeometric distribution with false discovery rate multiple testing correction was used to determine functional classes of genes significantly enriched. The pathway analysis was generated through the use of Ingenuity Pathway Analysis (Ingenuity Systems, Redwood City, CA).

Balancing, randomisation and Quality Control (QC) of samples.

20 *Target population:* The population used to train the assay was matched to reflect the general population properties from the SEER and CRUK databases. The following properties were being considered:

- **Gender.** The gender prevalence amongst in the Stage II population is approximately between 50-60% male (56% in the UK and 57% in the US).
- 25 • **Tumour location** (distal/proximal). The prevalence in the Stage II population is approximately 55%-65% proximal and 35%-45% distal.
- **Patient age.** According to NCI's SEER Cancer Statistics Review Colon and Rectum Section, from 2001-2005, 0.1% of patients were diagnosed under age 20; 1.0% between 20 and 34; 3.7% between 35 and 44; 11.6% between 45 and
- 30 54; 18.3% between 55 and 64; 25.1% between 65 and 74; 28.2% between 75 and 84 and 12.2% 85+ years of age.
- **Recurrence-free survival rate.** The rate of recurrence-free survival in the Stage II population is reported to be between 13%-22% (Gattaj *et al*, European Journal of Cancer, 2006) and ~30% from the SEER database.

Pre-balancing: Pre-balancing was performed so that the sample set put forward for hybridization was balanced with respect to selected clinical covariates whilst maintaining the general population statistics presented above. This excludes recurrence-free survival, which was intentionally enriched to increase the power of the biomarker discovery. The training set did not contain any samples with events after 5 years, whereas this was not a constraint in the validation set. The rationale for not using samples that recur after 5 years for signature generation (i.e. in the training set) is to avoid introducing additional heterogeneity in the sample population when performing the biomarker discovery.

The main aim of the balancing procedure was to reduce the association (if any) between the endpoint (high/low risk represented as a binary variable) and any of the factors listed below. Any association between these factors and the high/low risk endpoint would introduce a confounding that could limit the clinical utility of the assay. 603 colorectal samples were subjected to the pre-balancing in order to reduce strong associations between prognosis and any of the following factors: Gender; Tumor location within the bowel; Patient age; Contributing Centre; FFPE block age (date of surgery); Tumor content; and RNA quality.

Continuous parameters were tested using a Kolmogorov-Smirnoff test and categorical parameters were tested using a chi-squared test. A p -value ≥ 0.4 for all parameters was required to achieve balancing. 504 samples remained after balancing (335 low risk and 169 high risk) and were put forward to array profiling.

Randomization of samples during array profiling: Randomization of samples was performed to avoid confounding between known technical and biological factors, primarily the endpoint of interest (prognosis). In this study operator, hybridization-wash-stain (HWS) kit lot, array lot and array batch were considered together with the contributing center and the prognosis. Samples were first randomized into array batches such that each array batch had the same proportion of prognosis and contributing center. Operators were then assigned to each array batch according to availability. Each array batch was then assigned a HWS kit, ensuring that each operator used the same proportion of each kit. Array lots were allocated to each array batch, ensuring that they were evenly distributed amongst the array batches.

Quality control of the training data: QC procedures were applied on the resulting arrays, primarily based on values in the Affymetrix RPT files that contain various quality-related parameters. Limits were calculated based on visual inspection

of the distribution for each parameter for all samples: % present calls ($\geq 20\%$ required); Image artefacts were identified to remove arrays with noticeable blotches ; Outliers were detected from principal component analysis (PCA) based on the Q residuals and Hotelling's T^2 .

- 5 Assessment of gender genes was used to determine if observed expression levels matched the known gender in clinical information

The following Affymetrix quality parameters were also considered during the visual inspection of the distributions; broadly categorised as follows: RNA Quality; Signal Quality & Detection call; Background & Noise; and Background Homogeneity

- 10 A total of 319 colorectal samples passed the QC procedure. Due to preliminary results suggesting a heterogeneity introduced by the rectal samples, the rectal samples were removed to form a 249 colon-only set, which was put forward for final (post-QC) balancing.

- 15 *Final post-QC balancing:* The 249 colon samples passing QC were balanced using the same principles as the initial pre-balancing, with the addition of criteria for the % present call distribution to be similar in both low risk and high risk groups (this information is only available after hybridisation). A final set of 215 samples remained after QC and balancing.

- 20 The final colon set with 215 samples has the following properties compared to the known population distribution: Gender: 53% male (50-60% in population); Tumor location (distal/proximal): 62% proximal (55-65% proximal in population); Patient age: Closely follows the continuous distribution of the population; and Recurrence-free survival rate: 34% poor prognosis (high risk). Intentionally enriched compared to population around 15-20%.

- 25 *Quality control of the validation set and future sample sets:* Using a tailor-made QC procedure on the training set is an important step in order to facilitate the identification of biomarkers from a high-quality data set. However for prediction of future samples, QC has to be applied on a one-sample-at-a-time basis. Also, the QC procedure cannot be too specific to the data set and the system where the data has
30 been generated. For this purpose a separate evaluation was performed using 40 samples replicated across two systems and scanners to identify QC parameters that are stable across systems. The AvgSigA parameter (average signal of the absent probe sets) was determined to be the most stable parameter across the different systems and hence the best candidate for a system-independent QC procedure. For this parameter,

higher values imply lower quality and lower values imply higher quality. The AvgSigA values are strongly negatively correlated to the % present call parameter which is a commonly used QC parameter and was the primary QC parameter used on the training set. The lower acceptance value of % present calls from the training set was set to 20%, which corresponds approximately to an upper acceptance value of 43 for the AvgSigA parameter for this data set. To accommodate younger FFPE samples, it was decided not to introduce a lower threshold on the AvgSigA (which will allow inclusion of higher-quality samples). The final inclusion range derived from this study was hence $\text{AvgSigA} \leq 43$, which was the QC metric applied to the independent validation set and is the QC that will be applied to future samples.

Identifying probe sets that are stable over FFPE block age: It was recognized that mRNA transcripts are likely to degrade at different rates and to different levels in FFPE samples, which could result in a signature generated from old material not performing as expected on fresh FFPE material. Therefore two independent longitudinal studies were performed to identify probe sets that are stable over FFPE block age. In the first study, 9 FFPE blocks were serially sectioned and analyzed by DNA microarray at seven time points in a 16-week timeframe following fixation. These samples were supplemented by a second longitudinal study at three 6 month intervals in a one year timeframe in which 8 FFPE blocks ranging from 6 months to 4 years of age which were serially sectioned and analyzed by DNA microarray resulting in 113 individual samples for analysis. 5014 transcripts were identified that either did not undergo further degradation with time or decayed at an equivalent rate following fixation. This list of probe sets was subsequently used for signature generation. A separate manuscript for presenting the details from this study is in preparation.

Estimating the precision of the classifier during model development: The ability of a classifier to consistently produce the same output from technical replicates is an important aspect of an assay when used in a test setting. For this purpose, a set of 39 reference samples, which are technical replicates of the same colorectal cancer cell line (HCT114), were hybridized together with the clinical samples. During model development, this set was predicted as an external test set during cross-validation in order to estimate the relative variance at each step in the model development process. No information was shared between the training set and the 39 sample reference set during cross-validation. The standard deviation from the predicted signature scores were calculated and visualized as the average with 95% confidence limits. The

variability is low for longer signatures, which then gradually increases over the feature selection procedure, which is also reflected in lower accuracy (AUC) for the shorter signatures. At the selected signature length (634 probe sets), the model shows both high precision and accuracy.

5 *Permutation analysis of the classification performance:* Permutation analysis was performed to evaluate what classification performance one can expect by chance from a data set with similar properties. This was performed by randomly reshuffling the true class labels (i.e. the true prognosis) and subsequently repeating the entire model development process (with filtering, normalization, feature selection and
10 classification). The signature performance is significantly better than chance at longer signature lengths and specifically at the selected one where the number of probe sets is 634. Additionally, the permutation test reveals any underlying bias in the data set and/or the methodology used to develop the classifier. The median AUC over the random labels is 0.5, denoting chance, which confirms that there is no evident bias in
15 the procedure used.

RESULTS

Development of a Stage II Colon Cancer Prognostic Signature From FFPE

Tissues. Disease-free survival at 5 years was used as the primary end point for this study. After balancing for clinical factors and applying quality control criteria to the
20 initial data set, a training set of 215 patients (142 low-risk and 73 high-risk patients) was identified. Fifty percent variance-intensity filtering, RefRMA normalization, RFE feature selection, and partial least squares classification were performed under 10 repeats of five-fold cross validation for estimation of the classification performance. Cross validation indicated a 634-transcript signature to be optimal for prognostic
25 classification. A receiver operating characteristic curve with an AUC of 0.68 ($P < .001$) was generated, indicating a significant association between signature score and prognosis (FIG. 3A). The observed AUC was significantly higher than random in the permutation analysis and displayed a low variance in the evaluation of the precision from technical replicates. A threshold of 0.465 for dichotomization of the signature
30 prediction scores was established from the Youden J statistics, yielding an HR of 2.62 ($P < .001$; FIG. 3B). Table 4 contains a summary of the classification performance over the signature generation during cross validation.

Table 4. Classification Performance of the Training and Independent Validation Sets

Data Set	AUC	Sensitivity	Specificity	NPV	PPV	HR
----------	-----	-------------	-------------	-----	-----	----

Train(95 % CI)	0.682(0.6 43-0.720)	0.478(0.4 07-0.549)	0.791(0.73 7-0.845)	0.858(0.8 45-0.872)	0.365(0.3 17-0.413)	2.618(2.0 41-3.195)
Val.(95 % CI)	0.684(0.5 94-0.761)	0.718(0.6 17-0.811)	0.559(0.42 3-0.673)	0.867(0.8 28-0.900)	0.331(0.2 50-0.434)	2.526(1.5 36-4.154)

The 95% CIs are ± 2 standard deviations from cross validation (training set) or bootstrapping with 1,000 repeats (validation set); 80% and 20% priors have been used when calculating the NPVs and PPVs, respectively. The threshold $t = 0.465$ was used for dichotomization of the signature score. Abbreviations: AUC, area under the receiver operating characteristics curve; HR, hazard ratio; NPV, negative predictive value (negative is low risk); PPV, positive predictive value (positive is high risk)

Independent Validation of the Stage II Colon Cancer Prognostic Signature:

The prognostic signature was applied to an independent validation set of 144 patients enriched for recurrence (85 low-risk and 59 high-risk patients) using the threshold score identified in the training set. The sample analysis was run separately and at a later time to the training set. The signature predicted disease recurrence with an HR of 2.53 ($P < .001$) in the high-risk group (FIG. 4 and Table 4). The signature also predicted cancer-related death with an HR of 2.21 ($P < .0084$) in the high-risk group (FIG. 5).

The fact that the signature described herein was developed from FFPE derived tumor material facilitates a large scale validation strategy based on retrospective analysis of existing FFPE tumor banks.

The hazard ratio is an expression of the hazard or chance of events occurring in the stage II colon cancer patients identified by the classifier as high risk as a ratio of the hazard of the events occurring in the patients identified by the classifier as low risk. There was a significantly lower probability of recurrence for the group predicted to have good prognosis compared to those predicted to have poor prognosis, within 5-years post surgery. The negative predictive value is the proportion of patients with negative test results who are correctly diagnosed (predicted negative). In a prognostic setting, the NPV is dependent on the prevalence of disease recurrence. The positive predictive value is the proportion of patients with positive test results who are correctly diagnosed (predictive positive). In a prognostic setting, the PPV is dependent on the prevalence of disease recurrence. Based on a population prevalence of 20% poor prognosis samples, this would imply that patients with a predicted poor

prognosis have a 33% probability of recurrence whereas patients with a predicted good prognosis have a 13% probability of recurrence within 5 years.

Assessment of Signature Independence From Known Prognostic Factors:

For a prognostic assay to be useful, it must perform independently from known prognostic factors used in the clinic. Therefore the independence of the assay was assessed in both a univariate and multivariate analysis (Table 5).

Table 5. Comparison of Transcript Signature to Standard Pathologic Parameters in the Independent Validation Set

		Univariate			Multivariate		
		HR	CI	p	HR	CI	p
Tumor Stage	(T4 vs T3)	1.23	0.667-2.269	0.5067	1.617	0.84-3.110	0.1501
Patient Age		1.039	1.01-1.069	0.0086	1.046	1.014-1.078	0.0041
Tumor Grade	II	0.815	0.456-1.456	0.4895	1.274	0.48-3.383	0.6265
	III	1.326	0.654-2.689	0.434	2.161	0.636-7.339	0.2169
Tumor Location (Proximal vs Distal)		1.766	1.075-2.901	0.0248	2.158	1.224-3.804	0.0078
Gender		1.165	0.713-1.901	0.5426	0.971	0.549-1.720	0.9204
Mucinous subtype		0.825	0.418-1.627	0.5787	0.896	0.433-1.856	0.7682
No. of Nodes Retrieved		1.007	0.983-1.032	0.5678	1.014	0.988-1.041	0.2824
Prognostic Signature		2.526	1.536-4.154	< 0.001	2.551	1.471-4.423	< 0.001

Both the univariate and multivariate analyses have been performed using Cox

proportional hazards regression with *P* values coming from a log-likelihood test. For tumor grade, grade 1 has been used as the reference point for calculating the HR. Patient age and number of nodes retrieved are analyzed as continuous factors. The interpretation of the HR of patient age is the increased risk for a change in 1 year of age, and correspondingly, the interpretation of the HR of number of nodes retrieved is the increased risk for an increase of one retrieved node. Abbreviation: HR, hazard ratio.

The prediction of prognosis was significant in both the univariate ($P < .001$) and multivariate ($P < .001$) analysis, demonstrating that the signature provided prognostic information in addition to conventional risk factors. Furthermore, the independence of the signature was assessed with the addition of lymphovascular invasion in the samples where this had been recorded (100 of 144 samples in the validation set). The signature performed independently in the univariate ($P < .001$) and multivariate analysis ($P < .001$).

Functional Analysis of the Genes in the Prognostic Signature: Next it was asked if the assay detected biologic processes known to be relevant to colon cancer recurrence. The 634 probe sets were analyzed using Ingenuity Pathway Analysis, and a list of statistically significant pathways were identified, the most significant of which was *IGF-1* signaling.

DISCUSSION

As disclosed herein a DNAmicroarray-based assay was developed that identifies patients at higher risk of recurrence after surgery for stage II colon cancer. Specifically, the signature identified a high-risk cohort with an HR of recurrence of 2.53 and an HR of cancer-related death of 2.21 in an independent validation set. Validation of a prognostic assay using a completely separate set is necessary to avoid overestimations of the performance of the signature from the training set. The HR of 2.53 for recurrence compares favorably with histologic factors currently used to make decisions in the clinic, which typically have anHR of approximately 1.5 or less. Moreover, the signature does not require individual interpretation and may offer a more standardized approach than conventional histopathologic factors. Importantly, the assay is performed on FFPE tissue and, therefore, is easily applied in current medical practice.

Although several DNAmicroarray-based prognostic tests in several cancer types have been published, only one has been introduced into clinical practice, and to date, none is used in colon cancer. This may be a result of two major factors. First, many of the signatures have been developed from fresh or frozen tissue. Second, inappropriate study methodology has resulted in a failure to validate the test in an independent data set.

Regarding the use of frozen tissue samples, although this tissue type provides excellent microarray data, a test generated from this tissue is unlikely to perform adequately in FFPE tissue. This can create difficulty in collecting enough samples to

develop and independently validate a prognostic test. In addition, implementation of fresh tissue– based assays requires a change in clinical practice, because samples need to be collected at the time of surgery.

5 FFPE is the standard for tumor archiving, and numerous tumor banks already exist for assay development. Importantly, no change in sample collection and processing is required for the development and clinical implementation of FFPE-based assays.

10 The disclosed methods were developed to work with FFPE tissue but using a DNA microarray platform, thereby vastly increasing the number of detectable mRNA transcripts and biologic processes relative to quantitative polymerase chain reaction technology. As a result of using FFPE material with a microarray platform, several methodologic issues needed to be considered. Formalin fixation results in the degradation of mRNA transcripts through the cross linking of RNA to protein. Most of this degradation occurs immediately, but some transcripts continue to degrade with time. The DNAmicroarray platform used for the study has probe sets designed to the 3' end of mRNA transcripts to enhance the ability to detect degraded transcripts. In addition, a separate set of colon cancer samples was analyzed over time to ensure we did not incorporate probe sets that detected unstable or differentially stable mRNA transcripts as part of the signature.

20 The predictive value of the signature is above and beyond known prognostic clinical covariates. This performance can largely be attributed to the initial balancing of prognosis against biologic and technical factors that was performed as part of establishing a suitable training set. Biologic factors considered include known prognostic factors such as pT stage and grade, as well as other nonprognostic factors that may have affected gene expression including tumor location, patient age, and sex. 25 Technical factors such as FFPE block age and the contributing center were also balanced between high- and low-risk samples in the training set. In addition, randomization of operators and reagent kits was performed to avoid confounding between technical factors and known clinical factors. This minimized the risk that the assay was dependent on the operator or relied on the use of samples from specific 30 centers or the use of specific batches of reagents. Because the assay was developed to be independent from known prognostic factors, we believe that it may be possible to develop a multiparametric test that incorporates several factors to give an even more accurate prognostic indicator.

Functional analysis of the gene signature revealed that *IGF-1* signaling, *TGF- β* signaling, and *HMGB1* signaling were among the most significant pathways identified. All of these have been previously reported to confer a poor prognosis in colon cancer through promoting tumor growth, invasion, and metastasis and preventing apoptosis. In conclusion, disclosed herein is a validated and robust prognostic DNA microarray signature for stage II colon cancer from FFPE stored tumor tissue.

The disclosed signature can help physicians to make more informed clinical decisions regarding the risk of relapse and the potential to benefit from adjuvant chemotherapy. (Andre *et al.*, *Annals of Surgical Oncology* 13:887-898, 2006; Diaz-Rubio *et al.*, *Clin. Transl. Oncol.* 7: 3-11, 2005; Monga *et al.*, *Ann. Surg. Oncol.* 13: 1021-1134, 2006; Sobrero, *Lancet Oncol.* 7: 515-516, 2006). Furthermore, many patients want to know their likelihood of cure and the risks/benefits of treatment. (Gill *et al.*, *J. Clin. Oncol.* 22: 1797-1806, 2004; Kinney *et al.*, *Cancer* 91: 57-65, 2001; Carney *et al.*, *Ann. R. Coll. Surg. Engl.* 88: 447-449, 2006; Salkeld, *Health Expect* 7: 104-1014, 2004). Being able to predict the patient's prognosis provides the physician and the patient with a better assessment of the risks/benefits and the choice of therapy. The ability to offer individualized patient care will hopefully result in improved survival and quality of life for these patients.

In the past, many studies have implicated sample size as the primary reason for lack of convincing statistical evidence and point to larger trials being required to prove the benefit of adjuvant treatment. Using validated prognostic markers, such as the gene signature generated in this study, stage II patients can be stratified into high and low risk sub-populations. This approach may assist in improved clinical trial design by focusing on those patients at high risk of recurrence and therefore more likely to derive a benefit from adjuvant therapy. Thus, the *Colorectal Cancer DSA*TM may be a useful research tool for stratifying patients for inclusion in clinical trials, for decision-making regarding adjuvant and neo-adjuvant treatment, and for the identification of novel pathways or molecular targets for additional drug development

The prognostic signature reported in Table 6 accurately predicted for relapse for stage II colon cancer and is evaluated on an independent FFPE validation set. The overall accuracy for prediction of recurrence was substantial for this heterogeneous disease. Based on a population prevalence of 20% poor prognosis samples, this would imply that patients with a predicted poor prognosis have a 33% probability of

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 μ g 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 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
10 of a diagnosis of colon cancer, wherein the expression level, or a decision score derived 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:

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 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
25 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 same side of the threshold permits classification of the colon cancer sample.

30 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 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 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.

10 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, comprising:

15 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 indicative of a known response to treatment, wherein the expression level, or a decision score
20 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:

25 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 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:

5 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 indicative
10 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 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:

15 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)
20 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 survival.

25

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

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
30 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.

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
5 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
10 for a subject, comprising:
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
15 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.

20

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

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

25

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
30 is determined with PCR and/or or microarray-based methods.

23. The method of any one of claims 1-22, wherein detecting the expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 comprises detecting the expression levels for MUM1 and SIGMAR1

transcripts.

24. The method of any one of claims 1-23, wherein detecting the expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 comprises detecting the expression levels for MUM1, SIGMAR1, ARSD, SULT1C2 and PPFIBP1 transcripts.

25. The method of any one of claims 1-24, wherein detecting the expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 comprises detecting the expression levels for ARSD, CXCL9, PCLO, SLC2A3, FCGBP, SLC2A14, SLC2A3, BCL9L and antisense sequences of MUC3A, OLFM4 and RNF39 transcripts.

26. The method of any one of claims 1-25, wherein detecting the expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 comprises detecting the expression levels for the transcripts listed in Table 1.

27. The method of any one of claims 1-26, wherein detecting the expression level of at least 2 colon cancer-related nucleic acid molecules listed in Table 6 comprises detecting the expression levels for the transcripts listed in Table 2.

28. 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: 1-636 or its complement.

29. The nucleic acid probe according to claim 28, wherein the probe is labeled.

30. The nucleic acid probe according to claim 29, wherein the probe is radiolabeled, fluorescently-labeled, biotin-labeled, enzymatically-labeled, or chemically-labeled.

31. A set of probes, for detecting a gene expression signature for colon

cancer, 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: 1-636 or its complement.

5

32. The set of probes of claim 31, wherein the probes are labeled.

33. The set of probes of claim 32, wherein the probes are radiolabeled, fluorescently-labeled, biotin-labeled, enzymatically-labeled, or chemically-labeled.

10

34. The set of probes of any one of claims 31-33, wherein the set contains at least one probe complementary to each of the transcripts listed in Table 6.

15

35. The set of probes of claim 34, comprising at least one probe complementary to each transcript in a subset of the transcripts listed in Table 6, wherein the subset comprises at least 1%, 5%, 10%, 25%, 50%, 75% or 95% of the transcripts in Table 6.

20

36. 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 31-35.

25

37. 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: 1-636 or its complement; and

30

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.

38. 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: 1-636.

39. The set of primer pairs of claim 38, wherein the set of primer pairs
5 comprises primer pairs suitable for use in amplification of a subset of nucleic acid sequences set forth as SEQ ID NOs: 1-636 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.

10 40. 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 at least 50 transcripts listed in Table 6 are detected.

15

41. The method of claim 40, wherein from 400 to 800 transcript expression levels are detected.

42. The method of claim 40 or 41, wherein from 500 to 700 transcript
20 expression levels are detected.

43. The method of any one of claims 40-42, wherein at least 100 transcripts from Table 6 are detected.

25 44. The method of any one of claims 40-43, wherein at least 200 transcripts from Table 6 are detected.

45. The method of any one of claims 40-44, wherein at least 300 transcripts from Table 6 are detected.

30

46. The method of any one of claims 40-45, wherein at least 400 transcripts from Table 6 are detected.

47. The method of any one of claims 40-46, wherein at least 500

transcripts from Table 6 are detected.

48. The method of any one of claims 40-47, wherein at least 600 transcripts from Table 6 are detected.

5

49. The method of any one of claims 40-48, wherein all transcripts listed in Table 6 are detected.

50. The method of any one of claim 40-49, wherein the transcripts listed in Table 6 comprise the transcripts listed in Table 1.

10

51. The method of any one of claims 40-50, wherein the colon cancer specimen is a formalin fixed paraffin-embedded tissue sample.

52. The method of any one of claims 40-51, 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.

15

53. The method of claim 52, further comprising, selecting adjuvant chemotherapy where the patient is determined to be in the high risk group.

20

54. 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 at least 50 transcripts listed in Table 6, or a decision score derived therefrom, in a low risk or high risk group.

25

55. The method of claim 54, wherein the level of less than 1000 transcripts are detected in the gene expression profile.

30

56. The method of claim 54 or 55, wherein the level of less than 800 transcripts are detected in the gene expression profile.

57. The method of any one of claims 54-56, wherein the level of less than 700 transcripts are detected in the gene expression profile.

5 58. The method of any one of claims 54-57, wherein the specimen is classified based on the expression level of at least 100 transcripts from Table 6.

59. The method of any one of claims 54-58, wherein the specimen is classified based on the expression level of at least 200 transcripts from Table 6.

10

60. The method of any one of claims 54-59, wherein the specimen is classified based on the expression level of at least 300 transcripts from Table 6.

15 61. The method of any one of claims 54-60, wherein the specimen is classified based on the expression level of at least 400 transcripts from Table 6.

62. The method of any one of claims 54-61, wherein the specimen is classified based on the expression level of at least 500 transcripts from Table 6.

20 63. The method of any one of claims 54-62, wherein the specimen is classified based on the expression level of at least 600 transcripts from Table 6.

64. The method of any one of claims 54-63, wherein the expression level of all transcripts listed in Table 6 are used to classify the specimen.

25

65. The method of any one of claims 54-64, wherein the transcripts listed in Table 6 comprise the transcripts listed in Table 1.

30 66. The method of any one of claims 54-65, wherein the colon cancer specimen is a formalin fixed paraffin-embedded tissue sample.

67. The method of any one of claims 54-66, further comprising, selecting adjuvant chemotherapy where the patient is determined to be in the high risk group.