



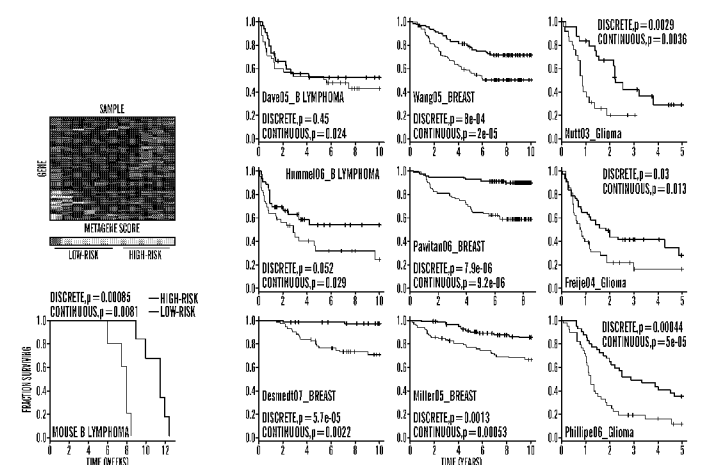
- (51) **International Patent Classification:**
C12Q 1/68 (2006.01) *C12N 15/11* (2006.01)
- (21) **International Application Number:**
PCT/US2009/000695
- (22) **International Filing Date:**
3 February 2009 (03.02.2009)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
61/029,730 19 February 2008 (19.02.2008) US
61/029,990 20 February 2008 (20.02.2008) US
- (71) **Applicant (for all designated States except US):** THE JACKSON LABORATORY [US/US]; 600 Main Street, Bar Harbor, ME 04609 (US).
- (72) **Inventors; and**
- (75) **Inventors/Applicants (for US only):** MILLS, Kevin [US/US]; 41 Meadow Drive, Bar Harbor, ME 04609 (US). WOO, Yong [US/US]; 10 Brae Drive, Ellsworth, ME 04605 (US).
- (74) **Agents:** RESNICK, David, S. et al.; Nixon Peabody LLP, 100 Summer Street, Boston, MA 02110-2131 (US).

(81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report (Rule 48.2(g))

(54) **Title:** DIAGNOSTIC AND PROGNOSTIC METHODS FOR CANCER**FIG. 4**

(57) **Abstract:** Described herein are methods for determining or predicting a prognosis for an individual having cancer. Also described are methods of determining a type or sub-type of a cancer. The methods are based upon the identification of genes or gene profiles that correlate with the aggressiveness and/or type of a cancer. The methods involve the measurement of the expression of a set of prognostic genes in a subject's tumor. In some embodiments, a metagene score is determined for the subject's tumor, and the metagene score is compared with a set or database of reference metagene scores from individuals with known diagnosis, prognosis or outcome. In such embodiments, the distance of the subject's metagene's score from the median reference metagene score is predictive of the relative aggressiveness of the subject's tumor. Also provided are methods of identifying genes prognostic or diagnostic for a given disease or disorder.

DIAGNOSTIC AND PROGNOSTIC METHODS FOR CANCER

FIELD OF THE INVENTION

[001] The present invention relates to diagnostic and prognostic predictors of cancer. Another aspect of the present invention relates to cancer biomarker discovery.

GOVERNMENT SUPPORT

[002] This invention was made with government support under CA115665 and RR018789 awarded by the National Institutes of Health (NIH). The Government of the United States has certain rights in this invention.

CROSS REFERENCE TO RELATED APPLICATIONS

[003] This application claims priority to US provisional applications 61/029,730 filed February 19, 2008 and 61/029,990 filed February 20, 2008, each of which is incorporated herein by reference in its entirety.

BACKGROUND

[004] A key challenge in the clinical management of cancers, and other diseases, is the identification of diagnostic and prognostic indicators and indicators that predict how individual patients are likely to respond to specific therapeutic regimens. The formulation of better diagnostic methods and the development of better individualized treatment strategies demand a detailed understanding of the origin, etiology, and function of cancer associated chromosomal abnormalities.

[005] Cancer is often characterized by chromosomal abnormalities, changes in gene expression and gene copy number. Loss of heterozygosity (LOH) is a common event in cancers (Tischfield and Shao, 2003). Nonhomologous end joining (NHEJ) is a critical pathway for DNA double strand break (DSB) repair. Cells with mutations in NHEJ genes display chromosomal instability, hypersensitivity to ionizing radiation, increased cellular senescence and proliferation defects. Many hematologic malignancies have been characterized for their chromosomal abnormalities, which often have clinically predictive value. Chromosomal translocations involving immunoglobulin loci are among the most commonly observed cytogenetic lesions in B-lymphoid lymphomas and leukemias, and often directly promote transformation by the activation of cellular proto-oncogenes (Mills et al., 2003). Rearrangement or structural abnormalities at chromosomes 4q21 and 11q23 are a very common theme in a variety of lymphoid neoplasms, including diffuse large cell lymphoma

(DLCL), mantle cell lymphoma (MCL), chronic lymphocytic leukemia (CLL), acute lymphoblastic leukemia (ALL), and Richter's syndrome. A common translocation partner for 11q23 is chromosome 4q21, very frequently noted in pediatric ALL, where t(4;11)(q21;q23) serves as an important prognostic indicator (Pui et al., 2004; Rooney et al., 2004a). The most common pediatric cancer, preB-ALL, shows chromosomal abnormalities at 4q21 and 11q23. Pre-B cell acute lymphoblastic leukemia (preB-ALL) represents the most common overall childhood cancer, accounting for approximately 25% of the total cancer burden in patients under 15 years of age, and approximately 35% of total cancer cases in children age 2-3 years (NCI) (Biondi et al., 2000; Carroll et al., 2003a; Pui et al., 2004). In the United States there are roughly 2400 newly diagnosed cases of ALL in children each year (NCI).

[006] A number of common chromosomal abnormalities have been identified in pediatric B-ALL, and generally define different biological subsets of patients. Approximately one third of ALL cases exhibit supernumerary chromosomes, with hyperdiploidy predicting positive 5-year event free survival (EFS) rates as high as 90%. In addition to increased chromosome number, several distinct types of chromosomal translocations are associated with childhood ALL (Carroll et al., 2003b). The t(1;19)(q23;q13) translocation results in fusion of the E2A and PBX genes, and is generally associated with a positive response to therapy. Translocations involving chromosome 11q23, to any of a variety of chromosomal targets, result in oncogenic fusions of the MLL gene (11q23) to one of more than 30 possible partners. The most common translocation seen in infant ALL patients, (t4;11)(q21;q23), results in an MLL/AF4 fusion oncogene. Rearrangements at 11q23 occur in approximately 6% of all ALL cases, and are associated with poor prognosis, showing 5 year EFS rates as low as 20% in patients under 3 years of age at the time of diagnosis (Armstrong et al., 2003; Biondi et al., 2000; Pui et al., 2003). In addition to 11q23 translocations, 11q23 deletions have been noted, and these tend to associate with more complex karyotypes, although the reasons for this are currently unknown (Moorman et al., 2005). Thus, deletions appear to demarcate a subset of ALL cases biologically distinct from those with 11q23 translocations. It is clear that the subset of ALL with del(11q23) is functionally distinct, and therefore elucidation of the mechanisms by which these different lesions arise, as well as the unique downstream consequences, are important for understanding their role in the initiation or progression of lymphoma and leukemia.

[007] Deletions at 11q23 are also frequently associated with Richter's syndrome, which describes the transformation of chronic lymphocytic leukemia (CLL) to aggressive, high grade non-Hodgkin's lymphoma (NHL) (Robak, 2004; Tsimberidou and Keating, 2005; Yee

et al., 2005). Richter's transformation is associated with a sudden and rapid clinical deterioration, possible splenomegaly, rapid growth of a lymphoid mass, usually at a single site, and other variable characteristics, such as occasional leptomeningeal disease, or extranodal involvement. While del(11q23) is correlated with Richter's transformation, the role of this lesion in the pathogenesis of the disease is unknown. Several observations point to a role for corrupted DNA repair and cell cycle control in the etiology of Richter's transformed NHL. These include loss of the tumor suppressor p53, amplification of the c-myc proto-oncogene, loss of heterozygosity for Atm, which encodes an important DNA damage responsive factor, and gene expression profiles that associate del(11q23) with aberrations in multiple cell cycle control and genome stability pathways (Tsimberidou and Keating, 2005; Yee et al., 2005).

[008] It has long been recognized that genomic instability represents one of the hallmarks of a wide variety of tumor types, including lymphoid tumors. Genomic instability can be increased by mutations in DNA polymerases, DNA repair enzymes (mismatch repair enzymes, nucleotide-excision repair enzymes), proteins that are involved in chromosomal stability (chromatin structure and condensation proteins, kinetochore proteins, spindle proteins) and proteins that regulate apoptosis and cell cycle in response to DNA damage (p53 and pRb) (Beckman and Loeb, 2005).

[009] Genomic instability has been reported (Tsafrir et al., 2006) for patients deficient in the double strand break (DSB) repair proteins Artemis (Art) and DNA ligase IV (lig4), which show an increased lymphoid cancer incidence (Moshous et al., 2003). Further, genomic instability plays a role in breast and ovarian cancer, where BRCA1 and BRCA2 mutations establish cancer predisposition (Shivji and Venkitaraman, 2004). Other syndromes which are caused by defects in DNA repair show increased cancer occurrence, e.g. Xeroderma pigmentosum, hereditary non-polyposis colon cancer and Bloom's syndrome (Risinger and Groden, 2004).

[010] For solid tumors it has been suggested that chromosomal copy number correlates with disease outcome and further it has been suggested that the expression of specific genes are involved in the disease process (Popescu and Zimonjic, 1997). With the advance of microarray gene expression profiling it has become feasible to analyze the relationship between copy number and gene expression of tumors, e.g. prostate, glioblastoma, multiple myeloma and colon carcinoma (Nigro et al., 2005; Walker et al., 2006; Wolf et al., 2004) (Tsafrir et al., 2006). With a loss of genomic regions loss of gene expression can be assumed,

however the correlation between copy number and gene expression is not simple for tumorigenesis and tumor progression.

[011] It has been demonstrated that gene ontology and expression profiles within tissues of interest can be used to further define candidate genes, however no simple method has been established to utilize both, copy number data and gene expression data for diagnostic and prognosis of cancer. Bussey et al. have shown that with changes in gene copy numbers, genes within the affected region also show changes in expression and that combining copy number and expression data improves the diagnostic and prognostic interpretation (Bussey et al., 2006).

SUMMARY OF THE INVENTION

[012] One aspect of the present invention relates to a group of cancer biomarkers and subgroups thereof for the prognosis of cancer and diagnosis of a cancer type in a subject. Another aspect of the present invention relates to a method to identify disease biomarkers, in diseases associated with chromosomal abnormalities such as cancer, the method comprising using a method termed herein as "Genome Association Analysis of Transcription and Comparative Oncogenomics" or "GAATCO." Further described herein are computer systems and computer readable media comprising instructions to diagnose and/or predict a prognosis for a subject having cancer.

[013] The GAATCO approach for biomarker discovery uses an integrative analysis framework to detect core gene expression signatures associated with copy-number alterations. In particular, the method for biomarker discovery as disclosed herein comprises a three step process; the first step involves identifying: (a) chromosomal abnormalities and the copy number of genes affected by such chromosomal abnormalities (e. g. copy number lesions or amplifications) in an animal disease model as compared to an animal model without such copy-number abnormalities, and (b) the expression level of genes in a biological sample obtained from the animal disease model (i.e. a global gene expression profile is obtained). The second step involves gene copy number-gene expression correlation analysis where the copy number of each gene is correlated with the average expression level for each gene (i.e. reference level) using bioinformatics analysis and distinguishing gene loci within the copy number affected region and showing correlated expression from genes outside the copy number affected region with the genes which are expressed by the loci of genes with copy number changes, resulting in a "cancer biomarker" gene set. The third step involves filtering this "cancer biomarker" gene set by a cross-species comparison, where genes with

no known human orthologs are discarded (i.e. genes which are specific to the animal of the animal model) from the gene set to obtain a “core cancer biomarker” gene set.

[014] In some embodiments, the core prognostic disease biomarker gene set can be validated by comparison with lists of differentially regulated genes obtained from human samples. As an illustrative example only, a core prognostic cancer biomarker gene list can be validated by comparing the core prognostic cancer biomarker gene list with gene expression data sets of human cancer samples (i.e. gene profile data sets) and which, for example, have been grouped to correlate with the survival of the humans with cancer. Stated another way, a gene identified as one of the core prognostic cancer biomarkers can be validated by identifying if it is differentially expressed in a human cancer and has been shown to be an indicator of survival of a human subject with such a cancer. For example, a gene or gene set identified as one or more of the core cancer biomarkers identified in an animal model of cancer is validated if it can be identified in a human cancer and the core cancer biomarker(s) correlate with a cancer stage or cancer type.

[015] Another aspect of the present invention relates to compositions and methods for the diagnosis and prognosis of cancer and/or provides a diagnostic test that is sensitive and specific.

[016] The inventors have discovered a method that significantly increases the diagnostic accuracy of identifying the type of cancer in a subject with cancer, for example a lymphoma, leukemia, glioma, breast cancer or lung cancer and other diseases with gene copy number variations (Freeman et al., 2006; Redon et al., 2006), using gene expression analysis of a group of biomarkers to identify or differentiate the cancer sub-type (e.g., Burkitt lymphoma versus large B-cell lymphoma).

[017] Accordingly the inventors have discovered a method for improved diagnosis and/or prognosis of cancer, such as for example lymphoma, glioma, breast cancer and lung cancer, among others in a subject by assessing the expression level of a group of cancer biomarkers or a subgroup thereof. In some embodiments, the methods as disclosed herein can be used to predict the progression of a tumor, for example, one can use the prognostic cancer biomarkers as disclosed herein to monitor the expression of the cancer biomarkers in a subject at a first timepoint as compared to a second timepoint, where a change in the level of expression of the cancer biomarkers in the subject from the first timepoint as compared to the second timepoint can be used to assess the progression of the tumor.

[018] One aspect described herein relates to a method for determining a prognosis for an individual having cancer, the method comprising: (a) measuring the level of expression of at

least 5 nucleic acid sequences encoding proteins selected from Group I in a biological sample from the individual, wherein Group I consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); DUT (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); FOXM1 (SEQ ID NO: 57); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LSM6 (SEQ ID NO: 72); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106);

RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SETD6 (SEQ ID NO: 120); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150) and ZFYVE26 (SEQ ID NO: 151); and (b) determining a metagene score for the patient based on the expression of the at least 5 nucleic acid sequences; (c) comparing the metagene score determined in step (b) with a set of reference metagene scores based on the expression of the at least 5 nucleic acid sequences in a population of individuals having cancer; wherein a metagene score from the patient greater than the median of the set of reference metagene scores indicates a poor prognosis; and wherein a metagene score from the patient less than the median of the set of reference metagene scores indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

[019] In one embodiment of this aspect and all other aspects described herein, the metagene score is designated on a scale of -1 to +1, with the median metagene score set at zero.

[020] Another aspect described herein is a method for diagnosing cancer in a subject, the method comprising: (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group I in a biological sample from the subject, wherein Group I consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO:

15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); DUT (includes EG:1854) (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); FOXM1 (SEQ ID NO: 57); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LSM6 (SEQ ID NO: 72); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SETD6 (SEQ ID NO: 120); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124);

STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150) and ZFYVE26 (SEQ ID NO: 151); and (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; (c) comparing the normalized gene expression level to a reference expression database for each set of nucleic acid sequences measured; wherein the correlation to the reference database identifies that the subject is likely to have cancer of a specific subtype or stage, and/or identifies a prognosis for the disease. In one embodiment, steps (b) and (c) comprise: normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; and comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis or outcome, wherein the comparing identifies the reference expression level in the database closest in value to the normalized expression level for the patient tumor, whereby the identification of the closest reference expression value indicates the tumor in the reference database that has the closest reference expression score is most similar to the patient tumor and therefore most similar in sub-type or stage to the patient tumor.

[024] In one embodiment of this aspect and all other aspects described herein, the level of expression is determined by measuring the protein level of the gene product.

[025] In another embodiment of this aspect and all other aspects described herein, the level of expression is determined by measuring the mRNA level of the gene product.

[026] Another aspect described herein is a method for determining a prognosis for an individual having cancer, the method comprising: (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group II in a biological sample from the individual, wherein Group II consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6

(SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24) CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CEP70 (SEQ ID NO: 28) CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31) CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MCM7 (SEQ ID NO: 79) MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO:

130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151); and (b) determining a metagene score for the patient based on the expression of the at least 5 nucleic acid sequences; (c) comparing the metagene score determined in step (b) with a set of reference metagene scores based on the expression of the at least 5 nucleic acid sequences in a population of individuals having cancer; wherein a metagene score from the patient greater than the median of the set of reference metagene scores indicates a poor prognosis; and wherein a metagene score from the patient less than the median of the set of reference metagene scores indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

[027] In one embodiment of this aspect and all other aspects described herein, the metagene score is designated on a scale of -1 to +1, with the median metagene score set at zero.

[028] Also described herein is a method for diagnosing cancer in a subject, the method comprising; (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group II in a biological sample from the subject, wherein Group II consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24) CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CEP70 (SEQ ID NO: 28); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31) CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2

(SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151); and (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; (c) comparing the normalized gene expression level to a reference database for each nucleic acid sequence measured; wherein the correlation to the reference database identifies that the subject is likely to have cancer of a specific subtype or stage and/or identifies a prognosis for the disease. In one embodiment, steps (b) and (c) comprise: normalizing the

level of expression of all nucleic acid sequences measured in (a) using a reference expression level; and comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis or outcome, wherein the comparing identifies the reference expression level in the database closest in value to the normalized expression level for the patient tumor, whereby the identification of the closest reference expression value indicates the tumor in the reference database that has the closest reference expression score is most similar to the patient tumor and therefore most similar in sub-type or stage to the patient tumor.

[029] Another aspect described herein relates to a method for determining a prognosis for an individual having cancer, the method comprising: (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group III in a biological sample from the individual, wherein Group III consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); CCNA2 (SEQ ID NO: 19); CCR1 (SEQ ID NO: 22); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 340); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LRDD (SEQ ID NO: 71); MAF (SEQ ID NO: 74); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MPHOSPH9 (SEQ ID NO: 80); MUM1 (SEQ ID NO: 82); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118);

SETD6 (SEQ ID NO: 120); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TOPBP1 (SEQ ID NO: 135); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151); and (b) determining a metagene score for the patient based on the expression of the at least 5 nucleic acid sequences; (c) comparing the metagene score determined in step (b) with a set of reference metagene scores based on the expression of the at least 5 nucleic acid sequences in a population of individuals having cancer; wherein a metagene score from the patient greater than the median of the set of reference metagene scores indicates a poor prognosis; and wherein a metagene score from the patient less than the median of the set of reference metagene scores indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

[030] In one embodiment of this aspect and all other aspects described herein, the metagene score is designated on a scale of -1 to +1, with the median metagene score set at zero.

[031] Another aspect described herein relates to a method for diagnosing cancer in a subject, the method comprising: (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group III in a biological sample from the subject, wherein Group III consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); CCNA2 (SEQ ID NO: 19); CCR1 (SEQ ID NO: 22); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 340); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LRDD (SEQ ID NO:

71); MAF (SEQ ID NO: 74); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MPHOSPH9 (SEQ ID NO: 80); MUM1 (SEQ ID NO: 82); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SETD6 (SEQ ID NO: 120); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TOPBP1 (SEQ ID NO: 135); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151); and (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; (c) comparing the normalized gene expression level to a reference database for each nucleic acid sequence measured; wherein the correlation to the reference database identifies that the subject is likely to have cancer of a specific subtype or stage and/or identifies a prognosis for the disease. In one embodiment, steps (b) and (c) comprise: normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; and comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis or outcome, wherein the comparing identifies the reference expression level in the database closest in value to the normalized expression level for the patient tumor, whereby the identification of the closest reference expression value indicates the tumor in the reference database that has the closest reference expression score is most similar to the patient tumor and therefore most similar in sub-type or stage to the patient tumor.

[032] Another aspect described herein relates to a method for determining a prognosis for an individual having cancer, the method comprising: (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group IV in a biological sample from the individual, wherein Group IV consists of gene sequences: C12ORF29 (SEQ

ID NO: 12); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCR1 (SEQ ID NO: 22); CDC25A (SEQ ID NO: 26); CHAF1B (SEQ ID NO: 29); CHEK1 (SEQ ID NO: 31); CSF1R (SEQ ID NO: 35); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DUT (includes EG:1854) (SEQ ID NO: 45); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HRAS (SEQ ID NO: 65); IDH3A (SEQ ID NO: 67); ITGB2 (SEQ ID NO: 69); LSM6 (SEQ ID NO: 72); MAF (SEQ ID NO: 74); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); RAB31 (SEQ ID NO: 103); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RRM1 (SEQ ID NO: 111); SAE2, also called UBA2 (SEQ ID NO: 113); SFXN3 (SEQ ID NO: 121); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WDHD1 (SEQ ID NO: 148); and WEE1 (SEQ ID NO: 150); and (b) determining a metagene score for the patient based on the expression of the at least 5 nucleic acid sequences; (c) comparing the metagene score determined in step (b) with a set of reference metagene scores based on the expression of the at least 5 nucleic acid sequences in a population of individuals having cancer; wherein a metagene score from the patient greater than the median of the set of reference metagene scores indicates a poor prognosis; and wherein a metagene score from the patient less than the median of the set of reference metagene scores indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

[033] In one embodiment of this aspect and all other aspects described herein, the metagene score is designated on a scale of -1 to +1, with the median metagene score set at zero.

[034] Another aspect described herein relates to a method for diagnosing cancer in a subject, the method comprising: (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group IV in a biological sample from the subject, wherein Group IV consists of gene sequences: C12ORF29 (SEQ ID NO: 12); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCR1 (SEQ ID NO: 22); CDC25A (SEQ ID NO: 26); CHAF1B (SEQ ID NO: 29); CHEK1 (SEQ ID NO: 31); CSF1R (SEQ ID NO: 35); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DUT (includes EG:1854)

(SEQ ID NO: 45); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HRAS (SEQ ID NO: 65); IDH3A (SEQ ID NO: 67); ITGB2 (SEQ ID NO: 69); LSM6 (SEQ ID NO: 72); MAF (SEQ ID NO: 74); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); RAB31 (SEQ ID NO: 103); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RRM1 (SEQ ID NO: 111); SAE2, also called UBA2 (SEQ ID NO: 113); SFXN3 (SEQ ID NO: 121); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WDHD1 (SEQ ID NO: 148); and WEE1 (SEQ ID NO: 150); and (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; (c) comparing the normalized gene expression level to a reference database for each nucleic acid sequences measured; wherein the correlation to the reference database identifies that the subject is likely to have cancer of a specific subtype or stage and/or identifying a prognosis of the disease. In one embodiment, steps (b) and (c) comprise: normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; and comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis or outcome, wherein the comparing identifies the reference expression level in the database closest in value to the normalized expression level for the patient tumor, whereby the identification of the closest reference expression value indicates the tumor in the reference database that has the closest reference expression score is most similar to the patient tumor and therefore most similar in sub-type or stage to the patient tumor.

[035] Another aspect described herein comprises a method for determining a prognosis for an individual having cancer, the method comprising: (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group V in a biological sample from the individual, wherein Group V consists of gene sequences: RECQL4 (SEQ ID NO: 108) and at least one gene selected from the following: BRIP1 (SEQ ID NO: 10); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); MCM7 (SEQ ID NO: 79); MRE11A (SEQ ID NO: 81); NAP1L4 (SEQ ID NO: 83); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95);

RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135); or UNG (SEQ ID NO: 143); and (b) determining a metagene score for the patient based on the expression of the at least 5 nucleic acid sequences; (c) comparing the metagene score determined in step (b) with a set of reference metagene scores based on the expression of the at least 5 nucleic acid sequences in a population of individuals having cancer; wherein a metagene score from the patient greater than the median of the set of reference metagene scores indicates a poor prognosis; and wherein a metagene score from the patient less than the median of the set of reference metagene scores indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

[036] In one embodiment of this aspect and all other aspects described herein, the metagene score is designated on a scale of -1 to +1, with the median metagene score set at zero.

[037] Another aspect described herein relates to a method for diagnosing cancer in a subject, the method comprising: (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group V in a biological sample from the subject, wherein Group V consists of gene sequences: RECQL4 (SEQ ID NO: 108) and at least one gene selected from the following: BRIP1 (SEQ ID NO: 10); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); MCM7 (SEQ ID NO: 79); MRE11A (SEQ ID NO: 81); NAP1L4 (SEQ ID NO: 83); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135); or UNG (SEQ ID NO: 143) in a biological sample from the subject; and (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; (c) comparing the normalized gene expression level to a reference database for each nucleic acid sequence measured; wherein the correlation to the reference database identifies that the subject is likely to have cancer of a specific subtype or stage and/or identifies a prognosis for the disease. In one embodiment, steps (b) and (c) comprise: normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; and comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis or outcome, wherein the comparing identifies the reference expression level in the database closest in value to the normalized expression level for the patient tumor, whereby the identification of the closest reference expression value indicates the tumor in the reference

database that has the closest reference expression score is most similar to the patient tumor and therefore most similar in sub-type or stage to the patient tumor.

[038] Another aspect described herein relates to a method for determining a prognosis for an individual having cancer, the method comprising: (a) measuring the level of expression of at least 3 nucleic acid sequences encoding proteins selected from Group VI in a biological sample from the individual, wherein Group VI consists of gene sequences: C12orf29 (SEQ ID NO: 12); C14orf104 (SEQ ID NO:13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); or C5ORF15 (SEQ ID NO: 18); and (b) determining a metagene score for the patient based on the expression of the at least 5 nucleic acid sequences; (c) comparing the metagene score determined in step (b) with a set of reference metagene scores based on the expression of the at least 5 nucleic acid sequences in a population of individuals having cancer; wherein a metagene score from the patient greater than the median of the set of reference metagene scores indicates a poor prognosis; and wherein a metagene score from the patient less than the median of the set of reference metagene scores indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

[039] In one embodiment of this aspect and all other aspects described herein, the metagene score is designated on a scale of -1 to +1, with the median metagene score set at zero.

[040] Another aspect described herein relates to a method for diagnosing cancer in a subject, the method comprising: (a) measuring the level of expression of at least 3 nucleic acid sequences encoding proteins selected from Group VI in a biological sample from the subject, wherein Group VI consists of gene sequences: C12orf29 (SEQ ID NO: 12); C14orf104 (SEQ ID NO:13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); or C5ORF15 (SEQ ID NO: 18); and (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; (c) comparing the normalized gene expression level to a reference database for each nucleic acid sequence measured; wherein the correlation to the reference database identifies that the subject is likely to have cancer of a specific subtype or stage and/or identifies a prognosis for the disease. In one embodiment, steps (b) and (c) comprise: normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; and comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis

or outcome, wherein the comparing identifies the reference expression level in the database closest in value to the normalized expression level for the patient tumor, whereby the identification of the closest reference expression value indicates the tumor in the reference database that has the closest reference expression score is most similar to the patient tumor and therefore most similar in sub-type or stage to the patient tumor.

[041] Another aspect described herein comprises a method for determining a prognosis for an individual having cancer, the method comprising: (a) measuring the level of expression of at least 3 nucleic acid sequences encoding proteins selected from Group VII in a biological sample from the individual, wherein Group VII consists of gene sequences: MCM7 (SEQ ID NO: 79); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); RFC2 (SEQ ID NO: 109; or TOPBP1 (SEQ ID NO: 135); and (b) determining a metagene score for the patient based on the expression of the at least 5 nucleic acid sequences; (c) comparing the metagene score determined in step (b) with a set of reference metagene scores based on the expression of the at least 5 nucleic acid sequences in a population of individuals having cancer; wherein a metagene score from the patient greater than the median of the set of reference metagene scores indicates a poor prognosis; and wherein a metagene score from the patient less than the median of the set of reference metagene scores indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

[042] In one embodiment of this aspect and all other aspects described herein, the metagene score is designated on a scale of -1 to +1, with the median metagene score set at zero.

[043] Another aspect described herein relates to a method for diagnosing cancer in a subject, the method comprising: (a) measuring the level of expression of at least 3 nucleic acid sequences encoding proteins selected from Group VII in a biological sample from the subject, wherein Group VII consists of gene sequences: MCM7 (SEQ ID NO: 79); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); RFC2 (SEQ ID NO: 109; or TOPBP1 (SEQ ID NO: 135); and (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; (c) comparing the normalized gene expression level to a reference database for each nucleic acid sequence measured; wherein the correlation to the reference database identifies that the subject is likely to have cancer of a specific subtype or stage and/or identifies a prognosis for the disease. In one embodiment, steps (b) and (c) comprise: normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; and comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of

reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis or outcome, wherein the comparing identifies the reference expression level in the database closest in value to the normalized expression level for the patient tumor, whereby the identification of the closest reference expression value indicates the tumor in the reference database that has the closest reference expression score is most similar to the patient tumor and therefore most similar in sub-type or stage to the patient tumor.

[044] Another aspect described herein is a method for determining a prognosis for an individual having cancer, the method comprising: (a) measuring the level of expression of a nucleic acid sequence encoding a protein selected from FOXM1 (SEQ ID NO: 57) or RECQL4 (SEQ ID NO: 108) in a biological sample from the subject; and (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; (c) comparing the normalized gene expression level to a reference database for each nucleic acid sequence measured; wherein the correlation to the reference database identifies that the subject is likely to have cancer of a specific subtype or stage and/or identifies a prognosis for the disease. In one embodiment, steps (b) and (c) comprise: normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level; and comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis or outcome, wherein the comparing identifies the reference expression level in the database closest in value to the normalized expression level for the patient tumor, whereby the identification of the closest reference expression value indicates the tumor in the reference database that has the closest reference expression score is most similar to the patient tumor and therefore most similar in sub-type or stage to the patient tumor.

[045] In one embodiment of this aspect and all other aspects described herein, the level of expression is the level of gene transcript expression.

[046] In another embodiment of this aspect and all other aspects described herein, the level of expression is the level of protein expression.

[047] In another embodiment of this aspect and all other aspects described herein, at least 10 nucleic acid sequences are measured.

[048] In another embodiment of this aspect and all other aspects described herein, at least 20 nucleic acid sequences are measured.

[049] In another embodiment of this aspect and all other aspects described herein, at least 30 nucleic acid sequences are measured.

[050] In another embodiment of this aspect and all other aspects described herein, at least 40 nucleic acid sequences are measured.

[051] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of positively correlated genes of Group VIII, wherein Group VIII consists of: ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ATAD5 (SEQ ID NO: 7); BRIP1 (SEQ ID NO: 10); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLUAP1 (SEQ ID NO: 34); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DUT (includes EG:1854) (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); FAM128B (SEQ ID NO: 53); FBXO21 (SEQ ID NO: 55); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); LRDD (SEQ ID NO: 71); LSM6 (SEQ ID NO: 72); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NSMCE4A (SEQ ID NO: 89); ORC5L (SEQ ID NO: 91); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SETD6 (SEQ ID NO: 120); SMARCD2 (SEQ ID NO: 123); TAF4 (SEQ ID NO: 128); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WARS2 (SEQ ID NO: 146); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149) or WEE1 (SEQ ID NO: 150).

[052] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of negatively correlated genes of Group IX, wherein Group IX consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BSDC1 (SEQ ID NO: 11); C5ORF15 (SEQ ID NO: 18); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CLEC10A (SEQ ID NO: 33); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DSE (SEQ ID NO: 44); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EVI2A (SEQ ID NO: 52); FAM129A (SEQ ID NO: 54); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); NCF2 (SEQ ID NO: 85); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); OPTN (SEQ ID NO: 90); PBXIP1 (SEQ ID NO: 92); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); ZFYVE26 (SEQ ID NO: 151).

[053] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of positively regulated genes of Group X, wherein Group X consists of: ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ATAD5 (SEQ ID NO: 7); BRIP1 (SEQ ID NO: 10); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); CCNA2 (SEQ ID NO: 19); CDC25A (SEQ ID NO: 26); CEP70 (SEQ ID NO: 28); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLUAP1 (SEQ ID NO: 34); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); E2F1

(SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); FAM128B (SEQ ID NO: 53); FBXO21 (SEQ ID NO: 55); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HDDC2 (SEQ ID NO: 64); ID3 (SEQ ID NO: 66); LRDD (SEQ ID NO: 71); MAZ (SEQ ID NO: 77); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NSMCE4A (SEQ ID NO: 89); ORC5L (SEQ ID NO: 91); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PPP1R8 (SEQ ID NO: 97); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2 (also called UBA2) (SEQ ID NO: 113); SMARCD2 (SEQ ID NO: 123); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TOPBP1 (SEQ ID NO: 135); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WARS2 (SEQ ID NO: 146); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149) and WEE1 (SEQ ID NO: 150).

[054] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of negatively correlated genes of Group XI, wherein Group XI consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BSDC1 (SEQ ID NO: 11); C5ORF15 (SEQ ID NO: 18); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CLEC10A (SEQ ID NO: 33); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DSE (SEQ ID NO: 44); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EVI2A (SEQ ID NO: 52); FAM129A (SEQ ID NO: 54); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); NCF2 (SEQ ID NO: 85); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); OPTN (SEQ ID NO: 90); PBXIP1 (SEQ ID NO: 92); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID

NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TRPV2 (SEQ ID NO: 140); UGCG (SEQ ID NO: 142); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); ZFYVE26 (SEQ ID NO: 151).

[055] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of positively correlated genes of Group XII, wherein Group XII consists of: ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ATAD5 (SEQ ID NO: 7); BRIP1 (SEQ ID NO: 10); CCNA2 (SEQ ID NO: 19); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CISD1 (SEQ ID NO: 32); CLUAP1 (SEQ ID NO: 34); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EIF5A2 (SEQ ID NO: 50); FAM128B (SEQ ID NO: 53); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); ID3 (SEQ ID NO: 66); LRDD (SEQ ID NO: 71); MAZ (SEQ ID NO: 77); MPHOSPH9 (SEQ ID NO: 80); MUM1 (SEQ ID NO: 82); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NSMCE4A (SEQ ID NO: 89); ORC5L (SEQ ID NO: 91); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); SETD6 (SEQ ID NO: 120); TAF4 (SEQ ID NO: 128); TOPBP1 (SEQ ID NO: 135); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149) and WEE1 (SEQ ID NO: 150).

[056] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of negatively correlated genes of Group XIII, wherein Group XIII consists of: consisting of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ARHGEF3 (SEQ ID NO: 6); ATP6V0A1 (SEQ ID NO: 9); BSDC1 (SEQ ID NO: 11); CCR1 (SEQ ID NO: 22); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CLEC10A (SEQ ID NO: 33); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DSE (SEQ ID NO: 44); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EVI2A (SEQ ID NO: 52); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); HBP1 (SEQ ID NO: 62); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); MAF (SEQ ID NO: 74); MAP3K8 (SEQ ID NO: 76); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); OPTN (SEQ ID NO: 90); PBXIP1 (SEQ ID NO: 92); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107);

SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); ZFYVE26 (SEQ ID NO: 151).

[057] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of positively correlated genes of Group XIV, wherein Group XIV consists of: C12ORF29 (SEQ ID NO: 12); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CDC25A (SEQ ID NO: 26); CHAF1B (SEQ ID NO: 29); CHEK1 (SEQ ID NO: 31); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DUT (includes EG:1854) (SEQ ID NO: 45); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HRAS (SEQ ID NO: 65); IDH3A (SEQ ID NO: 67); LSM6 (SEQ ID NO: 72); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); ORC5L (SEQ ID NO: 91); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RRM1 (SEQ ID NO: 111); SAE2, also called UBA2 (SEQ ID NO: 113); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WDHD1 (SEQ ID NO: 148); WEE1 (SEQ ID NO: 150).

[058] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of negatively correlated genes of Group XV, wherein Group XV consists of: CCR1 (SEQ ID NO: 22); CSF1R (SEQ ID NO: 35); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); ITGB2 (SEQ ID NO: 69); MAN2A1 (SEQ ID NO: 75); OPTN (SEQ ID NO: 90); PBXIP1 (SEQ ID NO: 92); RAB31 (SEQ ID NO: 103); SFXN3 (SEQ ID NO: 121); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127) and TNIP1 (SEQ ID NO: 134).

[059] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of replication genes of subgroup A, wherein subgroup A consists of: CCNA2; CDC25A; CHEK1; DNA2; ORC5L; POLD1; POLE; RFC2; WEE1; CCNB2 or RECQL4.

[060] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of DNA repair genes of subgroup B, wherein subgroup B consists of: CHEK1; MRE11A; UNG; WEE1 or RECQL4.

[061] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of nuclear factor genes of subgroup C, wherein subgroup C consists of: STAT4; TNIP1; HBP1; PBXIP1; CCNA2; CDC25A; CHEK1; DNMT1; MAZ; MRE11A; NEK2; ORC5L; POLD1; POLE; PPP1R8; RFC2; RRM1; UNG; WEE1; CHAF1B; RECQL4; TOPBP1; LSM6; WDHD1; NCAPH; GPSM2 or MBD3.

[062] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from the group of cytoplasmic factor genes of subgroup D, wherein subgroup D consists of: HCK; MAN2A1; SQSTM1; OPTN; RAB31; SFXN3; HADH; IDH3A; PFKM; USP1; CCNB2; SAE2; MPHOSPH9 or TRAIP.

[063] In another embodiment of this aspect and all other aspects described herein, the at least 5 nucleic acid sequences encoding the proteins are selected from at least one nucleic acid sequence in each subgroup of genes, wherein the subgroups consist of: Subgroup A, B, C, or D.

[064] In another embodiment of this aspect and all other aspects described herein, the cancer is leukemia.

[065] In another embodiment of this aspect and all other aspects described herein, the cancer is lymphoma.

[066] In another embodiment of this aspect and all other aspects described herein, the cancer is a neoplasm.

[067] In another embodiment of this aspect and all other aspects described herein, the cancer is selected from the group consisting of: a breast cancer, a colon cancer, a glioma, a lung cancer or is a hyperproliferative disorder.

[068] In another embodiment of this aspect and all other aspects described herein, the cancer is selected from the group consisting of: a carcinoma, a sarcoma, a metastatic cancer, a lymphatic metastases, a blood cell malignancy, a multiple myeloma, an acute leukemia, a chronic leukemia, a head and neck cancer, a mouth cancer, a larynx cancer, a thyroid cancer, a respiratory tract cancer, a small cell carcinoma, a non-small cell cancer, a ductal carcinoma, a gastrointestinal cancer, an esophageal cancer, a stomach cancer, a colorectal cancer, a

pancreatic cancer, a liver cancer, a genitourinary cancer, an urologic cancer, a bladder cancer, a prostate cancer, an ovarian carcinoma, an uterine cancer, an endometrial cancer, a kidney cancer, a renal cell carcinoma, a brain cancer, a neuroblastoma, an astrocytic brain tumors, a metastatic tumor cell invasion in the central nervous system, a bone cancer, osteomas, a skin cancer, a malignant melanoma, a squamous cell carcinoma, a basal cell carcinoma, a hemangiopericytoma or a Kaposi's sarcoma or combinations thereof.

[069] In another embodiment of this aspect and all other aspects described herein, the reference expression level is the level of the nucleic acid or protein measured from biological samples obtained from a plurality of subjects.

[070] In another embodiment of this aspect and all other aspects described herein, the reference expression level is the level of the nucleic acid or protein of at least one gene from biological samples obtained from at least 50 subjects.

[071] In another embodiment of this aspect and all other aspects described herein, the reference expression level is the level of one or more normalizing genes (e.g., housekeeping genes such as HPRT).

[072] In another embodiment of this aspect and all other aspects described herein, the reference expression level for each nucleic acid or protein measured is the level of the nucleic acid from a biological sample obtained from the subject at a first time point.

[073] In another embodiment of this aspect and all other aspects described herein, the biological sample is from the subject at a second time point, and the reference expression level is the level of the nucleic acid or protein from a biological sample obtained from the subject at a first time point.

[074] In another embodiment of this aspect and all other aspects described herein, where the method indicates that the cancer is an aggressive cancer, there is the further step wherein the individual is directed to be treated with an appropriate anti-cancer therapy for an aggressive cancer.

[075] In another embodiment of this aspect and all other aspects described herein, where the method indicates that the cancer is a non-aggressive cancer, the method includes the further step wherein the individual is directed to be treated with an appropriate anti-cancer therapy for a non-aggressive cancer.

[076] Another aspect described herein relates to a method to identify a prognostic disease biomarker, the method comprising: (a) identifying a copy-number variation in a genome of a subject, and measuring the copy number of the genes affected by the copy number variation in a biological sample from a subject with a gene copy-number variation; (b) comparing the

copy number of the genes measured in (a) with a reference copy number for each gene measured, and identifying genes where the copy number in the subject with the copy-number variation is different as compared with the reference sample; (c) measuring gene expression levels in a biological sample from the subject with a gene copy number variation; (d) performing a copy number-gene expression correlation analysis for each gene identified in (b); and (e) performing a copy number-gene expression correlation analysis for each gene identified in (c); wherein genes with an expression level dependent on copy number variation are identified as a prognostic disease biomarker in a subject.

[077] Another aspect described herein relates to a method to identify a prognostic disease biomarker, the method comprising: (a) identifying a chromosomal abnormality in the genome of a subject, and measuring the copy number of the genes affected by the chromosomal abnormality in a biological sample obtained from the subject with a chromosomal abnormality; (b) comparing the copy number of the genes measured in (a) with a reference copy number for each gene measured, and identifying genes where the copy number in the subject with the chromosomal abnormality is different as compared to the reference sample; (c) measuring the gene expression levels in a biological sample from the subject; (d) performing a copy number-gene expression correlation analysis for each gene identified in (c); and wherein a gene with whose expression level is changed proportionally to the change in copy number identifies the gene as a prognostic disease biomarker.

[078] In one embodiment of this aspect and all other aspects described herein, the copy number-gene expression correlation analysis is determined using the algorithm:

$$T_j = \beta_0 + \epsilon_j \quad (1)$$

$$T_j = \beta_0 + \beta_1 D_j + \epsilon_j \quad (2)$$

$$T_j = \beta_0 + \beta_2 C_j + \epsilon_j \quad (3)$$

$$T_j = \beta_0 + \beta_1 D_j + \beta_2 C_j + \epsilon_j \quad (4)$$

wherein T_j is expression, D_j is the copy-number, ϵ_j is the random error of the gene in sample j ; C_j is the copy-number of the CNA in sample j ; β_1 is the regression parameters measuring local association; β_2 is the regression parameters measuring global association; β_0 is an intercept regression parameter.

[079] In another embodiment of this aspect and all other aspects described herein, the reference copy number is the copy number from a subject without the chromosomal abnormality.

[080] In another embodiment of this aspect and all other aspects described herein, the difference in copy number is an increase in copy number.

[081] In another embodiment of this aspect and all other aspects described herein, the difference in copy number is a decrease in copy number.

[082] In another embodiment of this aspect and all other aspects described herein, the difference in copy number is at least a 0.5-fold change. Alternatively, in other embodiments, the difference in copy number is at least a 1.5-fold change; at least a 2.0-fold change; at least a 3.0-fold change; at least a 4.0-fold change; or at least a 5.0-fold change or greater.

[083] In another embodiment of this aspect and all other aspects described herein, the method further comprises performing a cross-species gene ontology comparison of the genes identified as a prognostic disease biomarker, wherein if at least a 1:1 ratio of the prognostic disease biomarker to human or subject to be treated (non-human animal) ortholog is present, then the prognostic disease biomarker is identified as a core prognostic disease biomarker.

[084] In another embodiment of this aspect and all other aspects described herein, the copy number-gene expression correlation analysis is determined by step 2 of the Genome Association Analysis of Transcription and Comparative Oncogenomics method as described herein.

[085] In another embodiment of this aspect and all other aspects described herein, the gene copy number variation is a chromosomal abnormality.

[086] In another embodiment of this aspect and all other aspects described herein, the gene copy number variation is present in a disease.

[087] In another embodiment of this aspect and all other aspects described herein, the gene copy number variation is present in a cancer.

[088] In another embodiment of this aspect and all other aspects described herein, the gene copy number variation is present in a nervous system disorder, metabolic disorder or in a cardiovascular disorder.

[089] In another embodiment of this aspect and all other aspects described herein, the disease is cancer.

[090] In another embodiment of this aspect and all other aspects described herein, the biological sample is cancer tissue.

[091] In another embodiment of this aspect and all other aspects described herein, the cancer tissue is a biopsy cancer tissue.

[092] In another embodiment of this aspect and all other aspects described herein, the subject is an animal model of a disease.

[093] In another embodiment of this aspect and all other aspects described herein, the subject is an in vitro model system, such as organ culture, tissue slice or cell culture.

[094] In another embodiment of this aspect and all other aspects described herein, the animal model comprises a mouse model.

[095] In another embodiment of this aspect and all other aspects described herein, the mouse model comprises a cancer model.

[096] In another embodiment of this aspect and all other aspects described herein, the method further comprises the step of comparing the gene identified as a prognostic disease biomarker, or a core prognostic disease biomarker, with a gene expression profile of genes known to be correlated with the survival of a subject with the disease, wherein if the prognostic biomarker is a gene which has also been identified to be correlated with the survival of a subject with the disease, it validates the prognostic biomarker.

[097] In another embodiment of this aspect and all other aspects described herein, the subject is a human.

[098] In another embodiment of this aspect and all other aspects described herein, the subject is selected from the group consisting of: a farm animal, companion animal and domestic animal.

[099] In another embodiment of this aspect and all other aspects described herein, the method further comprises performing a gene ontology analysis of the genes identified as a prognostic disease biomarker, and using such to identify novel disease pathways.

[0100] In another embodiment of this aspect and all other aspects described herein, the method identifies therapeutic targets for disease.

[0101] In another embodiment of this aspect and all other aspects described herein, the method identifies a therapeutic pathway or strategy.

[0102] Another aspect described herein relates to a kit comprising primers which specifically hybridize to the nucleic acid sequences of at least 5 genes selected from the group of genes encoding proteins of Group I, wherein Group I consists of proteins; ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ

ID NO: 24); CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); DUT (includes EG:1854) (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); FOXM1 (SEQ ID NO: 57); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LSM6 (SEQ ID NO: 72); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SETD6 (SEQ ID NO: 120); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132);

TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150) and ZFYVE26 (SEQ ID NO: 151).

[0103] Another aspect described herein relates to a kit comprising primers which specifically hybridize to the nucleic acid sequences of at least 5 genes selected from the group of genes encoding proteins of Group II, wherein Group II consists of proteins; ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24) CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CEP70 (SEQ ID NO: 28); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31) CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAPIL4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID

NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151).

[0104] Another aspect described herein relates to a kit comprising primers which specifically hybridize to the nucleic acid sequences of at least 5 genes selected from the group of genes encoding proteins of Group III, wherein Group III consists of proteins; ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); CCNA2 (SEQ ID NO: 19); CCR1 (SEQ ID NO: 22); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 340); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LRDD (SEQ ID

NO: 71); MAF (SEQ ID NO: 74); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MPHOSPH9 (SEQ ID NO: 80); MUM1 (SEQ ID NO: 82); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SETD6 (SEQ ID NO: 120); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TOPBP1 (SEQ ID NO: 135); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151).

[0105] Another aspect described herein relates to a kit comprising primers which specifically hybridize to the nucleic acid sequences of at least 5 genes selected from the group of genes encoding proteins of Group IV, wherein Group IV consists of proteins; C12ORF29 (SEQ ID NO: 12); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCR1 (SEQ ID NO: 22); CDC25A (SEQ ID NO: 26); CHAF1B (SEQ ID NO: 29); CHEK1 (SEQ ID NO: 31); CSF1R (SEQ ID NO: 35); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DUT (includes EG: 1854) (SEQ ID NO: 45); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HRAS (SEQ ID NO: 65); IDH3A (SEQ ID NO: 67); ITGB2 (SEQ ID NO: 69); LSM6 (SEQ ID NO: 72); MAF (SEQ ID NO: 74); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); RAB31 (SEQ ID NO: 103); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RRM1 (SEQ ID NO: 111); SAE2, also called UBA2 (SEQ ID NO: 113); SFXN3 (SEQ ID NO: 121); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TNIP1 (SEQ ID NO: 134); TOPBP1

(SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WDHD1 (SEQ ID NO: 148); and WEE1 (SEQ ID NO: 150).

[0106] Another aspect described herein relates to a kit comprising primers which specifically hybridize to the nucleic acid sequences of at least 5 genes selected from the group of genes encoding proteins of Group V, wherein Group V consists of proteins; RECQL4 (SEQ ID NO: 108) and at least one gene selected from the following: BRIP1 (SEQ ID NO: 10); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); MCM7 (SEQ ID NO: 79); MRE11A (SEQ ID NO: 81); NAP1L4 (SEQ ID NO: 83); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135); or UNG (SEQ ID NO: 143).

[0107] Another aspect described herein relates to a kit comprising primers which specifically hybridize to the nucleic acid sequences of at least 3 genes selected from the group of genes encoding proteins of Group VI, wherein Group VI consists of proteins; C12orf29 (SEQ ID NO: 12); C14orf104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); or C5ORF15 (SEQ ID NO: 18).

[0108] Another aspect described herein relates to a kit comprising primers which specifically hybridize to the nucleic acid sequence of at least 3 genes selected from the group of genes encoding proteins of Group VII, wherein Group VII consists of proteins; MCM7 (SEQ ID NO: 79); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135).

[0109] In one embodiment of this aspect and all other aspects described herein, the kit comprises primers which specifically hybridize to the nucleic acid sequences of at least 10 genes. Alternatively, in some embodiments the kit comprises primers which specifically hybridize to the nucleic acid sequences of at least 20 genes; at least 30 genes; at least 40 genes, or at least 50 genes.

[0110] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 1 but no more than 100 additional sets of primers.

[0111] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 1 but no more than 1000 additional sets of primers.

[0112] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 1 but no more than 5000 additional sets of primers.

[0113] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 10 additional sets of primers.

[0114] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 100 additional sets of primers.

[0115] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 1000 additional sets of primers.

[0116] In another embodiment of this aspect and all other aspects described herein, the kit comprises at least one primer which specifically hybridizes to the nucleic acid encoding the protein RECQL4 (SEQ ID NO: 108).

[0117] In another embodiment of this aspect and all other aspects described herein, the kit comprises at least one primer which specifically hybridizes to the nucleic acid encoding the protein FOXM1 (SEQ ID NO: 57).

[0118] Another aspect described herein relates to a kit comprising protein-binding probes which specifically bind to the protein of at least 5 proteins selected from Group I, wherein Group I consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); DUT (includes EG:1854) (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); FOXM1 (SEQ ID NO: 57); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71);

LSM6 (SEQ ID NO: 72); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SETD6 (SEQ ID NO: 120); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150) and ZFYVE26 (SEQ ID NO: 151).

[0119] Another aspect described herein relates to a kit comprising protein-binding probes which specifically bind to the protein of at least 5 proteins selected from Group II, wherein Group II consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18);

CCNA2 (SEQ ID NO: 19); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24) CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CEP70 (SEQ ID NO: 28) CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31) CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MCM7 (SEQ ID NO: 79) MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); UGCG (SEQ ID NO: 142); UNG

(SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151).

[0120] Another aspect described herein relates to a kit comprising protein-binding probes which specifically bind to the protein of at least 5 proteins selected from Group III, wherein Group III consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); CCNA2 (SEQ ID NO: 19); CCR1 (SEQ ID NO: 22); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 340); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LRDD (SEQ ID NO: 71); MAF (SEQ ID NO: 74); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MPHOSPH9 (SEQ ID NO: 80); MUM1 (SEQ ID NO: 82); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SETD6 (SEQ ID NO: 120); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TOPBP1 (SEQ ID NO: 135); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151).

[0121] Another aspect described herein relates to a kit comprising protein-binding probes which specifically bind to the protein of at least 5 proteins selected from Group III, wherein Group IV consists of: C12ORF29 (SEQ ID NO: 12); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCR1 (SEQ ID NO: 22); CDC25A (SEQ ID NO: 26); CHAF1B (SEQ ID NO: 29); CHEK1 (SEQ ID NO: 31); CSF1R (SEQ ID NO: 35); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DUT (includes EG:1854) (SEQ ID NO: 45); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HRAS (SEQ ID NO: 65); IDH3A (SEQ ID NO: 67); ITGB2 (SEQ ID NO: 69); LSM6 (SEQ ID NO: 72); MAF (SEQ ID NO: 74); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); RAB31 (SEQ ID NO: 103); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RRM1 (SEQ ID NO: 111); SAE2, also called UBA2 (SEQ ID NO: 113); SFXN3 (SEQ ID NO: 121); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WDHD1 (SEQ ID NO: 148); and WEE1 (SEQ ID NO: 150).

[0122] Another aspect described herein relates to a kit comprising protein-binding probes which specifically bind to the protein of at least 5 proteins selected from Group V, wherein Group V consists of: RECQL4 (SEQ ID NO: 108) and at least one gene selected from the following: BRIP1 (SEQ ID NO: 10); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); MCM7 (SEQ ID NO: 79); MRE11A (SEQ ID NO: 81); NAP1L4 (SEQ ID NO: 83); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135); or UNG (SEQ ID NO: 143).

[0123] Another aspect described herein relates to a kit comprising protein-binding probes which specifically bind to the protein of at least 3 proteins selected from Group VI, wherein Group VI consists of: C12orf29 (SEQ ID NO: 12); C14orf104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); or C5ORF15 (SEQ ID NO: 18).

[0124] Another aspect described herein relates to a kit comprising protein-binding probes which specifically bind to the protein of at least 3 proteins selected from Group VII, wherein

Group VII consists of: MCM7 (SEQ ID NO: 79); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135).

[0125] In one embodiment of this aspect and all other aspects described herein, the kit comprises protein-binding probes which specifically bind to at least 10 proteins.

[0126] In another embodiment of this aspect and all other aspects described herein, the kit comprises protein-binding probes which specifically bind to at least 20 proteins.

[0127] In another embodiment of this aspect and all other aspects described herein, the kit comprises protein-binding probes which specifically bind to at least 30 proteins.

[0128] In another embodiment of this aspect and all other aspects described herein, the kit comprises protein-binding probes which specifically bind to at least 40 proteins.

[0129] In another embodiment of this aspect and all other aspects described herein, the kit comprises protein-binding probes which specifically bind to at least 50 proteins.

[0130] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 1 but no more than 100 additional sets of protein-binding probes.

[0131] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 1 but no more than 1000 additional sets of protein-binding probes.

[0132] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 1 but no more than 5000 additional sets of protein-binding probes.

[0133] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 10 additional sets of protein-binding probes.

[0134] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 100 additional sets of protein-binding probes.

[0135] In another embodiment of this aspect and all other aspects described herein, the kit further comprises at least 1000 additional sets of protein-binding probes.

[0136] In another embodiment of this aspect and all other aspects described herein, the kit comprises at least one protein binding probe which specifically binds to the protein RECQL4 (SEQ ID NO: 108).

[0137] In another embodiment of this aspect and all other aspects described herein, the kit comprises at least one protein binding probe which specifically binds to the protein FOXM1 (SEQ ID NO: 57).

[0138] In another embodiment of this aspect and all other aspects described herein, the kit is an ELISA.

[0139] In another embodiment of this aspect and all other aspects described herein, the kit is a quantitative PCR kit.

[0140] In one aspect, described herein is a method for predicting prognosis for a cancer patient, the method comprising: (a) measuring gene expression for a prognostic gene set and a normalizing gene or gene set in a patient or subject with a tumor; (b) normalizing expression of the prognostic gene set relative to the normalizing gene or gene set; (c) determining a metagene score for the patient or subject using, for example, singular value decomposition, non-negative matrix factorization (NMF) or principal component analysis (PCA); and (d) comparing the metagene score with a reference metagene score normalized in the same manner as step (b), wherein a patient metagene score larger than the reference metagene score indicates a poor prognosis; and wherein a patient metagene score less than the reference metagene score indicates a good prognosis; and wherein the degree to which the score is greater than or less than the reference metagene score correlates with the relative aggressiveness of the tumor.

[0141] In one embodiment of this aspect and all other aspects described herein, the metagene score is designated on a scale of -1 to +1, with the median score set at zero. In this aspect, the degree to which the patient metagene score is greater than zero correlates with the relative aggressiveness of the tumor.

[0142] Also described herein is a computer system for predicting the prognosis of a cancer patient, the system comprising: (a) a computer processor, and (b) a computer readable physical storage medium having instructions recorded thereon sufficient to implement a process, employing the computer processor, for predicting the prognosis of a cancer patient, the instructions comprising:

- (i) instructions for accepting a plurality of input values characterizing gene expression in a tumor from said patient, wherein said values include values characterizing a normalizing gene and a prognostic gene set;
- (ii) instructions for normalizing values characterizing gene expression of genes in said prognostic gene set relative to said normalizing gene,
- (iii) instructions for determining, using for example singular value decomposition, non-negative matrix factorization (NMF) or principal component analysis (PCA), a metagene score for the patient;
- (iv) instructions for comparing said metagene score of said patient with a reference metagene score normalized in the same manner as step (ii) and (iii), and
- (v) instructions for transmitting to a user interface the results of said comparison,

wherein a patient metagene score larger than said reference metagene score indicates a poor prognosis; and wherein a patient metagene score less than said reference metagene score indicates a good prognosis; and wherein the degree to which the score is greater than or less than said reference metagene score correlates with the relative aggressiveness of the tumor.

[0143] Another aspect described herein relates to a computer readable physical storage medium having instructions recorded thereon sufficient to implement a process, employing a computer processor, for predicting the prognosis of a cancer patient, the instructions comprising:

- (i) instructions for accepting a plurality of input values characterizing gene expression in a tumor from said patient, wherein said values include values characterizing a normalizing gene and a prognostic gene set;
- (ii) instructions for normalizing values characterizing gene expression of genes in said prognostic gene set relative to said normalizing gene,
- (iii) instructions for determining, using, for example, singular value decomposition, non-negative matrix factorization (NMF) or principal component analysis (PCA) a metagene score for the patient;
- (iv) instructions for comparing said metagene score of said patient with a reference metagene score normalized in the same manner as in step (ii) and (iii), and
- (v) instructions for transmitting to a user interface the results of said comparison,

wherein a patient metagene score larger than said reference metagene score indicates a poor prognosis; and wherein a patient metagene score less than said reference metagene score indicates a good prognosis; and wherein the degree to which the score is greater than or less than said reference metagene score correlates with the relative aggressiveness of the tumor.

[0144] Also described herein is a computer system for predicting the prognosis of a cancer patient, the system comprising: (a) a computer processor, and (b) a computer readable physical storage medium having instructions recorded thereon sufficient to implement a process, employing the computer, for predicting the prognosis of a cancer patient, the instructions comprising:

- (i) instructions for accepting a plurality of input values characterizing gene expression in a tumor from said patient, wherein said values include values characterizing a normalizing gene and a prognostic gene set;

- (ii) instructions for normalizing values characterizing gene expression of genes in said prognostic gene set relative to said normalizing gene,
- (iii) instructions for determining, using, for example singular value decomposition, non-negative matrix factorization (NMF) or principal component analysis (PCA), a metagene score for the patient tumor;
- (iv) instructions for comparing said metagene score of said patient with a plurality of reference metagene scores in a database, determined for a plurality of reference tumors of known prognosis or outcome,

wherein said comparing identifies the reference metagene score in said database closest in value to the metagene score for said patient tumor, whereby the identification of the closest reference metagene score indicates the tumor in said reference database that has the closest reference metagene score is most similar to said patient tumor and therefore most similar in prognosis to said patient tumor; and

- (v) instructions for transmitting to a user interface the results of said comparison of step (iv), whereby a prognosis for said patient tumor is predicted.

[0145] Further described herein is a computer readable physical storage medium having instructions recorded thereon sufficient to implement a process, employing a computer processor, for predicting the prognosis of a cancer patient, the instructions comprising:

- (i) instructions for accepting a plurality of input values characterizing gene expression in a tumor from said patient, wherein said values include values characterizing a normalizing gene and a prognostic gene set;
- (ii) instructions for normalizing values characterizing gene expression of genes in said prognostic gene set relative to said normalizing gene,
- (iii) instructions for determining, using for example singular value decomposition, non-negative matrix factorization (NMF) or principal component analysis (PCA), a metagene score for the patient tumor;
- (iv) instructions for comparing said metagene score of said patient with a plurality of reference metagene scores in a database, determined for a plurality of reference tumors of known prognosis or outcome, wherein said comparing identifies the reference metagene score in said database closest in value to the metagene score for said patient tumor, whereby the identification of the closest reference metagene score indicates the tumor in said reference database that has the closest reference metagene score is most similar to said patient tumor and therefore most similar in prognosis to said patient tumor; and

(v) instructions for transmitting to a user interface the results of said comparison of step (iv), whereby a prognosis for said patient tumor is predicted.

BRIEF DESCRIPTION OF FIGURES

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawings(s) will be provided by the Office upon request and payment of the necessary fee.

Figures 1A-1B show a schematic diagram of human chromosomes 4 and 11. Figure 1 left panel shows the regions of conserved synteny of human chromosome 4 with mouse chromosome 5, and Figure 1 right panel shows the regions of conserved synteny of human chromosome 11 with mouse chromosome 9.

Figures 2A-2D show examples of the comparative analysis filters for a core conserved signature. Cross-species analysis identifies a core conserved gene signature. Figure 2A shows heat maps of gene-by-gene correlation matrices for the mouse and two human B lymphoma datasets. Positive (light gray) and negative (dark gray) correlations are indicated. Data are truncated at values of -0.5 and 0.5 . Figure 2B Heat maps of correlation agreement between the three datasets. Gray cells denote gene pairs with significant agreement ($p < 0.05$) in correlation between mouse and both human data sets. Genes are listed in the same order as A (top panel) or as ordered by k-means clustering (bottom panel). The core conserved set of 136 genes is shown in the overlaid box (dashed arrow). Figure 2C shows correlation heat maps for the core 136 genes. The core set can be subdivided into positively or negatively correlated fractions. Figure 2D shows schematic overview of the filtering steps to obtain the core, conserved 136-gene set.

Figures 3 shows functional enrichment by comparative filtering. Figure 3 shows a bar chart showing a fraction of prognostic genes for all genes ($n = 9816$), the unfiltered orthologous signature ($n = 321$), and the core conserved set ($n = 136$). Prognostic genes are defined as genes whose expression levels are correlated to tumor-free survival outcomes in more than three (> 3) cohorts among the collection of 7 human cohorts.

Figures 4A-4H show the association between the core prognostic cancer biomarkers and tumor-free survival among patients with B lymphoma, breast cancer and glioma. The 136-gene signature shows prognostic value in multiple tumor types. For each cohort, patients were bifurcated into two equal-sized groups based on core signature metagene scores. Kaplan-Meier survival estimates for the two groups are shown for (A) the mouse B lymphoma set; (B-C) two human B lymphoma (Dave et al., 2006; Hummel et al., 2006); (D-F) three human breast cancer (Desmedt et al., 2007; Miller et al., 2005; Wang et al., 2005); and (G-H) two human glioma datasets (Freije et al., 2004; Phillips et al., 2006). The human cohorts are obtained for example from Gene Expression Omnibus (available on the world wide web at ncbi.nlm.nih.gov/geo/).

Figures 5A-5C show an array CGH analysis of a *Lig4 Trp53* pro-B cell tumor. Figure 5 A shows a schematic diagram of an array CGH method, showing competitive hybridization of differentially labeled test and reference DNA. Figure 5B shows BAC-based array CGH analysis (3Mb resolution) of chromosomes 12 and 15, showing amplification encompassing *Igh* and *c-myc* respectively (vertical axis=test/reference signal ratio). Figure 5C shows BAC-based array CGH analysis of chromosomes 5 and 9, showing a putative amplicon on chr.5 and a putative deletion on chr.9.

Figure 6 shows the Compugen Array CGH comparison of normal male versus female genomic DNA. The top panel shows whole genome CGH plot. The lower panel shows an expanded view of chromosomes 15, 16, 17, 18, 19, X, and Y. Blue plot points represent positive $\log_2$ ratios and red plot points represent negative $\log_2$ ratios.

Figures 7A-7C show integrative oncogenomics to identify conserved transcriptional signatures associated with genetic aberrations. Figure 7A shows a schematic overview of Genome Association Analysis of Transcription and Comparative Oncogenomics (GAATCO). Figure 7B shows a schematic outline of the comparative analysis. Nodes (circles) indicate genes and edges (lines) significant coexpression. Comparative analysis involves identifying common co-expression links between mouse and human, as disclosed in the Examples and demonstrated in Figure 2. Figure 7C shows experiments used in this study. The top panel of Figure 7C shows the experimental design of primary murine B lymphoma study, and the bottom panel of Figure 7C shows a list of experiments used in this study and their sources, the references of which are incorporated herein in their entirety by reference.

Figure 8 shows histological validation using the proliferation marker Ki67. Ki67 index provides histological validation that the core signature is a faithful molecular marker of tumor cell proliferation. Average expression levels of genes, which positively correlate with aggressiveness of tumor in the core set, are compared to histology marker of proliferation, Ki67. The p-value for assessing the significance of correlation is displayed.

Figure 9 shows a schematic of an exemplary method as described herein. B6 (C57BL/6J) is a normal control DNA sample for the copy number, T1 through Tn are the DNA samples from tumors to be analyzed for copy number, for example from progenitor B lymphoma tumors of Art p53 double null mice. The DNA copy number can be established using e.g. array MEEBO CGH. From the same tumor samples RNA is isolated to examine the gene expression levels using for example Affymetrix GeneChips. The gene expression data is analyzed with e.g. ANOVA tools to determine the gene expression level of each gene in comparison to the average of all genes on the chip (normalization using a global pattern recognition algorithm, see Akilesh et al. 2003). Further the gene expression level is correlated to the gene copy number. The global gene expression signature is filtered using cross-species comparison to obtain the conserved gene signature.

Figure 10 shows an overview and validation of GAATCO analysis with a schematic representation of the GAATCO algorithm or approach, indicating multi-step filtering and cross-species analysis.

Figures 11A-11D show genomic analyses of mouse lymphomas with genome-wide transcriptional alterations associated with deletions on chromosome 6. Figure 11A shows DNA copy number (top panel) and gene expression (bottom panel) data for a collection of 18 AP lymphomas. AP refers to mice doubly deficient in the DNA double-strand break repair gene *Dclre1c*/Art and the tumor suppressor gene *Trp53*. Individual genes (horizontal axis) are displayed by chromosome position, for each tumor (vertical axis). DNA copy number gains (blue) and losses (red) are shown relative to a normal, diploid C57BL6/J mouse genome. Gene expression data is displayed relative to the mean (scaled to a value of 0) for each gene. Higher (green) and lower (red) than mean values are indicated for each gene. The vertical lines between the top and bottom panel denote genes with significant correlation between copy-number and expression levels (Holm $p < 0.05$). Figure 11B shows a bar chart showing p-values (as $-\log_{10}$ values) for association between CNAs and tumor-free survival time. The dotted lines correspond to p-value threshold of 0.01, 0.001, and 0.0001. Figure 11C shows chromosome ideograms

displaying genes with expression levels significantly correlated to deletions at chr.6C1-G3. Chromosome 6 deleted segment of the mouse genome shares conserved synteny with region of the human chromosomes 2, 3, and 12, implicated in B-lineage neoplasms. Locations of deletions found in human acute lymphoblastic leukemia (red squares) Burkitt's lymphoma (green squares) or diffuse large B-cell lymphoma (blue squares) are indicated.

Figure 12 shows an example for cancer typing. Left shows the Dave et al. cohort (Dave et al. 2006), and right the Hummel et al. cohort (Hummel et al. 2006). BL or mBL means Burkitt's lymphoma; DLBCL means Diffuse Large Cell B Lymphoma; non-mBL non-Burkitt's lymphoma; intermediate means that these tumors were classified as intermediate. To determine the tumor subtype the gene expression patterns of 136 genes were grouped based on unsupervised hierarchical clustering which effectively distinguished BL from DLBCL, and performed better than conventional molecular diagnostic methods.

DETAILED DESCRIPTION

[0146] As disclosed herein, the inventors have discovered a group of cancer biomarkers and subgroups for the diagnosis and prognosis of cancer in a subject. The inventors have also discovered a method for identifying disease biomarkers, such as cancer biomarkers for the diagnosis and prognosis of cancer in a subject, using a method termed herein as "Genome Association Analysis of Transcription for Comparative Oncogenomics" or "GAATCO." In particular embodiments, the methods as disclosed herein for biomarker discovery can be applied to any disease that involves gene copy number variations. Accordingly, the present invention relates to a method to direct treatment in a subject by measuring the expression of the cancer biomarkers or a subgroup thereof as disclosed herein in a subject and if the subject has a change in the level of expression of a cancer biomarker as compared to a reference sample, a clinician can administer to the subject an effective amount of an appropriate anti-cancer agent.

[0147] In another embodiment this method can be applied to identify disease pathways and consequently novel disease targets and treatment or therapy strategies.

[0148] In another embodiment this method can be used as a research tool to unravel biological mechanisms related to gene copy number variations.

Definitions

- [0149]** For convenience, certain terms employed in the entire application (including the specification, examples, and appended claims) are collected here. Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs.
- [0150]** The term "gene" as used herein can be a genomic gene sequence comprising transcriptional and/or translational regulatory sequences and/or a coding region and/or non-translated sequences (e.g., introns, 5'- and 3'- untranslated sequences and regulatory sequences). The coding region of a gene can be a nucleotide sequence coding for an amino acid sequence or a functional RNA, such as tRNA, rRNA, catalytic RNA, siRNA, miRNA and antisense RNA. A gene can also be an mRNA or cDNA corresponding to the coding regions (e.g. exons and miRNA) optionally comprising 5'- or 3' untranslated sequences linked thereto. A gene can also be an amplified nucleic acid molecule produced in vitro comprising all or a part of the coding region and/or 5'- or 3'- untranslated sequences linked thereto.
- [0151]** The terms "nucleic acid" or "oligonucleotide" or "polynucleotide" as used herein refer to nucleotides covalently linked together. A nucleic acid will generally have at least two nucleotides covalently linked in the manner of linkage in DNA and RNA in nature. An "oligonucleotide" nucleic acid is generally smaller than a "polynucleotide." Nucleic acids can be single stranded or double stranded, or can contain portions of both double stranded and single stranded sequence. As will be appreciated by those in the art, the depiction of a single strand also defines the sequence of the complementary strand. Thus, a nucleic acid as described herein also encompasses and fully describes the complementary strand of a depicted single strand. As will also be appreciated by those in the art, many variants of a nucleic acid can be used for the same purpose as a given nucleic acid. Thus, a nucleic acid also encompasses substantially identical nucleic acids and complements thereof. As will also be appreciated by those in the art, a single strand provides a probe that can hybridize to a target sequence under stringent hybridization conditions. Thus, a nucleic acid also encompasses a probe that hybridizes to a target nucleic acid under stringent hybridization conditions. The nucleic acid can be DNA, both genomic and cDNA, RNA, or a hybrid.
- [0152]** The phrases "hybridizing specifically to" or "specific hybridization" or "selectively hybridize to", refer to the annealing, binding, duplexing, or hybridizing of a nucleic acid molecule preferentially to a particular nucleotide sequence under stringent conditions when that sequence is present in a complex mixture (e.g., total cellular) DNA or RNA.

[0153] The term "stringent conditions" refers to conditions under which a probe will hybridize preferentially to its target subsequence, and to a lesser extent to, or not at all to, other sequences. "Stringent hybridization" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes* part I chapter 2 Overview of principles of hybridization and the strategy of nucleic acid probe assays, Elsevier, N.Y. Generally, highly stringent hybridization and wash conditions are selected to be about 5 °C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe.

[0154] The phrase "substantially identical," in the context of two nucleic acid sequences or polypeptide sequences, refers to two or more sequences or subsequences that have at least 80%, or at least 90, or 95% or higher nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using a comparison algorithm or by visual inspection. Preferably, the substantial identity exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably the sequences are substantially identical over at least about 150 residues. In a most preferred embodiment, the sequences are substantially identical over the entire length of the coding regions.

[0155] The terms "homology", "identity" and "similarity" refer to the degree of sequence similarity between two peptides or between two optimally aligned nucleic acid molecules. Homology and identity can each be determined by comparing a position in each sequence which can be aligned for purposes of comparison. For example, sequence comparison is based upon using a standard homology software in the default position, such as BLAST, version 2.2.14, which is suitable for determining percent sequence identity and sequence similarity as described in Altschul et al (1990) *J. Mol. Biol.* 215:403 410. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information and is well within the abilities of one skilled in the art to use for sequence comparisons.

[0156] When an equivalent position in the compared sequences is occupied by the same base or amino acid, then the molecules are identical at that position; when the equivalent site is occupied by similar amino acid residues (e.g., similar in steric and/or electronic nature, for example conservative amino acid substitutions), then the molecules can be referred to as homologous (similar) at that position. Expression as a percentage of homology/similarity or identity refers to a function of the number of similar or identical amino acids at positions shared by the compared sequences, respectfully. A sequence which is "unrelated" or "non-homologous" shares less than 40% identity, though preferably less than 25% identity with the sequences as disclosed herein. The term "substantially homologous" refers to sequences that are at least 90%, at least 95% identical, at least 96%, identical at least 97% identical, at least 98% identical or at least 99% identical. Homologous sequences can be the same functional gene in different species. Determination of homologs of the genes or peptides of the present invention can be easily ascertained by the skilled artisan.

[0157] As used herein, the term "sequence identity" means that two polynucleotide sequences or two amino acid sequences are identical (i.e., on a nucleotide-by-nucleotide or residue-by-residue basis) over the comparison window. Optimal alignment of sequences for comparison can be conducted, e.g., by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman (1988) *Proc. Natl. Acad. Sci. USA* 85:2444, by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, Wis.), or by visual inspection (see generally Ausubel et al., *supra*).

[0158] The term "specifically binds" when used to refer to binding-protein probes as used herein indicates that the binding preference (e.g., affinity for the target molecule/protein is at least 2 fold, more preferably at least 5 fold, and most preferably at least 10 or 20 fold over a non-specific (e.g. randomly generated molecule lacking the specifically recognized amino acid or amino acid sequence) target molecule or protein. Stated another way, the term "specifically bind" as used herein refers to the ability of a protein-binding probe to bind to a target protein with a greater affinity than non target proteins. For example, about 10%, about 20%, about 30%, about 40%, preferably about 50%, more preferably about 60%, more preferably about 70%, still more preferably about 80%, still more preferably about 90%, still more preferably about 100% or greater affinity for the target protein relative to non-target proteins. At a minimum, the term "specifically binds" refers to binding with a K_d of 10

micromolar or less, preferably 1 micromolar or less, more preferably 100 nM or less, 10 nM or less, or 1 nM or less. One example of a protein-binding probe is an antibody.

[0159] The term "biological sample" refers to a sample of biological tissue, cells, or fluid that, in a healthy and/or pathological state, comprises a nucleic acid or polypeptide that is to be detected according to the assays described herein. Such samples include, but are not limited to, cultured cells, primary cell preparations, sputum, amniotic fluid, blood, tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues (e.g., frozen or fixed sections taken for histological purposes). In some embodiments a biological sample is taken from a human patient, and in alternative embodiments the biological sample is taken from any mammal, such as rodents, animal models of diseases, commercial animals, companion animals, dogs, cats, sheep, cattle, and pigs, etc.

[0160] The term "biomarker" as used herein refers to measurable characteristics that indicate the phenotype of a particular cell or individual (i.e., a disease phenotype), for example expression levels of particular genes, mRNA or polypeptides. Examples of biomarkers include characteristics, whether morphological, functional or biochemical (enzymatic) characteristics that are particular to a cell type, or molecules expressed by the cell type. A biomarker may consist of any molecule found in a cell including, but not limited to, nucleic acids (nucleotides, DNA, RNA, mRNA, miRNA, snRNA), proteins (peptides and polypeptides), lipids, polysaccharides, and steroids. Markers may be detected by any method available to one of skill in the art.

[0161] As used herein, the term "prognostic disease biomarker" refers to a gene which is differentially expressed in a cancer or disease cell compared to a normal cell and thus acts as a biomarker for a disease in a subject. A prognostic disease biomarker can be used both to diagnose and determine prognosis of a disease in a subject. For example, a prognostic cancer biomarker is a gene or the protein expressed from it, the expression or level of which can identify a subject as likely to develop cancer, or that has cancer. A prognostic biomarker may be increased in expression compared to its expression in a normal (i.e., reference) cell, or alternatively a prognostic biomarker may be decreased in expression compared to that of a normal (i.e., reference) cell. The degree of increase (or decrease) in expression of a biomarker can correlate with the severity of disease (i.e., prognosis). Prognostic biomarkers that are nucleic acids or polypeptides can be detected and/or measured using e.g., nucleic acid array, RT-PCR, ELISA or any other method known to one of skill in the art.

[0162] The term “chromosomal abnormalities” is used herein to refer to abnormalities in the number or organization of chromosomes, and also refers to chromosomal alterations known in the art as chromosomal malformations, anomalies, or defects, or abnormal karyotype. In some instances, chromosomal abnormalities can be detected before birth by means of amniocentesis. The abnormalities can be either of number of chromosomes, or of composition of the individual chromosomes. In some embodiments, the chromosomal abnormalities are monosomy and trisomy, which are examples of numerical abnormalities. Chromosomal abnormalities also include translocations, for example abnormalities of structure where parts of one chromosome have been transferred to another. Chromosomal abnormalities are often linked with structural or functional defects of the animal body.

[0163] The term “copy-number variation” (also referred to in the art as “DNA copy number variation” or “CNV”) is used herein to refer to chromosomal aberrations, such as amplifications and deletions. In some embodiments, copy number variation (CNV) is a gain or loss of large segments of DNA sequence consisting of between ten thousand and five million base pairs. Numerous regions in the human genome have been identified as sites of large-scale copy number variations, some of which have been linked to disease phenotypes. Particular copy number variations have also been identified as correlating with protection against diseases, e.g., susceptibility to HIV infection and developing malaria.

[0164] As used herein, the term “reference level” or “reference expression level” refers to the normalized gene expression level for a given gene or normalizing gene across a range of biological samples (e.g., obtained from one or a plurality of individuals, or from a tissue or tissue bank) of the same tissue and disease status. As an illustrative example, the reference level for gene X is calculated as the average level of gene expression for gene X in a number of biological samples, for example, lung biological samples obtained from at least one subject. Typically, the average gene expression level or reference level is normalized to “0” value. For the reference level the expression level of at least one gene, preferably a “normalizing gene” such as a house-keeping gene (e.g. GAPDH (glyceraldehyde 3-phosphate dehydrogenase), SDHA (succinate dehydrogenase), HPRT (hypoxanthine phosphoribosyl transferase), HBS1L (HBS1-like protein), AHSP (alpha hemoglobin stabilizing protein), B2M (beta-2-microglobulin), beta-actin, beta-tubulin, large ribosomal protein, 18S ribosomal RNA, transferrin receptor, TATA-box-binding protein, beta-glucuronidase, cyclophilin A, porphobilinogen deaminase, ATP synthase 6 ; also see Trends in Genetics 19, 362-365 (2003)), is measured. Alternatively, the expression of many or all genes or any number of genes can be measured and used for normalization. Genes or proteins useful for

normalization include, but are not limited to, CALCR (calcitonin receptor), CRHR2 (Corticotropin-releasing hormone receptor 2), CYLC1 (Cylicin I), EDA (ectodysplasin A), GNRHR (GnRH receptor), HNF4A (hepatocyte nuclear factor 4, alpha), PAX8 (paired box 8), Mbd2 (methyl-CpG binding domain protein 2), ADAM22 (A disintegrin and metalloproteinase domain 22), Nf2 (neurofibromin 2), NPY2R (neuropeptide Y receptor Y2), TRIM15 (tripartite motif-containing 15), TRIM10 (tripartite motif-containing 10), RHAG (CD241 antigen), PDE10A (phosphodiesterase 10A) or ZBTB7A (Factor binding IST protein 1). Alternatively, the approach of global pattern recognition (GPR) normalization as described by Akilesh et al. (2003) can be used instead of housekeeping or normalizing genes.

[0165] As used herein the term “reference database” or “reference expression database” refers to a collection or database containing expression data and clinical data. The expression data can be generated from cancer tissue or cells, tumor tissue or cells, tumor or cancer biopsies, etc. and can be gene expression and/or protein expression of all or any number of genes. The clinical data refers to tumor classification (e.g. types, subtypes of tumors or cancer, stages of tumors or cancer), clinical information, survival data, treatment data, morphological data, clinical outcome, etc.. Such data may also be uploaded from public sources (e.g. Gene Expression Omnibus(GEO) of the National Center for Biotechnology Information; ArrayExpress (European Bioinformatics Institute, Cambridge, UK; Parkinson et al. 2007); Molecular Diagnosis of Burkitt’s Lymphoma resource <http://llmpp.nih.gov/BL>; Gene Expression Data Portal (GEDP) <http://caarraydb.nci.nih.gov>; Integrated Tumor Transcriptome Array and Clinical data Analysis (ITTACA); Cancer Gene Expression Database (CGED); ACTuDB: a database for the integrated analysis of array-CGH and clinical data for tumors; Genomics Institute of the Novartis Foundation data (available on the world wide web at wombat.gnf.org/index.html; SAGE Tag libraries from the Cancer Genome Anatomy Project, available on the world wide web at cgap.nci.nih.gov/SAGE), proprietary databases or newly generated. The more data points included in the reference database, the more accurate the prediction will be. A variety of computational tools known to those of skill in the art can be applied for the prediction. Expression data from at least 50 subjects are included in the reference expression database.

[0166] The terms “metascore” or “metagene score” or “metagene value” or “metavalue” are used interchangeably herein and refer to a weighted average value from expression data for a sample or disease type. In one embodiment, the metascore is defined as a range between -1 and + 1. For this, the weighted average gene expression score is used to rank the scores, with the lowest value designated as -1 and the highest value designated as +1. The value "0" is

defined as the median score between the highest and the lowest values. A lower score (i.e. -1) indicates a better overall prognosis, whereas a higher score (i.e. +1) indicates a less favorable prognosis and more aggressive disease progression. As used herein, the metascore for the 51 genes listed in Table 4 and for the 136 genes listed in Table 2 were calculated using singular value decomposition (SVD), which is a method of factorization of a rectangular real or complex matrix. Methods for using a singular value decomposition algorithm are well known in the art or can be found in the following references: Bau, D and Trefethen, L (1997),

Numerical linear algebra, Philadelphia: Society for Industrial and Applied Mathematics, ISBN 978-0-89871-361-9; Demmel, J, and Kahan, W (1990), *Journal on Scientific and Statistical Computing* 11 (5): 873–912; Eckart, C., & Young, G. (1936). *Psychometrika* I, 211-218; Golub, GH and Kahan, W (1965), *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* 2 (2): 205–224; Golub, GH. and Van Loan, CF (1996), *Matrix Computations (3rd ed.)*, Johns Hopkins; GSL Team (2007), "*§13.4 Singular Value Decomposition*", GNU Scientific Library. Reference Manual ; Halldor, B and Venegas, SA. (1997). "A manual for EOF and SVD analyses of climate data". McGill University, CCGCR Report No. 97-1, Montréal, Québec, 52pp; Hansen, P. C. (1987). The truncated SVD as a method for regularization. *BIT*, 27, 534-553; Horn, RA. and Johnson, CR (1985). "Matrix Analysis". Section 7.3. Cambridge University Press.; Horn, RA. and Johnson, CR (1991). *Topics in Matrix Analysis*, Chapter 3. Cambridge University Press. ISBN 0-521-46713-6; Strang G (1998). "Introduction to Linear Algebra". Section 6.7. 3rd ed., Wellesley-Cambridge Press. ISBN 0-9614088-5-5; Stewart, G. W. (1993), *SIAM Review* 35 (4): 551–566; Wall, ME., A Rechtsteiner, LM. Rocha (2003). A Practical Approach to Microarray Data Analysis. D.P. Berrar, W. Dubitzky, M. Granzow, eds. pp. 91-109, Kluwer: Norwell, MA. A metagene score can also be determined by other methods, such as non-negative matrix factorization (NMF), for example see following references: Tamayo, P, Scanfeld, D, Ebert, BL, Gillette, MA., Roberts, CWM, Mesirov, JP 2007, *Proc. Natl. Acad. Sci. USA* 104 (14), pp. 5959-5964; D. D. Lee and H. S. Seung (1999), *Nature* 401 (6755): 788–791; J. Zhang; J. Wang, *Neural Networks and Signal Processing*, 2008 International Conference on 7-11 June 2008, p163 – 167. Other methods for determining a metagene score include, but are not limited to, non-negative matrix factorization (NMF) or principal component analysis (PCA).

[0167] Furthermore, the range of the metascore can be defined by one of skill in the art using any other useful convention, for example from 0-10, or 0-100 instead of the convention (-1 to +1) described herein by the inventors.

[0168] The term “reduced” or “reduce” or “decrease” as used herein generally means a decrease by a statistically significant amount. However, for avoidance of doubt, “reduced” means a decrease by at least 10% as compared to a reference level, for example a decrease by at least about 20%, or at least about 30%, or at least about 40%, or at least about 50%, or at least about 60%, or at least about 70%, or at least about 80%, or at least about 90% or up to and including a 100% decrease (i.e. substantially absent or below levels of detection), or any decrease between 10-100% as compared to a reference level, as that term is defined herein.

[0169] The terms “increased” or “increase” as used herein generally means an increase by a statically significant amount; for the avoidance of any doubt, “increased” means an increase of at least 10% as compared to a reference level, for example an increase of at least about 20%, or at least about 30%, or at least about 40%, or at least about 50%, or at least about 60%, or at least about 70%, or at least about 80%, or at least about 90% or up to and including a 100% increase or any increase between 10-100% as compared to a reference level, or at least about a 2-fold, or at least about a 3-fold, or at least about a 4-fold, or at least about a 5-fold or at least about a 10-fold increase, or any increase between 2-fold and 10-fold or greater as compared to a reference level, as that term is defined herein.

[0170] The term “tissue” refers to a group or layer of similarly specialized cells, which together perform certain special functions. The term “tissue” is intended to include, blood, blood preparations such as plasma and serum, groups of cells, epithelial tissue, connective tissue, bones, joints, muscles, muscle tissue (e.g., smooth muscle, skeletal muscle, cardiac muscle), nervous tissue, and organs, among others.

[0171] The terms “subject”, “patient” and “individual” are used interchangeably herein, and refer to an animal, for example a human, to whom treatment, including prophylactic treatment is provided. The term “subject” as used herein refers to human and non-human animals. The term “non-human animals” and “non-human mammals” are used interchangeably herein and include all vertebrates, e.g., mammals, such as non-human primates, (particularly higher primates), horse, camel, sheep, dog, rodent (e.g. mouse or rat), guinea pig, goat, pig, cat, rabbits, and cattle. In one embodiment, the subject is human. In another embodiment, the subject is an experimental animal or animal substitute as a disease model. In some embodiments, the subject is a domestic animal, for example a companion animal (e.g., a pet), such as a dog or a feline pet, among others.

[0172] The terms “malignancy” and “cancer” are used interchangeably herein and refer to any disease of an organ or tissue in mammals characterized by poorly controlled or uncontrolled multiplication of cells in that tissue. A “malignancy” or “cancer” is generally also

characterized by the ability to invade into neighboring tissues or to metastasize.

Malignancies or cancer include disorders which result in the formation of a tumor, but also include non-tumor-forming hyperproliferative diseases such as the hematological neoplasms, including the leukemias. In particular embodiments, the term "cancer" can also encompass diseases involving benign neoplasms, dysplasias, hyperplasias as well as neoplasms showing metastatic growth or any other transformations, e.g. leukoplakias which often precede a breakout of cancer.

[0173] As used herein, the term "tumor" refers to an abnormal mass of cells or tissue, that are characterized, at least in part, by uncontrolled, progressive multiplication of cells. In some embodiments, such cells are characterized by neoplastic uncontrolled cell multiplication which is rapid and continues even after the stimuli that initiated the new growth has ceased. The term "tumor" is used broadly to include the tumor parenchymal cells as well as the supporting stroma, including the angiogenic blood vessels that infiltrate the tumor parenchymal cell mass. The term tumor is either a malignant tumor i.e., a cancer having the ability to metastasize (i.e. a metastatic tumor), or a benign or nonmalignant (i.e. non-metastatic) tumor.

[0174] The term "aggressive cancer" refers to a cancer which has at least one of the following properties: a rapid growth rate, a recurrent cancer (return after treatment), a cancer with potential to metastasize (spread) rapidly or a cancer which is resistant to at least one cancer therapy useful against cancers of that general type. Aggressive cancers usually require aggressive treatment while less aggressive cancers may be candidates for close clinical observation or other less aggressive cancer treatments. For certain types of cancer, the aggressiveness of the cancer can be graded on a clinically accepted scale. For example, an aggressive prostate cancer is a prostate cancer with a high Gleason score and is resistant to monotherapy. For example, for lymphoma, stages III and IV are considered as aggressive. A number of such clinically accepted grading scales are known to those of skill in the art.

[0175] As used herein, the term "metastasis", "metastases" or "metastatic tumor" refers to a secondary tumor that grows separately elsewhere in the body from the primary tumor (e.g., solid tumor) and has arisen from detached, transported cells. The primary tumor, as used herein, refers to a tumor that originated in the location or organ in which it is present and did not metastasize to that location from another location. As used herein, a "malignant tumor" is one having the properties of invasion and metastasis and showing a high degree of anaplasia. The term anaplasia was originally used to describe dedifferentiated cells. Anaplastic cells

display an immature or less differentiated phenotype and are pleomorphic. The presence of less differentiated or specialized cells, anaplastic cells is typical for most malignant tumors.

[0176] The term “copy number” or “gene copy number” as used herein refers to the number of molecules of a particular gene sequence in a DNA sequence or genome.

[0177] As used herein, the term “expression profile” as used herein refers to the amount or level of a plurality of different nucleic acids or proteins expressed in a cell or a population of cells or a tissue.

[0178] As used herein, the term “gene profile” as used herein refers to the amount or level of a plurality of different nucleic acids expressed in a cell or a population of cells or a tissue. As used herein, the terms “treat” or “treatment” or “treating” refer to both therapeutic treatment and prophylactic or preventative measures, wherein the objective is to prevent or slow the development of the disease, e.g., slow the development of a tumor, the spread of cancer, or reducing at least one effect or symptom of a condition, disease or disorder associated with inappropriate proliferation or a cell mass (e.g., cancer). Treatment is generally “effective” if one or more symptoms or clinical markers are reduced as that term is defined herein.

Alternatively, treatment is “effective” if the progression of a disease is reduced or halted.

That is, “treatment” includes not just the improvement of symptoms or markers, but also a cessation of the disorder, slowing of progress or preventing worsening of symptoms that would be expected in the absence of treatment. Beneficial or desired clinical results include, but are not limited to, alleviation of one or more symptom(s), diminishment of extent of disease, stabilization of the disease (i.e., not worsening), delay or slowing of disease progression, amelioration or palliation of the disease state, and remission (whether partial or total). “Treatment” can also mean prolonging survival as compared to expected survival if not receiving treatment. Those in need of treatment include those already diagnosed with cancer, as well as those likely to develop secondary tumors due to metastasis. A therapeutically or prophylactically significant reduction in a symptom is, e.g. at least about 10%, at least about 20%, at least about 30%, at least about 40%, at least about 50%, at least about 60%, at least about 70%, at least about 80%, at least about 90%, or even 100% (i.e., absence of symptom) in a measured parameter (e.g., symptom) as compared to a control or non-treated subject.

Measured or measurable parameters include clinically detectable markers of disease, for example, elevated or depressed levels of a biological marker, as well as parameters related to a clinically accepted scale of symptoms or markers for a disease or disorder. It is preferred that the treatment results in a decrease in the expression of a positive cancer biomarker by at least about 10%. A reduction in the rate of proliferation of the cancer cells by at least about

10% would also be considered effective treatment by the methods as disclosed herein. As alternative examples, a reduction in a symptom of cancer, for example, a slowing of the rate of growth of the cancer by at least about 10% or a cessation of the increase in tumor size, or a reduction in the size of a tumor by at least about 10% or a reduction in the tumor spread (i.e. tumor metastasis) by at least about 10% would also be considered as effective treatments by the methods as disclosed herein.

[0179] The term “effective amount” as used herein refers to the amount of a therapeutic agent or pharmaceutical composition that alleviates at least one or more symptom of the disease or disorder, and relates to a sufficient amount of pharmacological composition to provide the desired effect. The phrase “therapeutically effective amount” as used herein, e.g., of any composition as disclosed herein means a sufficient amount of the composition to treat a disorder, at a reasonable benefit/risk ratio applicable to any medical treatment. The term “therapeutically effective amount” therefore refers to an amount of the composition as disclosed herein that is sufficient to effect a therapeutically or prophylactically significant reduction in a symptom or clinical marker associated with cancer or any other disease associated with copy number variations. With reference to the treatment of a subject with cancer, the term “therapeutically effective amount” refers to the amount that is safe and sufficient to prevent or delay the development and further growth of a tumor or the spread of metastases in cancer patients. The amount can thus cure or cause the cancer to go into remission, slow the course of cancer progression, slow or inhibit tumor growth, slow or inhibit tumor metastasis, slow or inhibit the establishment of secondary tumors at metastatic sites, or inhibit the formation of new tumor metastases. The effective amount for the treatment of cancer depends on the tumor to be treated, the severity of the tumor, the drug resistance level of the tumor, the species being treated, the age and general condition of the subject, the mode of administration and so forth. Thus, it is not possible to specify the exact “effective amount”. However, for any given case, an appropriate “effective amount” can be determined by one of ordinary skill in the art using only routine experimentation. When using an experimental animal model, efficacy of treatment is evidenced when a reduction in a symptom of the cancer, for example a reduction in the size of the tumor or a slowing or cessation of the rate of growth of the tumor occurs earlier in treated, versus untreated animals. By “earlier” is meant that a decrease, for example in the size of the tumor occurs at least 5% earlier, but preferably more, e.g., one day earlier, two days earlier, 3 days earlier, or more.

[0180] The term “agent” refers to chemicals, small molecules, nucleic acid sequences, nucleic acid analogues, proteins, peptides, aptamers, antibodies, or functional fragments of such agents that can be used for treating cancer.

[0181] As used herein the term "user interface" is used to describe e.g., a display, hard-copy readout, or any other readout known to those of skill in the art.

[0182] The articles "a" and "an" are used herein to refer to one or to more than one (i.e., to at least one) of the grammatical object of the article. By way of example, "an element" means one element or more than one element.

[0183] It should be understood that this invention is not limited to the particular methodology, protocols, and reagents, etc., described herein and as such can vary. The terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention, which is defined solely by the claims.

Diagnosis and prognosis of cancer and other copy number variant associated diseases

[0184] The present invention relates to compositions and methods for the diagnosis and prognosis of cancer and copy number variant associated diseases, and provides for a diagnostic test that is sensitive and specific.

[0185] The inventors have discovered a method that significantly increases the diagnostic accuracy of identifying a subject having cancer or any other copy number variant associated disease, for example a lymphoma, leukemia, glioma, breast cancer, lung cancer, nervous system disease, autoimmune disease or metabolic disease using gene expression analysis of a group of biomarkers. Accordingly the inventors have discovered a method for enhanced diagnosis and/or prognosis of cancer (or any other copy number variant associated disease) in a subject by assessing the expression level of a group of biomarkers or a subgroup thereof, enabling improved typing and staging of cancer or any other copy number variant associated disease in a subject compared to other methods.

[0186] For example, for diagnosis and/or prognosis a cohort of tumor samples will be collected, and the gene expression level will be measured for a group of biomarkers and additional genes, for example housekeeping genes or “normalization” genes. A computational model for predicting a “good” or “poor” prognosis will be trained based on these collected expression data sets. Further, publicly available data sets can be included into this set. The collected data set will be used to predict diagnosis and/or prognosis of individual subjects.

[0187] Using mouse models of cancer, the inventors have discovered a prognostic signature of biomarker genes, such as cancer biomarker genes. The inventors have further identified a

subset of these prognostic cancer biomarkers as a core prognostic biomarker signature. The inventors have discovered a method to identify disease biomarkers using an analysis method, herein termed Genome Association Analysis of Transcription and Comparative Oncogenomics (GAATCO), to identify genes with differentially expressed in a cancer sample. GAATCO is a three-step strategy to detect core gene expression signatures associated with copy-number alterations. The inventors first (1) detect genomic regions of copy-number alterations (copy number lesions) where genes within the region exhibit correlated expression alterations; (2) for each copy-number lesion, the inventors detect gene(s) which are outside the lesion but show correlated expression levels to its copy-number pattern. (3) The inventors apply comparative analysis to filter the correlated signature(s) down to a functionally relevant set of genes. Possible statistical tools applied are disclosed in Example 8.

[0188] For example the inventors identified the copy number status of all genes of the diseased tissue sample (e. g. tumor sample) using a reference sample with normal gene copy number (e.g. nontumorous tissue sample). From the same tissue which was used for gene copy number analysis, gene expression level was also determined. Next, the genes expressed within the gene copy number changes were filtered from the set, i.e. discounted. The rest of the genes were analyzed regarding their correlation to the gene copy number using standard bioinformatics tools, e.g. ANOVA. The resulting gene set was then admitted to comparative analysis with another species, e.g. human sequences (orthologous methods). This results in a valuable gene signature set for diagnosis and prognosis of the disease originally addressed. For validation the inventors used the obtained gene signature to test it with the data set from e.g., human cancer gene sets. Using such a method on mouse models of cancer, such as mouse models of leukemia and lymphoma, the inventors have identified a group of cancer biomarkers (or subgroups thereof as diagnostic and/or prognostic cancer biomarkers which are set forth in Table 1.

[0189] For example, a tumor biopsy of a subject is obtained and applied for gene expression analysis using any group of the cancer biomarkers (e. g. Group I to XV or Group A-D). The gene expression is normalized and a metagene score is determined. The metagene score of the subject is compared to the reference database. Datasets with the same or similar metascores are used to predict the prognosis, e.g. chance of survival.

Table 1: Group I (151 gene set) of cancer biomarkers.

TABLE 1: GROUP I (151 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	1	ABCA3	21
SEQ ID NO:	2	ACADVL	37
SEQ ID NO:	3	ACP6	51205
SEQ ID NO:	4	ALG6	29929
SEQ ID NO:	5	APH1B	83464
SEQ ID NO:	6	ARHGEF3	50650
SEQ ID NO:	7	ATAD5	79915
SEQ ID NO:	8	ATF6	22926
SEQ ID NO:	9	ATP6V0A1	535
SEQ ID NO:	10	BRIP1	83990
SEQ ID NO:	11	BSDC1	55108
SEQ ID NO:	12	C12ORF29	91298
SEQ ID NO:	13	C14ORF104	55172
SEQ ID NO:	14	C14ORF124	56948
SEQ ID NO:	15	C16ORF80	29105
SEQ ID NO:	16	C18ORF24	220134
SEQ ID NO:	17	C21ORF45	54069
SEQ ID NO:	18	C5ORF15	56951
SEQ ID NO:	19	CCNA2	890
SEQ ID NO:	20	CCNB2	9133
SEQ ID NO:	21	CCPG1	9236
SEQ ID NO:	22	CCR1	1230
SEQ ID NO:	23	CD47	961
SEQ ID NO:	24	CD80	941
SEQ ID NO:	25	CD83	9308
SEQ ID NO:	26	CDC25A	993
SEQ ID NO:	27	CDT1	81620
SEQ ID NO:	28	CEP70	80321
SEQ ID NO:	29	CHAF1B	8208
SEQ ID NO:	30	CHCHD7	79145
SEQ ID NO:	31	CHEK1	1111
SEQ ID NO:	32	CISD1	55847
SEQ ID NO:	33	CLEC10A	10462
SEQ ID NO:	34	CLUAP1	23059
SEQ ID NO:	35	CSF1R	1436
SEQ ID NO:	36	CST3	1471
SEQ ID NO:	37	CTSO	1519
SEQ ID NO:	38	CX3CR1	1524
SEQ ID NO:	39	CXCR7	57007
SEQ ID NO:	40	DAB2	1601
SEQ ID NO:	41	DDX39	10212
SEQ ID NO:	42	DNA2	1763

TABLE 1: GROUP I (151 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	43	DNMT1	1786
SEQ ID NO:	44	DSE	29940
SEQ ID NO:	45	DUT (includes EG:1854)	1854
SEQ ID NO:	46	E2F1	1869
SEQ ID NO:	47	E2F8	79733
SEQ ID NO:	48	EBI2	1880
SEQ ID NO:	49	EHD4	30844
SEQ ID NO:	50	EIF5A2	56648
SEQ ID NO:	51	ETAA1	54465
SEQ ID NO:	52	EVI2A	2123
SEQ ID NO:	53	FAM128B	80097
SEQ ID NO:	54	FAM129A	116496
SEQ ID NO:	55	FBXO21	23014
SEQ ID NO:	56	FLII	2314
SEQ ID NO:	57	FOXMI	2305
SEQ ID NO:	58	GABARAP	11337
SEQ ID NO:	59	GALNT11	63917
SEQ ID NO:	60	GPSM2	29899
SEQ ID NO:	61	HADH	3033
SEQ ID NO:	62	HBP1	26959
SEQ ID NO:	63	HCK	3055
SEQ ID NO:	64	HDHC2	51020
SEQ ID NO:	65	HRAS	3265
SEQ ID NO:	66	ID3	3399
SEQ ID NO:	67	IDH3A	3419
SEQ ID NO:	68	IL2RA	3559
SEQ ID NO:	69	ITGB2	3689
SEQ ID NO:	70	LAMB3	3914
SEQ ID NO:	71	LRDD	55367
SEQ ID NO:	72	LSM6	11157
SEQ ID NO:	73	LY6E	4061
SEQ ID NO:	74	MAF	4094
SEQ ID NO:	75	MAN2A1	4124
SEQ ID NO:	76	MAP3K8	1326
SEQ ID NO:	77	MAZ	4150
SEQ ID NO:	78	MBD3 (includes EG:53615)	53615
SEQ ID NO:	79	MCM7	4176
SEQ ID NO:	80	MPHOSPH9	10198
SEQ ID NO:	81	MRE11A	4361
SEQ ID NO:	82	MUM1	84939
SEQ ID NO:	83	NAP1L4	4676
SEQ ID NO:	84	NCAPH	23397
SEQ ID NO:	85	NCF2	4688

TABLE 1: GROUP I (151 Gene set)

SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	86	NEK2	4751
SEQ ID NO:	87	NEU1	4758
SEQ ID NO:	88	NRBP1	29959
SEQ ID NO:	89	NSMCE4A	54780
SEQ ID NO:	90	OPTN	10133
SEQ ID NO:	91	ORC5L	5001
SEQ ID NO:	92	PBXIP1	57326
SEQ ID NO:	93	PFKM	5213
SEQ ID NO:	94	PKMYT1	9088
SEQ ID NO:	95	POLD1	5424
SEQ ID NO:	96	POLE	5426
SEQ ID NO:	97	PPP1R8	5511
SEQ ID NO:	98	PRDM1	639
SEQ ID NO:	99	PRR13	54458
SEQ ID NO:	100	PSCDBP	9595
SEQ ID NO:	101	QDPR	5860
SEQ ID NO:	102	R3HCC1	203069
SEQ ID NO:	103	RAB31	11031
SEQ ID NO:	104	RACGAP1	29127
SEQ ID NO:	105	RAD51	5888
SEQ ID NO:	106	RAD54B	25788
SEQ ID NO:	107	RASGRP3	25780
SEQ ID NO:	108	RECQL4	9401
SEQ ID NO:	109	RFC2	5982
SEQ ID NO:	110	RNF41	10193
SEQ ID NO:	111	RRM1	6240
SEQ ID NO:	112	SAE1	10055
SEQ ID NO:	113	SAE2, also called UBA2	10054
SEQ ID NO:	114	SAMHD1	25939
SEQ ID NO:	115	SCAMP2	10066
SEQ ID NO:	116	SCARB2	950
SEQ ID NO:	117	SCPEP1	59342
SEQ ID NO:	118	SEMA4A	64218
SEQ ID NO:	119	SERINC1	57515
SEQ ID NO:	120	SETD6	79918
SEQ ID NO:	121	SFXN3	81855
SEQ ID NO:	122	SLAMF1	6504
SEQ ID NO:	123	SMARCD2	6603
SEQ ID NO:	124	SQSTM1	8878
SEQ ID NO:	125	STAB1	23166
SEQ ID NO:	126	STAT4	6775
SEQ ID NO:	127	STX11	8676
SEQ ID NO:	128	TAF4	6874

TABLE 1: GROUP I (151 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	129	TDRD7	23424
SEQ ID NO:	130	TENC1	23371
SEQ ID NO:	131	TIPIN	54962
SEQ ID NO:	132	TMEM186	25880
SEQ ID NO:	133	TMEM214 (also called FLJ20254; FLJ39682)	54867
SEQ ID NO:	134	TNIP1	10318
SEQ ID NO:	135	TOPBP1	11073
SEQ ID NO:	136	TRAIP	10293
SEQ ID NO:	137	TRIB2	28951
SEQ ID NO:	138	TRIM28	10155
SEQ ID NO:	139	TROAP	10024
SEQ ID NO:	140	TRPV2	51393
SEQ ID NO:	141	TUBB6	84617
SEQ ID NO:	142	UGCG	7357
SEQ ID NO:	143	UNG	7374
SEQ ID NO:	144	USP1	7398
SEQ ID NO:	145	USP32	84669
SEQ ID NO:	146	WARS2	10352
SEQ ID NO:	147	WBP2	23558
SEQ ID NO:	148	WDHD1	11169
SEQ ID NO:	149	WDR76	79968
SEQ ID NO:	150	WEE1	7465
SEQ ID NO:	151	ZFYVE26	23503

Table 2: Group II (subgroup of 136 genes from Group I) cancer biomarkers.

TABLE 2: GROUP II (136 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	1	ABCA3	21
SEQ ID NO:	2	ACADVL	37
SEQ ID NO:	3	ACP6	51205
SEQ ID NO:	4	ALG6	29929
SEQ ID NO:	5	APH1B	83464
SEQ ID NO:	6	ARHGEF3	50650
SEQ ID NO:	7	ATAD5	79915
SEQ ID NO:	8	ATF6	22926
SEQ ID NO:	9	ATP6V0A1	535
SEQ ID NO:	10	BRIP1	83990
SEQ ID NO:	11	BSDC1	55108
SEQ ID NO:	12	C12ORF29	91298

TABLE 2: GROUP II (136 Gene set)

SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	13	C14ORF104	55172
SEQ ID NO:	14	C14ORF124	56948
SEQ ID NO:	15	C16ORF80	29105
SEQ ID NO:	16	C18ORF24	220134
SEQ ID NO:	17	C21ORF45	54069
SEQ ID NO:	18	C5ORF15	56951
SEQ ID NO:	19	CCNA2	890
SEQ ID NO:	21	CCPG1	9236
SEQ ID NO:	22	CCR1	1230
SEQ ID NO:	23	CD47	961
SEQ ID NO:	24	CD80	941
SEQ ID NO:	25	CD83	9308
SEQ ID NO:	26	CDC25A	993
SEQ ID NO:	28	CEP70	80321
SEQ ID NO:	30	CHCHD7	79145
SEQ ID NO:	31	CHEK1	1111
SEQ ID NO:	32	CISD1	55847
SEQ ID NO:	33	CLEC10A	10462
SEQ ID NO:	34	CLUAP1	23059
SEQ ID NO:	35	CSF1R	1436
SEQ ID NO:	36	CST3	1471
SEQ ID NO:	37	CTSO	1519
SEQ ID NO:	38	CX3CR1	1524
SEQ ID NO:	39	CXCR7	57007
SEQ ID NO:	40	DAB2	1601
SEQ ID NO:	41	DDX39	10212
SEQ ID NO:	42	DNA2	1763
SEQ ID NO:	43	DNMT1	1786
SEQ ID NO:	44	DSE	29940
SEQ ID NO:	46	E2F1	1869
SEQ ID NO:	47	E2F8	79733
SEQ ID NO:	48	EBI2	1880
SEQ ID NO:	49	EHD4	30844
SEQ ID NO:	50	EIF5A2	56648
SEQ ID NO:	51	ETAA1	54465
SEQ ID NO:	52	EVI2A	2123
SEQ ID NO:	53	FAM128B	80097
SEQ ID NO:	54	FAM129A	116496
SEQ ID NO:	55	FBXO21	23014
SEQ ID NO:	56	FLII	2314
SEQ ID NO:	58	GABARAP	11337
SEQ ID NO:	59	GALNT11	63917
SEQ ID NO:	60	GPSM2	29899

TABLE 2: GROUP II (136 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	61	HADH	3033
SEQ ID NO:	62	HBP1	26959
SEQ ID NO:	63	HCK	3055
SEQ ID NO:	64	HDDC2	51020
SEQ ID NO:	66	ID3	3399
SEQ ID NO:	68	IL2RA	3559
SEQ ID NO:	69	ITGB2	3689
SEQ ID NO:	70	LAMB3	3914
SEQ ID NO:	71	LRDD	55367
SEQ ID NO:	73	LY6E	4061
SEQ ID NO:	74	MAF	4094
SEQ ID NO:	75	MAN2A1	4124
SEQ ID NO:	76	MAP3K8	1326
SEQ ID NO:	77	MAZ	4150
SEQ ID NO:	79	MCM7	4176
SEQ ID NO:	80	MPHOSPH9	10198
SEQ ID NO:	81	MRE11A	4361
SEQ ID NO:	82	MUM1	84939
SEQ ID NO:	83	NAP1L4	4676
SEQ ID NO:	84	NCAPH	23397
SEQ ID NO:	85	NCF2	4688
SEQ ID NO:	86	NEK2	4751
SEQ ID NO:	87	NEU1	4758
SEQ ID NO:	88	NRBP1	29959
SEQ ID NO:	89	NSMCE4A	54780
SEQ ID NO:	90	OPTN	10133
SEQ ID NO:	91	ORC5L	5001
SEQ ID NO:	92	PBXIP1	57326
SEQ ID NO:	94	PKMYT1	9088
SEQ ID NO:	95	POLD1	5424
SEQ ID NO:	97	PPP1R8	5511
SEQ ID NO:	98	PRDM1	639
SEQ ID NO:	99	PRR13	54458
SEQ ID NO:	100	PSCDBP	9595
SEQ ID NO:	101	QDPR	5860
SEQ ID NO:	102	R3HCC1	203069
SEQ ID NO:	103	RAB31	11031
SEQ ID NO:	104	RACGAP1	29127
SEQ ID NO:	105	RAD51	5888
SEQ ID NO:	106	RAD54B	25788
SEQ ID NO:	107	RASGRP3	25780
SEQ ID NO:	108	RECQL4	9401
SEQ ID NO:	109	RFC2	5982

TABLE 2: GROUP II (136 Gene set)

SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	110	RNF41	10193
SEQ ID NO:	111	RRM1	6240
SEQ ID NO:	112	SAE1	10055
SEQ ID NO:	113	SAE2, also called UBA2	10054
SEQ ID NO:	114	SAMHD1	25939
SEQ ID NO:	115	SCAMP2	10066
SEQ ID NO:	116	SCARB2	950
SEQ ID NO:	117	SCPEP1	59342
SEQ ID NO:	118	SEMA4A	64218
SEQ ID NO:	119	SERINC1	57515
SEQ ID NO:	121	SFXN3	81855
SEQ ID NO:	122	SLAMF1	6504
SEQ ID NO:	123	SMARCD2	6603
SEQ ID NO:	124	SQSTM1	8878
SEQ ID NO:	125	STAB1	23166
SEQ ID NO:	126	STAT4	6775
SEQ ID NO:	127	STX11	8676
SEQ ID NO:	129	TDRD7	23424
SEQ ID NO:	130	TENC1	23371
SEQ ID NO:	131	TIPIN	54962
SEQ ID NO:	132	TMEM186	25880
SEQ ID NO:	133	TMEM214 (also called FLJ20254; FLJ39682)	54867
SEQ ID NO:	134	TNIP1	10318
SEQ ID NO:	135	TOPBP1	11073
SEQ ID NO:	137	TRIB2	28951
SEQ ID NO:	138	TRIM28	10155
SEQ ID NO:	139	TROAP	10024
SEQ ID NO:	140	TRPV2	51393
SEQ ID NO:	142	UGCG	7357
SEQ ID NO:	143	UNG	7374
SEQ ID NO:	144	USP1	7398
SEQ ID NO:	145	USP32	84669
SEQ ID NO:	146	WARS2	10352
SEQ ID NO:	147	WBP2	23558
SEQ ID NO:	148	WDHD1	11169
SEQ ID NO:	149	WDR76	79968
SEQ ID NO:	150	WEE1	7465
SEQ ID NO:	151	ZFYVE26	23503

Table 3: Group III (subgroup of 91 genes from Group I) cancer biomarkers.

TABLE 3: GROUP III (91 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	1	ABCA3	21
SEQ ID NO:	2	ACADVL	37
SEQ ID NO:	3	ACP6	51205
SEQ ID NO:	4	ALG6	29929
SEQ ID NO:	6	ARHGEF3	50650
SEQ ID NO:	7	ATAD5	79915
SEQ ID NO:	9	ATP6V0A1	535
SEQ ID NO:	10	BRIP1	83990
SEQ ID NO:	11	BSDC1	55108
SEQ ID NO:	19	CCNA2	890
SEQ ID NO:	22	CCR1	1230
SEQ ID NO:	24	CD80	941
SEQ ID NO:	25	CD83	9308
SEQ ID NO:	27	CDT1	81620
SEQ ID NO:	28	CEP70	80321
SEQ ID NO:	32	CISD1	55847
SEQ ID NO:	33	CLEC10A	10462
SEQ ID NO:	34	CLUAP1	23059
SEQ ID NO:	35	CSF1R	1436
SEQ ID NO:	36	CST3	1471
SEQ ID NO:	38	CX3CR1	1524
SEQ ID NO:	39	CXCR7	57007
SEQ ID NO:	40	DAB2	1601
SEQ ID NO:	42	DNA2	1763
SEQ ID NO:	43	DNMT1	1786
SEQ ID NO:	44	DSE	29940
SEQ ID NO:	46	E2F1	1869
SEQ ID NO:	47	E2F8	79733
SEQ ID NO:	48	EBI2	1880
SEQ ID NO:	49	EHD4	30844
SEQ ID NO:	50	EIF5A2	56648
SEQ ID NO:	52	EVI2A	2123
SEQ ID NO:	53	FAM128B	80097
SEQ ID NO:	58	GABARAP	11337
SEQ ID NO:	59	GALNT11	63917
SEQ ID NO:	60	GPSM2	29899
SEQ ID NO:	61	HADH	3033
SEQ ID NO:	62	HBP1	26959
SEQ ID NO:	66	ID3	3399
SEQ ID NO:	68	IL2RA	3559
SEQ ID NO:	69	ITGB2	3689
SEQ ID NO:	71	LRDD	55367

TABLE 3: GROUP III (91 Gene set)

SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	74	MAF	4094
SEQ ID NO:	76	MAP3K8	1326
SEQ ID NO:	77	MAZ	4150
SEQ ID NO:	80	MPHOSPH9	10198
SEQ ID NO:	82	MUM1	84939
SEQ ID NO:	84	NCAPH	23397
SEQ ID NO:	86	NEK2	4751
SEQ ID NO:	87	NEU1	4758
SEQ ID NO:	88	NRBP1	29959
SEQ ID NO:	89	NSMCE4A	54780
SEQ ID NO:	90	OPTN	10133
SEQ ID NO:	91	ORC5L	5001
SEQ ID NO:	92	PBXIP1	57326
SEQ ID NO:	94	PKMYT1	9088
SEQ ID NO:	95	POLD1	5424
SEQ ID NO:	98	PRDM1	639
SEQ ID NO:	99	PRR13	54458
SEQ ID NO:	100	PSCDBP	9595
SEQ ID NO:	101	QDPR	5860
SEQ ID NO:	102	R3HCC1	203069
SEQ ID NO:	103	RAB31	11031
SEQ ID NO:	107	RASGRP3	25780
SEQ ID NO:	108	RECQL4	9401
SEQ ID NO:	109	RFC2	5982
SEQ ID NO:	115	SCAMP2	10066
SEQ ID NO:	116	SCARB2	950
SEQ ID NO:	117	SCPEP1	59342
SEQ ID NO:	118	SEMA4A	64218
SEQ ID NO:	120	SETD6	79918
SEQ ID NO:	122	SLAMF1	6504
SEQ ID NO:	124	SQSTM1	8878
SEQ ID NO:	125	STAB1	23166
SEQ ID NO:	126	STAT4	6775
SEQ ID NO:	127	STX11	8676
SEQ ID NO:	128	TAF4	6874
SEQ ID NO:	129	TDRD7	23424
SEQ ID NO:	133	TMEM214 (also called FLJ20254; FLJ39682)	54867
SEQ ID NO:	135	TOPBP1	11073
SEQ ID NO:	140	TRPV2	51393
SEQ ID NO:	141	TUBB6	84617
SEQ ID NO:	142	UGCG	7357
SEQ ID NO:	143	UNG	7374

TABLE 3: GROUP III (91 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	144	USP1	7398
SEQ ID NO:	145	USP32	84669
SEQ ID NO:	147	WBP2	23558
SEQ ID NO:	148	WDHD1	11169
SEQ ID NO:	149	WDR76	79968
SEQ ID NO:	150	WEE1	7465
SEQ ID NO:	151	ZFYVE26	23503

Table 4: Group IV (subgroup of 51 genes from Group I) cancer biomarkers.

TABLE 4: GROUP IV (51 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	12	C12ORF29	91298
SEQ ID NO:	19	CCNA2	890
SEQ ID NO:	20	CCNB2	9133
SEQ ID NO:	22	CCR1	1230
SEQ ID NO:	26	CDC25A	993
SEQ ID NO:	29	CHAF1B	8208
SEQ ID NO:	31	CHEK1	1111
SEQ ID NO:	35	CSF1R	1436
SEQ ID NO:	42	DNA2	1763
SEQ ID NO:	43	DNMT1	1786
SEQ ID NO:	45	DUT (includes EG:1854)	1854
SEQ ID NO:	60	GPSM2	29899
SEQ ID NO:	61	HADH	3033
SEQ ID NO:	62	HBPI	26959
SEQ ID NO:	63	HCK	3055
SEQ ID NO:	65	HRAS	3265
SEQ ID NO:	67	IDH3A	3419
SEQ ID NO:	69	ITGB2	3689
SEQ ID NO:	72	LSM6	11157
SEQ ID NO:	74	MAN2A1	4124
SEQ ID NO:	77	MAZ	4150
SEQ ID NO:	78	MBD3 (includes EG:53615)	53615
SEQ ID NO:	80	MPHOSPH9 (includes EG:10198)	10198
SEQ ID NO:	81	MRE11A	4361
SEQ ID NO:	84	NCAPH	23397
SEQ ID NO:	86	NEK2	4751
SEQ ID NO:	90	OPTN	10133

TABLE 4: GROUP IV (51 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	91	ORC5L (includes EG:5001)	5001
SEQ ID NO:	92	PBXIP1	57326
SEQ ID NO:	93	PFKM	5213
SEQ ID NO:	94	PKMYT1	9088
SEQ ID NO:	95	POLD1	5424
SEQ ID NO:	96	POLE	5426
SEQ ID NO:	97	PPP1R8	5511
SEQ ID NO:	103	RAB31	11031
SEQ ID NO:	108	RECQL4	9401
SEQ ID NO:	109	RFC2	5982
SEQ ID NO:	111	RRM1	6240
SEQ ID NO:	113	SAE2, also called UBA2	10054
SEQ ID NO:	121	SFXN3	81855
SEQ ID NO:	124	SQSTM1	8878
SEQ ID NO:	125	STAB1	23166
SEQ ID NO:	126	STAT4	6775
SEQ ID NO:	127	STX11	8676
SEQ ID NO:	134	TNIP1	10318
SEQ ID NO:	135	TOPBP1 (includes EG:11073)	11073
SEQ ID NO:	136	TRAIP	10293
SEQ ID NO:	143	UNG	7374
SEQ ID NO:	144	USP1	7398
SEQ ID NO:	148	WDHD1	11169
SEQ ID NO:	150	WEE1	7465

Table 5: Group V (subgroup of 14 genes from Group I) cancer biomarkers.

TABLE 5: GROUP V (14 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	10	BRIP1	83990
SEQ ID NO:	42	DNA2	1763
SEQ ID NO:	43	DNMT1	1786
SEQ ID NO:	79	MCM7	4176
SEQ ID NO:	81	MRE11A	4361
SEQ ID NO:	83	NAP1L4	4676
SEQ ID NO:	91	ORC5L	5001
SEQ ID NO:	95	POLD1	5424

TABLE 5: GROUP V (14 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	105	RAD51	5888
SEQ ID NO:	106	RAD54B	25788
SEQ ID NO:	108	RECQL4	9401
SEQ ID NO:	128	RFC2	5982
SEQ ID NO:	135	TOPBP1	11073
SEQ ID NO:	143	UNG	7374

Table 6: Group VI (subgroup of 7 genes from Group I) cancer biomarkers.

TABLE 6: GROUP VI (7 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	12	C12orf29	91298
SEQ ID NO:	13	C14orf104	55172
SEQ ID NO:	14	C14orf124	56948
SEQ ID NO:	15	C16orf80	29105
SEQ ID NO:	16	C18orf24	220134
SEQ ID NO:	17	C21orf45	54069
SEQ ID NO:	18	C5orf15	56951

Table 7: Group VII (subgroup of 6 genes from Group I) cancer biomarkers.

TABLE 7: GROUP VII (6 Gene set)			
SEQ ID NO:		Gene symbol	Entrez GeneID
SEQ ID NO:	79	MCM7	4176
SEQ ID NO:	91	ORC5L	5001
SEQ ID NO:	95	POLD1	5424
SEQ ID NO:	96	POLE	5426
SEQ ID NO:	109	RFC2	5982
SEQ ID NO:	135	TOPBP1	11073

[0190] In some embodiments, the group of cancer biomarkers or subgroups thereof can be detected at the level of gene expression, for example the gene transcript level such as mRNA

expression. In alternative embodiments, the group of cancer biomarkers as disclosed herein or subgroups thereof can be detected at the level of protein expression.

[0191] The inventors provide detailed guidance on the increase and/or decrease of gene expression and/or protein expression of a group of cancer biomarkers for the diagnosis and/or prognosis of cancer in a subject (e.g., leukemia, lymphoma, glioma, breast cancer, colon cancer, or lung cancer). In some embodiments, the cancer biomarkers or subgroups thereof as disclosed herein can be used for the diagnosis and/or prognosis of cancer in a subject where the cancer is a glioma, breast cancer, glioma, lung cancer, colon cancer, and/or any cancer that has a proliferating phenotype.

[0192] In one aspect of the present invention, the group of cancer biomarkers useful in the diagnosis and/or prognosis of cancer in a subject is a group of 151 genes as set forth in Table 1 (Group I). In another aspect of the present invention, the group of cancer biomarkers useful in the diagnosis and/or prognosis of cancer in a subject is a subgroup of 136 genes set forth in Table 2 (Group II). In another aspect of the present invention, the group of cancer biomarkers useful in the diagnosis and/or prognosis of cancer in a subject is a subgroup of 91 genes set forth in Table 3 (Group III). In another aspect of the present invention, the group of cancer biomarkers useful in the diagnosis and/or prognosis of cancer in a subject is a subgroup of 51 genes set forth in Table 4 (Group IV). In another aspect of the present invention, a cancer biomarker useful in the diagnosis and/or prognosis of cancer in a subject is the gene RECQL4 (SEQ ID NO: 108).

[0193] In some embodiments, the cancer biomarker gene RECQL4 (SEQ ID NO: 108) can be used as a sole cancer biomarker using the methods as disclosed herein, or alternatively it can be combined with at least 1 other cancer biomarker selected from Groups I-IV. Alternatively, RECQL4 (SEQ ID NO: 108) can be combined with at least 2, or at least 5 or more, cancer biomarkers selected from Groups I-IV. In some embodiments, the cancer biomarker gene RECQL4 (SEQ ID NO: 108) can be used as a sole cancer biomarker using the methods as disclosed herein, or alternatively it can be combined with at least 1 other cancer biomarker selected from Group V. Alternatively, RECQL4 (SEQ ID NO: 108) can be combined with at least 2 or at least 5 or more cancer biomarkers selected from Group V, for example, at least 1, or at least 2, or at least 3, or at least 4, or at least 5, or at least 6, or at least 7, or at least 8, or at least 9, or at least 10, 11, 12 or at least 13, in any combination of the following list of genes; BRIP1 (SEQ ID NO: 10); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); MCM7 (SEQ ID NO: 79); MRE11A (SEQ ID NO: 81); NAP1L4 (SEQ ID NO: 83); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID

NO: 106); RFC2 (SEQ ID NO: 109); TOPBP1 RFC2 (SEQ ID NO: 135); or UNG (SEQ ID NO: 143)

[0194] In some embodiments, the group of cancer biomarkers useful in the methods and compositions as disclosed herein can be identified by their NCBI Gene ID Numbers (Entrez ID) as shown in Tables 1-7.

[0195] The inventors have further discovered that taking a sub-group or sub-groups of genes from the initially identified group of biomarkers provides a much greater diagnostic and/or prognostic capability than chance alone. Preferably, a subgroup of cancer biomarkers comprises at least two, or at least three cancer biomarkers from the group of cancer biomarkers set forth in Tables 1 to 7. In some embodiments, a subgroup of cancer biomarkers comprises at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, 11, 12 or at least 13 cancer biomarkers from the group of cancer biomarkers as set forth in Table 1 to Table 7. As disclosed herein, a subgroup of cancer biomarkers comprises RECQL4 (SEQ ID NO: 108) and at least one, or at least two, or at least three cancer biomarkers from the group of cancer biomarkers set forth in Table 5.

[0196] It is noted that one can use any combination of Group I (151 gene set) cancer biomarkers as set forth in Table 1, or any combination of a subgroup of Group I cancer biomarkers, i.e. any combination of Group II (136 gene subset), Group III (91 gene subset), Group IV (51 gene subset), Group V (14 gene subset), Group VI (7 gene subgroup) or Group VII (6 gene subgroup) of biomarkers as set forth in Tables 2-7 in the methods as disclosed herein. In one embodiment, one can enhance the accuracy of diagnosis by adding additional genes to Group I cancer biomarkers as listed in Table 1 or subgroups (such as Groups II-V) thereof. In such an embodiment, the additional genes can be any gene, for example other cancer biomarker genes, and in particular any other cancer biomarker gene that is not listed in Table 1.

[0197] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of positively correlated genes from Group I (151 gene set), where the subgroup of positively correlated genes from Group I is referred to herein as Group VIII and consists of: ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ATAD5 (SEQ ID NO: 7); BRIP1 (SEQ ID NO: 10); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLUAP1 (SEQ ID NO: 34); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID

NO: 43); DUT (includes EG:1854) (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); FAM128B (SEQ ID NO: 53); FBXO21 (SEQ ID NO: 55); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); LRDD (SEQ ID NO: 71); LSM6 (SEQ ID NO: 72); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NSMCE4A (SEQ ID NO: 89); ORC5L (SEQ ID NO: 91); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SETD6 (SEQ ID NO: 120); SMARCD2 (SEQ ID NO: 123); TAF4 (SEQ ID NO: 128); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TOPBP1 (SEQ ID NO: 135); TRAIIP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WARS2 (SEQ ID NO: 146); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149) and WEE1 (SEQ ID NO: 150) or having substantial identity thereto.

[0198] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of positively correlated genes from Group II (136 subgroup gene set), where the subgroup of positively correlated genes from Group II is referred to herein as Group X and consists of: ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ATAD5 (SEQ ID NO: 7); BRIP1 (SEQ ID NO: 10); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); CCNA2 (SEQ ID NO: 19); CDC25A (SEQ ID NO: 26); CEP70 (SEQ ID NO: 28); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLUAP1 (SEQ ID NO: 34); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); FAM128B (SEQ ID NO: 53); FBXO21 (SEQ ID NO: 55); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HDDC2 (SEQ ID NO: 64); ID3 (SEQ ID NO: 66); LRDD (SEQ ID NO: 71); MAZ (SEQ ID NO: 77); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81);

MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NSMCE4A (SEQ ID NO: 89); ORC5L (SEQ ID NO: 91); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PPP1R8 (SEQ ID NO: 97); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SMARCD2 (SEQ ID NO: 123); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TOPBP1 (SEQ ID NO: 135); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WARS2 (SEQ ID NO: 146); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149) and WEE1 (SEQ ID NO: 150), or having substantial identity thereto.

[0199] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of positively correlated genes from Group III (91 subgroup gene set), where the subgroup of positively correlated genes from Group III is referred to herein as Group XII and consists of: ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ATAD5 (SEQ ID NO: 7); BRIP1 (SEQ ID NO: 10); CCNA2 (SEQ ID NO: 19); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CISD1 (SEQ ID NO: 32); CLUAP1 (SEQ ID NO: 34); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DUT (includes EG:1854) (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EIF5A2 (SEQ ID NO: 50); FAM128B (SEQ ID NO: 53); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); ID3 (SEQ ID NO: 66); LRDD (SEQ ID NO: 71); MAZ (SEQ ID NO: 77); MPHOSPH9 (SEQ ID NO: 80); MUM1 (SEQ ID NO: 82); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NSMCE4A (SEQ ID NO: 89); ORC5L (SEQ ID NO: 91); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); SETD6 (SEQ ID NO: 120); TAF4 (SEQ ID NO: 128); TOPBP1 (SEQ ID NO: 135); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149) and WEE1 (SEQ ID NO: 150) or having substantial identity thereto.

[0200] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of positively correlated genes from Group IV (51 subgroup gene set), where the subgroup of positively correlated genes from Group IV is referred to herein as **Group XIV** and consists of: C12ORF29 (SEQ ID NO: 12); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CDC25A (SEQ ID NO: 26); CHAF1B (SEQ ID NO: 29); CHEK1 (SEQ ID NO:

31); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DUT (includes EG:1854) (SEQ ID NO: 45); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HRAS (SEQ ID NO: 65); IDH3A (SEQ ID NO: 67); LSM6 (SEQ ID NO: 72); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); ORC5L (SEQ ID NO: 91); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RRM1 (SEQ ID NO: 111); SAE2, also called UBA2 (SEQ ID NO: 113); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WDHD1 (SEQ ID NO: 148); WEE1 (SEQ ID NO: 150) or having substantial identity thereto.

[0201] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of negatively correlated genes from Group I (151 gene set), where the subgroup of negatively correlated genes from Group I is referred to herein as Group IX and consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BSDC1 (SEQ ID NO: 11); C5ORF15 (SEQ ID NO: 18); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CLEC10A (SEQ ID NO: 33); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DSE (SEQ ID NO: 44); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EVI2A (SEQ ID NO: 52); FAM129A (SEQ ID NO: 54); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); NCF2 (SEQ ID NO: 85); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); OPTN (SEQ ID NO: 90); PBXIP1 (SEQ ID NO: 92); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TRPV2 (SEQ ID NO: 140);

TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); ZFYVE26 (SEQ ID NO: 151) or having substantial identity thereto.

[0202] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of negatively correlated genes from Group II (136 subgroup gene set), where the subgroup of negatively correlated genes from Group II is referred to herein as Group XI and consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BSDC1 (SEQ ID NO: 11); C5ORF15 (SEQ ID NO: 18); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CLEC10A (SEQ ID NO: 33); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DSE (SEQ ID NO: 44); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EVI2A (SEQ ID NO: 52); FAM129A (SEQ ID NO: 54); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); NCF2 (SEQ ID NO: 85); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); OPTN (SEQ ID NO: 90); PBXIP1 (SEQ ID NO: 92); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TRPV2 (SEQ ID NO: 140); UGCG (SEQ ID NO: 142); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); ZFYVE26 (SEQ ID NO: 151) or having substantial identity thereto.

[0203] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of negatively correlated genes from Group III (91 subgroup gene set), where the subgroup of negatively correlated genes from Group III is referred to herein as Group XIII and consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ARHGEF3 (SEQ ID NO: 6); ATP6V0A1 (SEQ ID NO: 9); BSDC1 (SEQ ID NO: 11); CCR1 (SEQ ID NO: 22); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CLEC10A (SEQ ID NO: 33); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID

NO: 39); DAB2 (SEQ ID NO: 40); DSE (SEQ ID NO: 44); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EVI2A (SEQ ID NO: 52); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); HBP1 (SEQ ID NO: 62); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); MAF (SEQ ID NO: 74); MAP3K8 (SEQ ID NO: 76); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); OPTN (SEQ ID NO: 90); PBXIP1 (SEQ ID NO: 92); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); ZFYVE26 (SEQ ID NO: 151) or having substantial identity thereto.

[0204] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of negatively correlated genes from Group IV (51 subgroup gene set), where the subgroup of negatively correlated genes from Group IV is referred to herein as Group XV and consists of: CCR1 (SEQ ID NO: 22); CSF1R (SEQ ID NO: 35); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); ITGB2 (SEQ ID NO: 69); MAN2A1 (SEQ ID NO: 75); OPTN (SEQ ID NO: 90); PBXIP1 (SEQ ID NO: 92); RAB31 (SEQ ID NO: 103); SFXN3 (SEQ ID NO: 121); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127) and TNIP1 (SEQ ID NO: 134) or having substantial identity thereto.

[0205] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of replication factor genes; CCNA2; CDC25A; CHEK1; DNA2; ORC5L; POLD1; POLE; RFC2; WEE1; CCNB2 and RECQL4 or having substantial identity thereto. (Subgroup A).

[0206] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of DNA repair genes; CHEK1; MRE11A; UNG; WEE1 and RECQL4 or having substantial identity thereto. (Subgroup B).

[0207] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of nuclear factor genes; STAT4; TNIP1; HBP1; PBXIP1; CCNA2; CDC25A; CHEK1; DNMT1; MAZ; MRE11A; NEK2; ORC5L; POLD1; POLE; PPP1R8; RFC2; RRM1; UNG; WEE1; CHAF1B; RECQL4; TOPBP1; LSM6; WDHD1; NCAPH; GPSM2 and MBD3, or having substantial identity thereto. (Subgroup C).

- [0208] In some embodiments, a subgroup of cancer biomarkers is selected from the group consisting of cytoplasmic genes; HCK; MAN2A1; SQSTM1; OPTN; RAB31; SFXN3; HADH; IDH3A; PFKM; USP1; CCNB2; SAE2; MPHOSPH9 and TRAIP, or having substantial identity thereto. (Subgroup D).
- [0209] In some embodiments, the cancer biomarker is FOXM1 (SEQ ID NO:57) or RECQL4 (SEQ ID NO: 108).
- [0210] In some embodiments, reduced expression of FOXM1 (SEQ ID NO:57) or RECQL4 (SEQ ID NO: 108) is a good prognosis.
- [0211] In one embodiment, the present invention provides gene groups, the expression profile of which can be used in methods to diagnose cancer in a subject with more than 60% accuracy, preferably more than 65%, more than 70% more than 75%, or preferably approximately 80%-95% accuracy.
- [0212] In one embodiment the present invention provides groups of genes, for which the expression profile provides a diagnostic and/or prognostic test to detect cancer in a subject. For example, in one embodiment, the present invention provides groups of genes, for which the expression profiles distinguish subjects with cancer from subjects without cancer.
- [0213] In one embodiment, the present invention provides a screening system for cancer by analyzing the expression profile of prognostic cancer biomarkers as disclosed herein. Such screening can be performed, for example on a biological sample taken from a subject, such as a tumor biopsy sample. For example, where the subject is diagnosed of having a lymphoma, the cancer biomarkers as disclosed herein can be analyzed on cells obtained from the subject from a blood sample to type the specific lymphoma, e.g. Burkitt lymphoma vs. diffuse large B-cell lymphoma (DLBCL) vs. follicular lymphoma vs. non-Hodgkin's lymphoma , etc..
- [0214] In some embodiments, one can analyze the expression of cancer biomarkers (or a subgroup of cancer biomarkers) as disclosed herein, as compared to the expression of the same group and/or a subgroup of cancer biomarkers in a reference sample, for example a reference biological sample, or to a reference level as defined herein.
- [0215] In some embodiments, the reference sample can be from a reference biological sample or a group of reference samples, for example such tissues can be normal tissue from the subject, or a biological sample from a subject that does not have cancer.
- [0216] In some embodiments, a metagene score of a cancer biomarker in the subject, such as the genes in Group I as disclosed herein in Table 1, or in Groups II, III, IV, V, VI or VII or subgroup A-D can be used to identify a subject with cancer. In one embodiment, a metagene value between 0.001 and +1 indicates poor prognosis of a subject, for example decreased

survival and/or a more aggressive tumor. In some embodiments, a metagene score ranging between 0.001 and +1, including all integers between 0.001 and +1.0, i.e. at least about +0.05, at least about +0.1, at least about +0.2, at least about +0.3, at least about +0.4, at least about +0.5, at least about +0.6, at least about +0.7, at least about +0.8, at least about +0.9, or even +1.0 identifies a poor prognosis of a subject, for example the likelihood of decreased survival and/or a more aggressive tumor as compared to a subject with a metavalues within the ranges of 0 and -1. It is important to note that the metagene score can also provide a scale of severity of disease, such that a score at the higher end of the range (i.e., closer to +1) indicates a cancer is likely more severe than a cancer with a metagene score at the lower end of the range (i.e., 0.1).

[0217] In some embodiments, a metagene value between 0 and -1 indicates that a subject has a good prognosis, such as an increased survival and a less aggressive or non-aggressive cancer. In some embodiments, a metagene score of between 0 and -1, including all integers between 0 and -1.0, i.e. at least about -0.05, at least about -0.1, at least about -0.2, at least about -0.3, at least about -0.4, at least about -0.5, at least about -0.6, at least about -0.7, at least about -0.8, at least about -0.9, or even -1.0 identifies a subject with a good prognosis, for example the likelihood of increased survival and a less aggressive tumor as compared to a subject with a metavalues in the ranges of 0.001 to +1.

[0218] In some embodiments, a metagene score of a cancer biomarker in the subject, such as the genes of Group I as disclosed herein in Table 1, or in Groups II, III, IV, V, VI or VII as disclosed herein in Tables 2-7 respectively; Groups VIII, IX, X, XI, XII, XIII, XIV, XV or subgroup A-D can be used to identify a change in the cancer prognosis of a subject with cancer. For example, a subject can be identified to have a metascore at a first timepoint which can be used as a baseline to assess if there is a change in the metascore at a second timepoint. As an illustrative example, if a subject is identified as having a metascore of +1 at a first timepoint, and thus the subject was identified as having an aggressive form of cancer (i.e. poor prognosis), a decrease in the metascore measured at a second timepoint can be used to determine an improvement in prognosis (i.e., a decrease in tumor aggressiveness). In some embodiments, a decrease in metagene value between 0.05 and 1.99 identifies a subject with increased or improved prognosis. For example, a decrease in a metagene score of between 0.05 and 1.99, including all integers between 0.05 and 1.99, i.e. a decrease of at least about 0.05, at least about 0.1, at least about 0.2, at least about 0.3, at least about 0.4, at least about 0.5, at least about 0.6, at least about 0.7, at least about 0.8, at least about 0.9, at least about 1.0, at least about 1.1, at least about 1.2, at least about 1.3, at least about 1.4, at least about

1.5, at least about 1.6, at least about 1.8, at least about 1.9, or even a decrease of about 1.99 of the metagene score taken at a second timepoint as compared to the metagene score taken at a first timepoint identifies the subject with a better prognosis (such as, for example, a decreased tumor aggressiveness, an increase in longevity and/or an increase in survival) at the second timepoint as compared to the time when the first metascore was taken.

[0219] The degree of change in expression of a cancer biomarker is compared to the reference database to identify the degree of severity (i.e., degree of aggressiveness of a cancer) of disease. For example, the cancer biomarker expression of the subject correlates with the expression level in the reference database of an aggressive tumor. Thus, the degree of correlation of a measured and normalized expression value in a cancer biomarker provides a scale for which a skilled clinician can predict the prognosis of a patient, as well as any improvements in prognosis exhibited upon anti-cancer treatment. Further the reference expression database can allow a skilled clinician to identify the cancer subtype.

[0220] It is contemplated herein that an expression profile comprises some cancer biomarkers that are positively correlated and others that are negatively correlated in subjects with cancer. Thus, it is not necessary that all of the biomarkers in an expression profile increase (or alternatively decrease), but rather there can be a heterogeneous mix of positively correlated and negatively correlated biomarkers in a single expression profile of an individual.

[0221] In some embodiments, reference levels useful in the methods as disclosed herein can be biological samples obtained from a subject or a group of subjects who have cancer, or are being treated for cancer. In some embodiments, reference levels can be obtained from databases, e.g., public databases such as Oncomine. In some embodiments, reference levels can be obtained from biological samples from tumor banks or cell banks. In some embodiments, reference levels can be obtained from biological samples from the same subject, for example the reference level can be the level in a biological sample obtained from the subject at one time point, for example an earlier (i.e. first) time point, which is useful as a reference level for comparison with a biological sample from the same subject obtained at a later (i.e. second) time point. Such embodiments are useful for prognosis, for example monitoring cancer disease progression or response to anti-cancer treatment, in a subject over a defined time period, for example from the time when the reference level (i.e. first biological sample) was obtained to the time when the second biological sample was obtained from the same subject.

[0222] In some embodiments, reference levels useful in the methods as disclosed herein are obtained from a population group (e.g., a group of individuals or subjects sharing a common

ethno-geographic origin). Reference levels can be reference levels from a population group such as those predicted to have representative levels of expression of one or more gene transcripts and/or proteins encoded. The reference level can also be expressed relative to a housekeeping gene, any other gene, or an expression profile.

[0223] In another embodiment, the present invention provides a group of genes that can be used as diagnostic or prognostic predictors of cancer in a subject. These genes were identified using probabilities with a t-test analysis and show differential gene expression in subjects with cancer. In particular, the expression profile of a group of genes comprising between 1 and 11, and all combinations in between, for example, 2, 3, 4, 5, 6, 7, 8, 9 and 10 gene transcripts selected from any of the genes listed in Group I (151 gene set) or any of the genes listed in Groups II (136 gene subgroup), or Group III (91 gene subgroup) or Group IV (51 gene subgroup) or Group V (14 gene subgroup), Group VI (7 gene subgroup) or Group VII (6 gene subgroup) can be used to diagnose the cancer type or the cancer prognosis, for cancers such as leukemia, lymphoma, glioma, lung cancer and breast cancer, among others. In such an embodiment, the expression pattern is compared to the gene expression pattern of the same group of genes in a control biological sample, such as a biological sample which does not have, or is not at risk of developing cancer.

[0224] In another embodiment, the expression profile of a group of genes comprising RECQL4 (SEQ ID NO: 108) and at least between 1 and 13, and all combinations in between, for example, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 gene transcripts selected from the 14 gene subset of Group V, can be used to diagnose a cancer type or determine a prognosis, for cancers such as leukemia, lymphoma, glioma, lung cancer and breast cancer, among others. In such an embodiment, the expression profile is compared to the gene expression profile of the same group of genes in a reference biological sample, such as a reference sample from an individual that does not have, or is not at risk of developing cancer.

[0225] In another embodiment, the prognostic cancer biomarker analysis can comprise analysis of a core cancer biomarker subgroup (here referred to "subgroup") of about 3 to 5, 5 to 7, 7 to 9 or 9 to 11 of any of the cancer biomarkers of Group I (151 gene set) as disclosed in Table 1. In some embodiments, the subgroup of cancer biomarkers useful in the diagnostic and prognostic methods described herein can be combined with other biomarker genes, for example but not limited to other biomarker genes for cancer. In some embodiments, the group of cancer biomarkers or subgroup thereof can be combined with any number of other genes, for example other biomarker genes such as cancer biomarkers comprising a group of

about 1, about 5, about 1-5, about 5-10, about 10-15, about 15-20, about 20-25, about 25-30 about 35-40 about 40-45 or about 45-50.

[0226] In another embodiment, the present invention provides a method for diagnosing a cancer type or subtype in a subject or for providing prognostic information for a subject with cancer, the method comprising obtaining a nucleic acid from a biological sample from the subject and measuring the gene transcript levels of at least 4, or at least 5 cancer biomarkers selected from the group of cancer biomarkers of Group I (151 gene set), or Group II (136 gene subgroup), or Group III (91 gene subgroup), or Group IV (51 gene subgroup), or Group V (14 gene subgroup), or Group VI (7 gene subgroup), or Group VII (6 gene subgroup), as disclosed in Tables 1-7 respectively, or Group VIII (81 gene subgroup of Group I) or Group IX (69 gene subgroup of Group I) or Group X (68 gene subgroup of Group II) or Group XI (68 gene subgroup of Group II) or Group XII (40 gene subgroup of Group III) or Group XIII (51 gene subgroup of Group III) or Group XIV (36 gene subgroup of Group IV) or Group XV (15 gene subgroup of Group IV), or Group A-D and comparing the level of the gene transcripts of the same group of cancer biomarkers in a reference sample, wherein a difference in the level of expression in the group of cancer biomarkers analyzed indicates the subject's risk of cancer progression. More specifically, an comparing the level of at least 4, or at least 5 cancer biomarkers, or more, selected from within one group in the biological sample from the subject to the reference sample and the reference panel identifies the cancer type or subtype of the subject and/or the risk of cancer progression.

[0227] A subject identified with cancer, may develop an aggressive cancer at any point in the future. Accordingly, such subjects can be selected for frequent follow up measurements of the levels of the gene transcripts of at least 4, or at least 5 cancer biomarkers as listed in Group I (151 gene set), or Group II (136 gene subgroup), or Group III (91 gene subgroup), or Group IV (51 gene subgroup), or Group V (14 gene subgroup), or Group VI (7 gene subgroup), or Group VII (6 gene subgroup) as disclosed in Tables 1-7 respectively, or Group VIII (81 gene subgroup of Group I) or Group IX (69 gene subgroup of Group I) or Group X (68 gene subgroup of Group II) or Group XI (68 gene subgroup of Group II) or Group XII (40 gene subgroup of Group III) or Group XIII (51 gene subgroup of Group III) or Group XIV (36 gene subgroup of Group IV) or Group XV (15 gene subgroup of Group IV), or Group A-D to allow aggressive treatment of cancer. Alternatively, the present invention provides methods to diagnose subjects who are at a lesser risk of developing aggressive cancer by analyzing the gene transcript levels of at least 4 or at least 5 cancer biomarkers as listed in Group I (151 gene set), or Group II (136 gene subgroup), or Group III (91 gene

subgroup), or Group IV (51 gene subgroup), or Group V (14 gene subgroup), or Group VI (7 gene subgroup) or Group VII (6 gene subgroup) as disclosed in Tables 1-7 respectively, or Group VIII (81 gene subgroup of Group I) or Group IX (69 gene subgroup of Group I) or Group X (68 gene subgroup of Group II) or Group XI (68 gene subgroup of Group II) or Group XII (40 gene subgroup of Group III) or Group XIII (51 gene subgroup of Group III) or Group XIV (36 gene subgroup of Group IV) or Group XV (15 gene subgroup of Group IV), or Group A-D to identify subjects not having or not at risk of having an aggressive cancer, which can be selected to not undergo frequent follow up measurements of the levels of the biomarkers or other alternative invasive cancer diagnostic methods.

[0228] In some embodiments, the methods to measure the expression profile of a group of cancer biomarkers, or subgroups thereof, measures the level of gene transcripts, such as mRNA expression. Methods to measure gene transcript levels are commonly known by persons of ordinary skill in the art, and are encompassed for use in the present invention, for example nucleic acid hybridization methods. In some embodiments, methods to measure gene transcript levels, for example mRNA, can use nucleic acid probes capable of hybridizing to the subject's gene/transcript sequences of the cancer biomarkers as disclosed herein. In some embodiments, methods to measure gene transcript expression can be nucleic acid probes that are immobilized on a surface, such as a nucleic acid binding chip to allow analysis, diagnosis, and prognosis, by hybridizing to the subject's gene/transcript sequences of the cancer biomarkers as disclosed herein. In some methods to measure gene transcript levels, for example serial analysis of gene expression (SAGE) or SuperSAGE is applied. In some embodiments, methods to measure gene transcript levels, can be performed by sequencing methods. Such methods include for example massive parallel sequencing (MPSS). In some embodiments, the methods to measure the expression profile of a group of cancer biomarkers, or subgroups thereof, measures the level of gene transcripts, such as mRNA expression. Methods to measure gene transcript levels are commonly known by persons of ordinary skill in the art, and are encompassed for use in the present invention, for example nucleic acid hybridization methods.

[0229] In alternative embodiments, the methods to measure the expression level of a group of biomarkers or subgroups thereof as disclosed herein can measure the expression level of a protein encoded by the cancer biomarker gene(s) as disclosed herein. Methods to measure protein expression level are commonly known by persons of ordinary skill in the art, and are encompassed for use in the present invention, for example use of antibodies targeting the proteins encoded by the cancer biomarker genes. In some embodiments, methods to measure

protein expression can use protein-binding probes, for example antibodies or protein-binding agents that are immobilized on a surface, such as a protein chip to allow analysis, diagnosis, and prognosis, by binding to the expressed protein biomarkers.

[0230] In another embodiment, the methods to measure the expression level of a group of biomarkers (or subgroups thereof) as disclosed herein are performed by analyzing the level of proteins encoded by the group of cancer biomarkers or subgroups thereof listed in Group I (151 gene set), or Group II (136 gene subgroup), or Group III (91 gene subgroup), or Group IV (51 gene subgroup), or Group V (14 gene subgroup), or Group VI (7 gene subgroup) or Group VII (6 gene subgroup) as disclosed in Tables 1-7 respectively, or Group VIII (81 gene subgroup of Group I) or Group IX (69 gene subgroup of Group I) or Group X (68 gene subgroup of Group II) or Group XI (68 gene subgroup of Group II) or Group XII (40 gene subgroup of Group III) or Group XIII (51 gene subgroup of Group III) or Group XIV (36 gene subgroup of Group IV) or Group XV (15 gene subgroup of Group IV), or Group A-D in a biological sample obtained from the subject.

[0231] In an alternative embodiment, the methods to measure the expression level of a group of cancer biomarkers (or subgroups thereof) as disclosed herein can be performed by analyzing the gene expression regulatory regions of the group of cancer biomarkers using nucleic acid polymorphisms, such as single nucleic acid polymorphisms (SNPs). Polymorphisms are known to be associated with increased and/or decreased expression of genes or polypeptides and can be used to indicate increased or decreased expression of a gene transcript in the subject. In addition, methylation patterns of the regulatory regions of the group of cancer biomarkers can be analyzed.

[0232] Provided herein in Tables 11 and 12 are lists of nucleic acid primers useful for analyzing expression of prognostic biomarkers and normalizing genes, respectively.

[0233] In another embodiment, the present invention provides methods and compositions for minimally invasive sample procurement methods for diagnosis and/or prognosis of cancer in a subject, by analyzing a group of cancer biomarkers or a subgroup thereof as disclosed herein by array-based gene expression profiling or measurement of the levels of protein encoded by the cancer biomarkers or subgroups thereof in biological sample from the subject. These methods can be used to diagnosis subjects who are already affected with cancer, such as leukemia or lymphoma, or are at high risk of developing cancer.

[0234] In some embodiments, the cancer is a lymphoid neoplasm, including diffuse large cell lymphoma (DLCL), mantle cell lymphoma (MCL), chronic lymphocytic leukemia (CLL), acute lymphoblastic leukemia (ALL), and Richter's syndrome and pediatric ALL where

t(4;11)(q21;q23) occurs. In some embodiments the cancer biomarkers are useful as prognostic and diagnostic indicators for all cancers for example, but not limited to, cancers with a high proliferation index, lymphoma, breast cancer, ovarian cancer, uterine cervical cancer, testicular germ cell tumor, glioma, neuroblastoma, astrocytoma, lung cancer and colon cancer.

[0235] In some embodiments, the cancer is a breast cancer, a colon cancer, a hyperproliferative disorder; a carcinoma, a sarcoma, a metastatic cancer, a lymphatic metastases, a blood cell malignancy, a multiple myeloma, an acute leukemia, a chronic leukemia, a head and neck cancer, a mouth cancer, a larynx cancer, a thyroid cancer, a respiratory tract cancer, a small cell carcinoma, a non-small cell cancer, a ductal carcinoma, a gastrointestinal cancer, an esophageal cancer, a stomach cancer, a colorectal cancer, a pancreatic cancer, a liver cancer, a lung cancer, a genitourinary cancer, an urologic cancer, a bladder cancer, a prostate cancer, a uterine cancer, an endometrial cancer, a kidney cancer, a renal cell carcinoma, a brain cancer, an astrocytic brain tumors, a metastatic tumor cell invasion in the central nervous system, a bone cancer, osteomas, a skin cancer, a malignant melanoma, a squamous cell carcinoma, a basal cell carcinoma, a hemangiopericytoma or a Kaposi's sarcoma or combinations thereof.

[0236] In some embodiments leukemia examples for subtypes are acute lymphocytic leukemia, chronic leukemia, myeloid leukemia, monocytic leukemia, acute myeloid leukemia, chronic myeloid leukemia, chronic lymphocytic leukemia (CLL), acute lymphoblastic leukemia (ALL), or pediatric ALL.

[0237] In some embodiments lymphoma subtypes are for example B- or T-cell neoplasms, mycosis fungoides (MF), follicular lymphoma, grade 1, diffuse large B-cell lymphoma (DLBCL), immunoblastic DLBCL, Burkitt lymphoma, marginal zone B-cell lymphoma, mantle cell lymphoma (MCL), follicular lymphoma, cutaneous T-cell lymphoma, non-Hodgkin's lymphoma (NHL), Richter's syndrome, mucosa-associated lymphoid tissue (MALT) lymphoma, Hodgkin's lymphoma, angioimmunoblastic T-cell lymphoma, nodular sclerosis type Hodgkin lymphoma, large cell diffuse lymphoma, small lymphocytic lymphoma (SLL) or nasal NK/T-cell lymphoma.

[0238] In some embodiments, subtypes of breast cancer are ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS), invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), adenocarcinoma, sarcoma or inflammatory breast cancer (IBC).

[0239] In some embodiments, subtypes of gliomas are pilocytic astrocytoma, astrocytoma grade I-IV, primary glioblastoma, malignant glioblastoma, oligoastrocytoma, oligodendroglioma or glioblastoma multiforme (GBM).

[0240] In some embodiments, the present invention further provides a group of cancer biomarkers on a microarray consisting of 4, or 5, or more of the cancer biomarkers as listed in Group I (151 gene set), or Group II (136 gene subgroup), or Group III (91 gene subgroup), or Group IV (51 gene subgroup), or Group V (14 gene subgroup), or Group VI (7 gene subgroup) or Group VII (6 gene subgroup) as disclosed in Tables 1-7 respectively, specifically intended for the diagnosis and/or prediction of cancer in a subject, or determining susceptibility of a subject to developing cancer.

[0241] In some embodiments, the present invention further provides a single biomarker, FOXM1 (SEQ ID NO: 57) or RECQL4 (SEQ ID NO: 108), specifically intended for the prediction of cancer in a subject, or determining susceptibility of a subject to developing cancer with a reduced expression level indicating a good prognosis or longer survival.

[0242] In some embodiments, the present invention relates to a method of diagnosing a cancer type or stage in a subject comprising measuring the expression profile of nucleic acid or protein biomarkers from a biological sample obtained from a subject to be diagnosed, in the expression of such cancer biomarkers in the biological sample from the subject is compared to the expression pattern of the same cancer biomarkers in a reference database, and the result will indicate a subject's likelihood of having or developing an aggressive cancer. For example, an increase in the protein and/or gene transcript expression of at least 4, or at least 5, cancer biomarkers in the biological sample from the subject as compared to the protein and/or gene transcript expression of the same cancer biomarkers in the reference sample is indicative of the subject with a likelihood of having cancer or developing an aggressive cancer as compared to a subject and/or a subject population from which the reference sample(s) was/were obtained.

[0243] In some embodiments, the present invention relates to methods of diagnosing a disease correlating to a gene copy number variation, such as nervous system disorders, metabolic diseases, infectious disease susceptibilities, autoimmune diseases and chromosome disorders.

[0244] Alternatively, the analysis of the metagene score can be used as a diagnostic or prognostic predictor. In general, a high metagene score correlates with a more aggressive cancer and/or a shorter life expectancy.

Biomarker Discovery: Identification of Disease biomarkers

[0245] Another aspect of the present invention relates to biomarker discovery, disease target discovery and disease pathway discovery. In particular, the method and algorithms as disclosed herein provide a method to identify biomarkers for diseases, such as diseases with chromosomal abnormalities (e.g., cancer). Accordingly, in some embodiments, the methods as described herein provide a method to identify genes for use as disease biomarkers, such as cancer biomarkers, on the basis of an integrative analysis framework to detect core gene expression signatures associated with copy-number alterations, which is referred to herein as Genome Association Analysis of Transcription and Comparative Oncogenomics (GAATCO).

[0246] As disclosed herein, the inventors have discovered a method to identify disease biomarkers, for example cancer biomarkers by using an integrative analysis framework to detect core gene expression signatures associated with copy-number alterations, which is referred to herein as Genome Association Analysis of Transcription and Comparative Oncogenomics (GAATCO). GAATCO is a three-step strategy. This method comprises; (1) identifying genes in regions of copy-number alterations (e. g. copy number lesions or amplifications) and the gene expression profiles in tissues from an animal model of a disease with chromosomal abnormalities, (2) correlating the copy number with the level of expression of each gene to identify a list of prognostic disease biomarkers and (3) filtering the list of disease biomarkers for genes which have orthologs in humans to identify a list of core prognostic disease biomarkers. (see Figure 7 and 10).

Step 1:

[0247] In one embodiment, the first step involves identifying genes in regions of copy-number alterations (e. g. copy number lesions or amplifications) and obtaining a gene expression profile of a biological sample from an animal model of a disease with chromosomal abnormalities.

[0248] Any method known by persons of ordinary skill to determine gene copy number can be used with the methods described herein. In one embodiment, gene copy number can be determined by, for example, whole genome hybridization approaches, comparative genomic hybridization (CGH), SNP array analysis, oligonucleotide arrays, microsatellite based analysis, bead chips (e.g. Illumina), quantitative PCR, Southern blot analysis and fluorescence in situ hybridization (FISH) or high efficiency sequencing technique, among others. In one embodiment, the array is a mouse Exonic Evidence-Based Oligonucleotide (MEEBO; Microarrays Inc.; Nashville, TN).

[0249] In some embodiments, RNA can be used for gene expression analysis. In alternative embodiments, the RNA or DNA can be obtained from any tissue, and the gene expression analysis is performed on disease tissue, for example but not limited to tumor tissue. In alternative embodiments, the DNA for copy number analysis and the RNA for gene expression analysis is obtained from cells of an in vitro model of a disease, such as cells taken from a subject with a disease and cultured in vitro, or cells genetically modified to have the disease phenotype.

[0250] Any methods and/or tools available for gene expression analysis can be used with the methods described herein. Some non-limiting examples include microarrays, e.g. Affymetrix GeneChip (Santa Clara, CA), Agilent array (Palo Alto, CA), bead technologies, e.g. Illumina bead chips (e.g. the Mouse-6 and MouseRef-8 Expression BeadChip), Luminex xMAP technology, cDNA arrays, oligonucleotide arrays, quantitative RT-PCR, SAGE, high efficiency sequencing techniques and Northern blot analysis.

[0251] Any method known by persons of ordinary skill to determine the level of gene expression can be used with the methods described herein. In some embodiments, mRNA expression can be detected on a DNA array, chip, beads, microspheres or a microarray. Methods of preparing nucleic acid arrays and their use are well known in the art. (See, for example U.S. Patent Numbers: 6,618,6796; 6,379,897; 6,664,377; 6,451,536; 548,257; U.S. 20030157485 and Schena et al. 1995 Science 20:467-470; Gerhold et al. 1999 Trends in Biochem. Sci. 24, 168-173; and Lennon et al. 2000 which are incorporated herein in their entirety by reference. Drug Discovery Today 5: 59-65, which are herein incorporated by reference in their entirety). Serial Analysis of Gene Expression (SAGE) or SuperSAGE, can also be performed (See for example U.S. Patent Application 20030215858).

[0252] In some embodiments, mRNA is extracted from the tissue sample to be tested, reverse transcribed, and fluorescent-labeled cDNA probes are generated. The microarrays capable of hybridizing to enzyme cDNA are then probed with the labeled cDNA probes, the slides scanned and fluorescence intensity measured. This intensity correlates with the hybridization intensity and expression levels.

[0253] In some embodiments, mRNA levels can be monitored by applying a cell lysate to beads, which capture the target RNAs by cooperative hybridization followed by signal amplification and detection.

[0254] Methods of "quantitative" amplification are well known to those of skill in the art, and include methods such as e.g., quantitative PCR. Detailed protocols for quantitative PCR are provided, for example, in Innis et al. (1990) PCR Protocols, A Guide to Methods and

Applications, Academic Press, Inc. N.Y.. Some exemplary primers useful for quantitative PCR can be found in Tables 11 and 12 provided herein to determine expression of prognostic biomarkers and normalizing genes, respectively.

[0255] Any suitable method for detecting mRNA expression levels in a sample can be used in connection with the methods described herein. For example, mRNA expression levels in a sample can be determined by generation of a library of expressed sequence tags (ESTs) from a sample. Enumeration of the relative representation of ESTs within the library can be used to approximate the relative representation of a gene transcript within the starting sample. The results of EST analysis of a test sample can then be compared to EST analysis of a reference sample to determine the relative expression levels of a selected polynucleotide, particularly a polynucleotide corresponding to one or more of the differentially expressed genes described herein.

[0256] In some embodiments, hybridization to arrays can be performed, where the arrays can be produced according to any suitable methods known in the art. For example, methods of producing large arrays of oligonucleotides are described in U.S. Pat. No. 5,134,854, and U.S. Pat. No. 5,445,934 using light-directed synthesis techniques, which are incorporated herein in their entirety by reference. Alternatively, microarrays are generated by deposition of pre-synthesized oligonucleotides onto a solid substrate, for example as described in International Application No. WO 95/355 05 which are incorporated herein in their entirety by reference. Methods for collection of data from hybridization of samples with an array are also well known in the art. For example, the polynucleotides of the cell samples can be generated using a detectable fluorescent label, and hybridization of the polynucleotides in the samples detected by scanning the microarrays for the presence of the detectable label. Methods and devices for detecting fluorescently marked targets on devices are known in the art.

Step 2:

[0257] In some embodiments, one can calculate gene-to-gene correlation between all pairs of genes in the signature for each dataset. In some embodiments, the correlation matrices are combined across species to a single conservation matrix (Figure 2 B1). In some embodiments, the inventors applied hierarchical clustering method to separated genes showing high conservation from the rest (Lee et al., 2004) (Figure 2 B2).

[0258] Without wishing to be bound by theory, it is known that copy number alterations exert phenotypic effects via transcriptional alterations, therefore in one embodiment, lesions containing genes with copy-number dependent expression changes can be selected, i.e.

correlated copy-number and expression levels of the genes within the locus of the copy number changes.

[0259] In some embodiments, gene signatures comprising genes, which are outside the lesions are selected and their expression level is correlated to their copy-number levels using statistical tools, e.g. ANOVA. Identifying genes outside the lesion can be complicated by linked genetic events on different genomic loci. For example, if two loci on different chromosomes are juxtaposed by translocation and amplified, they would show the same copy number alteration and should be considered de facto as one lesion although they are normally on different chromosomes, for example see Rooney et al., 2004b and Zhu et al., 2002. Excluding genes based on their genomic locations will not be effective in cases like this. Accordingly, the present invention relates to a statistical method to circumvent this problem, by using a linear model framework to first account for a copy-number to copy-number correlation before testing for copy-number to expression correlation. This strategy effectively accounts for linked copy number lesions within or across chromosomes and identify genes whose expression levels are correlated to copy number lesions.

GAATCO analysis.

[0260] In some embodiments, the correlation between gene copy number and the level of gene expression can be determined for each gene. In some embodiments, one can determine the copy number-gene expression correlation using the following formula and using a least square estimate of β_{jj} -the slope of expression level of gene j over DNA copy numbers of gene j :

$$H_0 : T_j = \mu_j + \epsilon_j \quad (1)$$

$$H_a : T_j = \mu_j + \beta_{jj} DC_j + \epsilon_j \quad (2)$$

[0261] In some embodiments, for all possible pairs between copy-number (CNA) and genes, one can test for CNA to expression (gene) association using the following linear models:

$$T_j = \beta_0 + \epsilon_j \quad (1)$$

$$T_j = \beta_0 + \beta_1 D_j + \epsilon_j \quad (2)$$

$$T_j = \beta_0 + \beta_2 C_j + \epsilon_j \quad (3)$$

$$T_j = \beta_0 + \beta_1 D_j + \beta_2 C_j + \epsilon_j \quad (4)$$

,where T_j denotes expression, D_j is copy-number, and ϵ_j is random error of the gene in sample j . C_j denotes copy-number of the CNA in sample j . β_1 and β_2 denote regression parameters

measuring local and global association, respectively. β_0 is an intercept regression parameter. One can use a forward, step-wise model selection strategy to test for statistical significance of β_2 . First, statistical significance of β_1 are tested by comparing model 1 and 2. Second, based on the model selected from the first step, statistical significance of β_2 , the parameter of interest, are tested either by comparing model 1 and 3 or by comparing model 2 and 4. Resulting p-values can be adjusted for multiple-testing by false discovery rate calculation, separately for each CNA (Storey and Tibshirani, 2003).

[0262] Statistical analyses can be performed by any means known by persons of ordinary skill in the art.

Step 3:

[0263] One can perform cross-species comparison of genes using any available method commonly known by persons of ordinary skill in the art. In some embodiments, the prognostic disease biomarker gene set is filtered to a “core prognostic disease biomarker” gene set by comparative analysis of mouse and human transcriptome. Cross-species conservation is a powerful biologic filter to detect functionally important genes (Stuart et al., 2003).

[0264] In some embodiments, gene ontology (GO) analysis can be performed to identify altered biological pathway(s) associated with gene signatures. In some embodiments, enrichment of GO biological processes of the prognostic disease biomarkers can be compared against a set of known NCBI Genes in human microarray data which is available to the public. In some embodiments, hyper-geometric tests can be used to assess statistical significance of enrichment using R/GOSTATS. In one embodiment, the gene ontology analysis is valuable for identifying novel targets and disease pathways.

[0265] In other embodiments, one can map probe sequences of the prognostic disease biomarker genes using a BLAT server (Kent, 2002) and filtered those with unambiguous mappings. Names, descriptions, and Entrez identifiers of known genes were imported from NCBI public database (<ftp://ftp.ncbi.nlm.nih.gov>).

[0266] In some embodiments, orthologous genes between *Mus musculus* and *Homo sapiens* can be identified using NCBI HomoloGene database. In some embodiments, mapping to Entrez GeneIDs for Affymetrix Genechips and Gene Ontology annotations for genes can also be obtained from the bioconductor repository, which is publicly available at world-wide-web-dot-bioconductor-org.

[0267] In some embodiments, orthologous genes between the animal model and the target subject can be identified using NCBI HomoloGene database. In some embodiments, orthologous genes from the animal disease model to other animals can be identified using any method known by persons of ordinary skill in the art, for example using canine or feline or other animal gene databases, so that orthologous genes in dogs and cats and other domestic animals can be identified.

[0268] In some embodiments, prognostic diseases biomarker genes selected as core prognostic diseases biomarker genes are those which show at least a one-to-one orthologous relationship according to NCBI HomoloGene database, for example at least one human or other animal ortholog is identified per animal disease model gene.

Validation of Biomarkers by Comparative Analysis with human survival data

[0269] As an illustrative example the GAATCO method of biomarker identification was applied to a mouse model of B lymphoma (e.g., the Art Trp53 mice as disclosed in Example 8). In the copy number analysis 19 genetic lesions were identified. In step two, the correlation of copy number and gene expression data from these samples using the regression analysis as disclosed herein identified 151 genes (Group I) which are referred to herein as “prognosis cancer biomarkers”, whereas using genome-wide based analysis (see Figure 12) 464 genes were identified. These prognostic cancer biomarkers were used in a comparative cross-species gene analysis to compare the mouse genes to human orthologs, which resulted in 51 genes, which are referred to as the “core prognostic cancer biomarkers”(Figure 2 C) and in 136 genes (Figure 12C). The 151 genes (shown in Table 1, Group I) comprise subsets of genes: 136 (Group II), 91 (Group III), 51 (Group IV), 14 (Group V), 7 (Group VI) and 6 genes (Group VII).

[0270] To evaluate the performance of the comparative filtering, the inventors tested whether the gene set become functionally enriched after the filtering. The inventors used association with cancer prognosis as a surrogate measure for the functional importance of a gene. For every gene and for every human cohort, the inventors tested for correlation between its expression level and survival outcomes by univariate cox regression analysis and found that 28 % of all genes show significant correlation to survival outcomes in multiple cohorts (≥ 3). While the unfiltered signature set ($n=185$) is moderately enriched in prognostic genes (39%, $p=0.019$), the inventors observed drastic enrichment in the core set (66 %, $p_1 \leq 5$) (Figure 2). To evaluate enrichments of specific pathways, the inventors also tested for enrichment of genes annotated with Gene Ontology (GO) biological processes. Figure 2B shows three

representative examples, where the inventors observed similar trends of moderate and drastic enrichment before and after filtering (Figure 2C). Since many cell-cycle related genes are critical determinants of cancer prognosis (Whitfield et al., 2006), the inventors tested whether the enrichment of prognostic genes was simply due to enrichment of cell-cycle related genes. The inventors repeated the enrichment analysis of prognostic genes using a subset of genes annotated with the GO terms. The results indicate that not all cell-cycle genes are prognostic genes (i.e. proportion of prognostic genes are less than 1), but the inventors observe a significant increase in the proportion with the core prognostic cancer biomarkers set. This demonstrates that comparative analysis is an unbiased way to enrich for functionally significant genes.

[0271] The B cell interaction network data provide independent evidence that FOXM1 is a highly connected “hub” gene in the pathway represented by the 136 signature. The inventors have shown that FOXM1 acts as a regulator of the 136 gene signature. The transcriptional changes of the 136 genes associated with the 6C1-G3 deletions in the mouse lymphomas were in part mediated by down-regulation of FOXM1. It is known that FOXM1 is a cell-cycle regulator (Laoukili et al., 2005, 2008a,b; Schuller et al., 2007; Zhang et al., 2008) and has been found upregulated in a number of cancers, e.g. glioblastoma, primary breast cancer, lung cancer and prostate cancer (Radhakrishnan and Gartel, 2008, *Nature Reviews Cancer*, Vol. 8, p242; Myatt and Lam, 2008, *Nature Reviews Cancer*, Vol.7, 847-859). FoxM1 has been reported to be prognostic for certain breast cancers (Bektas et al. 2008, *BMC Cancer*, Vol 8, p. 42), cervical cancer (Chiu et al. 2005, *Proc Amer Assoc Cancer Res*, Volume 46) and glioma (Zhang et al. 2008, *Cancer Research* 68, 8733-8742). For pulmonary squamous cell carcinoma FoxM1 has been described as a poor prognostic marker (Yang, D.K., Son, C.H., Lee, S.K., Choi, P.J., Lee, K.E., Roh, M.S., *Human Pathology* 2009, in press). However, this is the first time that FoxM1 has been shown to regulate the expression of a gene signature, the 136-gene signature and being a general prognostic marker for cancer. Furthermore, these findings indicate that a reduced expression level of FOXM1 is significantly associated with tumor-free survival in multiple human cancer cohort, and thus is a good prognostic marker for longer tumor-free survival (Table 10).

[0272] RECQL4 belongs to the RecQ family of DNA helicases has been associated with genome instability and certain human genetic disorders, Bloom syndrome (BS), Werner syndrome (WS), and Rothmund–Thompson syndrome (RTS) (Harrian and Bohr 2003, *Biochimie*, Vol. 85, Issue 11, p 1185-1193). The inventors show that the expression level of RECQL4 is significantly associated with tumor-free survival in multiple human cancer

cohorts and that reduced RECQL4 level is a good prognostic marker for longer tumor-free survival (Table 10).

Animal models of cancer

[0273] One aspect of the present invention provides a method to identify cancer biomarkers using the methods for biomarker discovery as disclosed herein. In some embodiments, the methods disclosed herein are used to identify cancer biomarkers using an animal model with chromosomal abnormalities, e.g. translocations or gene copy number changes where the animal develops tumors. Tumors from the animal model with chromosomal abnormalities can be used for gene expression profiling. Such chromosomal abnormalities can be complicons. Complicons arise as a result of amplification of large regions of the genome due to complex translocations. Complicons can arise when the DNA repair mechanism is impaired and broken chromosomes persist and recombine inappropriately with other chromosomes.

[0274] In some embodiments, the animal model is a mouse model, for example a mouse carrying one or more defects for example in the nonhomologous end joining (NHEJ) region, in DNA repair genes, DNA recombinase genes, or in cell cycle genes etc. (see Table 8). In some embodiments, animal models can arise by spontaneous mutation or can be created by ENU mutagenesis or specific genetic modification, like gene targeting. In other embodiments, animal models useful in the methods as disclosed herein include zebrafish, medaka, Drosophila and rat, among others. The usefulness of such alternative animal models for cancer research is well established in the art (Grabher and Look, 2006; M. Scott Lucia, 1998; Taniguchi et al., 2006; Vidal and Cagan, 2006).

Table 8: Table of mouse models of cancer with chromosomal abnormalities. An overview of various mouse models is described in (Friedberg and Meira, 2004)

Gene (loss of function)	Reference:
53BP1	(Morales et al., 2006; Ward et al., 2004)
53BP1 Trp53	(Morales et al., 2006; Ward et al., 2004)
Art (Artemis)	(Rooney et al., 2002)
Art Trp53	(Woo et al., 2007)
Atm	(Xu et al., 1996)
brca1 associated ring domain 1 (Bard1)	(McCarthy et al., 2003)
Brcal	(Liu et al., 2007; Mak et al., 2000; Snouwaert et al., 1999)

Gene (loss of function)	Reference:
Brca1/p53	(Liu et al., 2007)
Brca2	(Bennett et al., 2000; Evers and Jonkers, 2006)i
Cernunnos/XLF	(Zha et al., 2007)
Chaos3 (Mcm4)	(Shima et al., 2007)
CHEK2*1100delC	(Bahassi et al., 2007)
DNA Ligase I Gene1	(Friedberg and Meira, 2004)
DNA-PKcs (classical SCID)	(Bosma et al., 1988; Gao et al., 1998a; Kurimasa et al., 1999; Taccioli et al., 1998)
DNA-PKcs SCID mutations, Trp53	(Gladdy et al., 2003)
DNA-PKcs SCID, irradiated	(Williams et al., 2001)
DNA-PKcs Trp53	(Gladdy et al., 2003)
Exo1	(Wei, 2003)
Fanconi Anemia	(Houghtaling et al., 2003; Noll et al., 2002; Wong and Buchwald, 2002)
Gadd45a	(Hollander et al., 1999)
H2ax Trp53	(Celeste et al., 2003)
Ku70	(Gu et al., 1997; Ouyang et al., 1997)
Ku70 Trp53	
Ku80	(Nussenzweig et al., 1996)
Ku80 Trp53	
lig4 (DNA ligase IV)	(Barnes et al., 1998)
lig4p53 (DNA ligase IV;Trp53)	(Frank et al., 2000)
Msh2	http://www.phenogenomics.ca/databases/non_enu.html
Msh2/Apc(+/-)	(Friedberg and Meira, 2004)
Msh2/Trp53	(Friedberg and Meira, 2004)
mTerc Trp53	(Friedberg and Meira, 2004)
Mus81	(2003; Artandi and DePinho, 2000; Niida et al., 1998)
Mus81/Trp53	(Pamidi et al., 2007)
p19 (Arf)	(Thornton et al., 2005)
p27(kip1)-/-	(Payne and Kemp, 2003; Payne et al., 2008)
p27(kip1)+/-	(Payne and Kemp, 2003; Payne et al., 2008)
Pms2	(Friedberg and Meira, 2004)
Pms2/Apc(+/-)	(Friedberg and Meira, 2004)
Polbeta-/DNA-PKcs/scidp53-/- triple-mutant mice	(Niimi et al., 2005)
Rad50	(Luo et al., 1999)
Rad51	(Lim and Hasty, 1996)

Gene (loss of function)	Reference:
Rb conditional KO (liver)	(Mayhew et al., 2005; Mayhew et al., 2007)
TdT	(Gilfillan et al., 1993)
Trp53	(Donehower et al., 1992; Jacks et al., 1994)
Xpa	(Friedberg and Meira, 2004)
Xpa/Trp(+/-)	(Friedberg and Meira, 2004)
Xrcc1	(Friedberg and Meira, 2004)
Xrcc2	(Friedberg and Meira, 2004)
Xrcc4	(Gao et al., 1998b)
Xrcc4 conditional, Trp53	(Yan et al., 2006)
Xrcc4 transgenic mice crossed to Xrcc4 null crossed to CD21-Cre, Trp53	(Soulas-Sprauel et al., 2007)
Gene (gain of function)	Reference:
Eg5 overexpression	(Castillo et al., 2007)
AID overexpression	(Okazaki et al., 2003)
Balb/c mice treated with pristane	

[0275] Many mouse models of cancer show loss of the tumor suppressor p53, amplification of the c-myc proto-oncogene, loss of heterozygosity for Atm (which encodes an important DNA damage responsive factor), and gene expression profiles that associate del(11q23) with aberrations in multiple cell cycle control and genome stability pathways (Tsimberidou and Keating, 2005; Yee et al., 2005). DNA repair defective mice display immunologic and developmental defects. Mice deficient for each of the six known factors involved in the nonhomologous end joining (NHEJ) pathway of DNA double strand break (DSB) repair, which is critical for repair of general DSBs and for V(D)J recombination, have been isolated or engineered (Mills et al., 2003; Rooney et al., 2004a). The overall severity of the phenotype reflects the degree to which DSB repair and/or V(D)J recombination are compromised, with Artemis at the mild end of the spectrum and Lig4 at the severe end of the spectrum. Deficiency for Artemis (Art), encoding a DNA end processing factor important for preparing modified or irregular DNA ends for ligation, results in a variable SCID (severe combined immunodeficiency) phenotype with leaky T cell development, and an increase in sensitivity to ionizing radiation (Rooney et al., 2003; Rooney et al., 2004a). Deficiency for Lig4, encoding the essential DNA Ligase IV required for the final step of NHEJ, results in late

embryonic lethality, extensive neuronal apoptosis in the embryonic brain, a complete block in V(D)J recombination, profound cellular sensitivity to ionizing radiation, and spontaneous chromosomal instability (Bassing et al., 2002; Mills et al., 2003).

[0276] NHEJ/Trp53 doubly deficient mice develop B-lymphoid malignancies with copy number abnormalities in 4q21 and 11q23 conserved syntenic regions. Viability in *Lig4* null mice can be rescued by additional deficiency for the tumor suppressor p53, encoded by the *Trp53* gene (Frank et al., 2000; Gao et al., 2000). For simplicity these double null mice are referred to hereafter as *Lig4 Trp53*. Lack of p53 prevents apoptosis in developing neurons and *Lig4 Trp53* mice are born, but lack mature lymphocytes while retaining an enlarged pro-B cell compartment. By approximately 12 weeks after birth *Lig4 Trp53* mice succumb to aggressive progenitor B cell lymphomas with characteristic cytogenetic features (Frank et al., 2000; Gao et al., 2000; Zhu et al., 2002). These include a nonreciprocal der(12)t(12;15) translocation product; complex, multi-chromosome translocations with co-amplified *Igh* and *c-myc* which the inventors have termed complicons; and frequent copy number 'aberrations on chromosomes 5 and 9 (Figure 1). Like *Lig4 Trp53* mice, *Art Trp53* mice, rapidly develop pro-B cell malignancies with characteristic cytogenetic features, including complicons, but can incur either *c-myc* or *N-myc* amplification.

A mouse model for Richter's syndrome

[0277] While one clear oncogenic event in these tumors is thought to be hyperactivation of the *c-myc* allele, the recurrent nature of cytogenetic lesions on chromosomes 5 and 9 is indicative of functional roles for these, as well. Because these tumors arise in a p53-deficient setting, invariably incur *Myc*-family oncogene activation via gene amplification, and frequently exhibit deletions on chromosome 9 in a region of conserved synteny with human chromosome 11q23, *NHEJ/Trp53* mice represent an excellent molecular genetic model for Richter's syndrome, and for other human malignancies with CNA-type lesions at 11q23 and/or 4q21.

[0278] By way of a non-limiting example and as disclosed herein, the inventors have demonstrated use of a NHEJ/Trp53 mouse model, which develops lymphoid cancers with CNA-type lesions. In some embodiments, two mouse cancer models are used herein, and include (i) mice heterozygous for the Ligase IV gene, a member of the NHEJ gene family, and deficient for the tumor suppressor p53 (encoded by TP53 in humans and *Trp53* in mice) (Frank et al., 2000) and (ii) mice deficient for artemis (*Art*), a member of the NHEJ gene family, and *Trp53* (Rooney et al., 2004b). Both mouse models develop B cell lymphomas.

[0279] B cell lymphomas from these mouse models have been harvested and processed for DNA and RNA isolation using standard molecular biology techniques as described for example in *Molecular Cloning: A Laboratory Manual* (3rd edition; 2001); by Joseph Sambrook, Peter MacCallum and David Russell.

[0280] As disclosed herein, the inventors tested for the association of the core prognostic cancer biomarkers with overall clinical outcomes. In addition to one mouse B lymphoma, the inventors obtained nine datasets representing human cohorts of two B lymphoma, four breast cancers, three gliomas, and two lung cancers. For each cohort, the inventors derived a meta-gene score for each sample by a singular value decomposition technique. A meta-gene score comprising a single value represents the expression of 136 genes for each sample. The inventors performed cox regression analysis to test for their association with survival outcomes and found them to be significant in all ten mouse and human cohorts. The inventors then stratified them into two halves (i.e., low- and high-risk groups) based on their metagene scores (Figure 4 LEFT PANEL). The Kaplan-Meier analysis show significant risk-stratification in 8 out of 10 cohorts (Figure 4).

[0281] Examination of a prognostic cancer biomarker gene set as disclosed in Table 1 revealed several pathways important in tumor initiation and progression represented by the genes. Genes positively correlated with aggressiveness of tumor were predominantly mitotic cell cycle and DNA replication related: cyclins (CCNA2, CCNB2), DNA helicases (RECQL4, DNA2). Genes negatively correlated with tumor aggressiveness include transcriptional regulators (HBP1, SQSTM1), cell adhesion (ITGB2), apoptosis (HCK, etc). These data show that the inventors discovery and identification of the gene set represents a set of multiple pathways contributing to tumorigenesis.

Other uses for Cancer biomarkers

[0282] In view of the currently limited options for diagnostic and prognostic of cancers, the core prognostic cancer biomarkers or subgroups thereof as disclosed herein are useful for diagnosing cancer types and/or stages and for prognosis. In some embodiments, a subject diagnosed with cancer can be administered a therapeutic regimen to eliminate the cancer. In some embodiments, the core prognostic cancer biomarkers or subgroups thereof as disclosed herein are useful for identifying subjects with a poor-prognosis. Accordingly, subjects identified with a poor prognosis can be administered therapy, for example systematic therapy. In some embodiments, a subject identified to have a poor cancer prognosis can be administered a more aggressive cancer treatment regimen, for example, multiple anti-cancer

therapies simultaneously, such as, but not limited to administration of anti-cancer agents and radiotherapy or surgical resection. In the case of breast cancer, mastectomy is considered an aggressive treatment, which can be further combined with radiotherapy and/or chemotherapy and/or chemotherapeutic combinations. For other cancers, high-dose radiotherapy in combination with macroscopic total resection or chemotherapy are considered aggressive, or high doses of chemotherapy drugs combined with autologous stem cell transplantation may be considered as aggressive treatment. Aggressive (or non-aggressive) treatments appropriate for a specific cancer type are well known to those of skill in the art and therapy can be adjusted during treatment using the prognostic indicators and methods described herein.

[0283] In some embodiments, the compositions and methods as disclosed herein can also be used to identify subjects in need of frequent follow-up by a physician or clinician to monitor the cancer and risk of relapse, as well as cancer progression. For example, if a subject is diagnosed with cancer comprising a poor cancer prognosis, the subject can initiate treatment earlier when the disease is potentially more sensitive to treatment, or the subject can initiate a treatment specifically aimed at eliminating the cancer.

[0284] In further embodiments, the methods and compositions as disclosed herein are useful for identifying subjects with cancer expressing at least 5 core prognostic cancer biomarkers or subgroups thereof, which is useful to identify subjects most suitable or amenable to be enrolled in clinical trial for assessing a therapy specifically aimed at eliminating the cancer cells. Such an embodiment will permit more effective subgroup analyses and follow-up studies. Furthermore, the expression of the group of core prognostic cancer biomarkers as disclosed herein can be used to monitor such subjects enrolled in a clinical trial to provide a quantitative measure for the therapeutic efficacy of an anti-cancer therapy aimed at eliminating the cancer in a subject.

[0285] In further embodiments, the core prognostic cancer biomarkers as identified by the methods disclosed herein can be used to prepare antibodies that are specific markers of cancer cells disclosed herein. Polyclonal antibodies can be prepared by injecting a vertebrate animal with cells of this invention in an immunogenic form. Production of monoclonal antibodies is described in such standard references as U.S. Pat. Nos. 4,491,632, 4,472,500 and 4,444,887 which are incorporated herein by reference, and Methods in Enzymology 73B:3 (1981). Specific antibody molecules can also be produced by contacting a library of immunocompetent cells or viral particles with the target antigen, and growing out positively selected clones. See Marks et al., New Eng. J. Med. 335:730, 1996, and McGuinness et al., Nature Biotechnol. 14:1449, 1996. A further alternative is reassembly of random DNA

fragments into antibody encoding regions, as described in EP patent application 1,094,108 A, which is incorporated herein by reference. A further alternative are alternative binding proteins as described in Stumpp and Amstutz, *Current Opinion in Drug Discovery and Development* 2007, 10:153-159; Hey et al. *Trends in Biotechnology*, Volume 23, Issue 10, October 2005, Pages 514-522; Woodnutt et al. *Current Opinion in Drug Discovery and Development* 2008, 11:754-761.

[0286] The antibodies in turn can be used as diagnostic applications to identify a subject with cancer comprising fast proliferating cancer, or alternatively, antibodies can be used as therapeutic agents to prevent cell proliferation and/or kill cancer cells.

[0287] In further embodiments, the core prognostic cancer biomarkers as identified by the methods disclosed herein can be used to screen for agents or small molecules, which can normalize the gene expression of at least one of the 151 cancer biomarkers as disclosed in Table 1 to a gene expression level closer (i.e. at least 10% closer) to that of a reference sample.

[0288] In further embodiments, the core prognostic biomarkers as identified by the methods disclosed herein can be used to identify novel targets for diseases.

[0289] In further embodiments, the core prognostic biomarkers as identified by the methods disclosed herein can be used to identify disease pathways and develop research tools for furthering the understanding of disease mechanisms.

[0133] In some embodiments, the methods disclosed herein are useful to identify prognostic disease biomarkers for all diseases with chromosomal abnormalities or disease with copy number variations.

Kits

[0290] In some embodiments, the present invention provides a diagnostic kit(s) for determining the prognosis of a subject having cancer by gene expression analysis of at least 5 gene transcripts of the core prognostic cancer biomarkers as listed in Group I (151 gene set) disclosed in Table 1. In some embodiments, the methods use probes or primers comprising nucleotide sequences, which bind under high stringency or moderately stringent conditions to the different nucleic acid sequences selected from the group of genes encoded by nucleic acid sequences selected from the list of 151 genes (Group I) as shown in Table 1

[0291] In some embodiments, the present invention provides diagnostic kits for determining the prognosis of a subject having cancer by gene expression analysis of at least 5 gene transcripts of cancer biomarkers as listed in Group II (136 gene set) as disclosed in Table 2, or

alternatively Group X (positive gene subset from Group II) or Group XI (negative gene subset from Group II), where the methods use probes or primers comprising nucleotide sequences, which bind under high stringency or moderately stringent conditions to the different nucleic acid sequences selected from the group of genes encoded by nucleic acid sequences selected from the list of 136 genes listed in Group II, X or XI. Alternative sequencing methods, for example SAGE or SuperSAGE can be used to determine gene expression levels.

[0292] In another embodiment, the present invention provides diagnostic kits for determining the prognosis of a subject having cancer by gene expression analysis of at least 5 gene transcripts of cancer biomarkers as listed in Group III (91 gene set) as disclosed in Table 3, or alternatively Group XII (positive gene subset from Group III) or Group XIII (negative gene subset from Group III), where the methods use probes or primers comprising nucleotide sequences, which bind under high stringency or moderately stringent conditions to the different nucleic acid sequences selected from the group of genes encoded by nucleic acid sequences selected from the list of 91 genes listed in Group III, XII or XIII.

[0293] In some embodiments, the present invention provides diagnostic kits for determining the prognosis of a subject having cancer by gene expression analysis of at least 5 gene transcripts of cancer biomarkers as listed in Group IV (51 gene set) as disclosed in Table 4, or alternatively Group XIV (positive gene subset from Group IV) or Group XV (negative gene subset from Group IV), where the methods use probes or primers comprising nucleotide sequences, which bind under high stringency or moderately stringent conditions to the different nucleic acid sequences selected from the group of genes encoded by nucleic acid sequences selected from the list of 51 genes listed in Group IV, XIV or XV.

[0294] In some embodiments, the present invention provides diagnostic kits for determining the prognosis of a subject having cancer by gene expression analysis of at least 3, or at least 4, or at least about 5 gene transcripts of cancer biomarkers as listed in Group V (14 gene set) as disclosed in Table 5, where one of genes is RECQL4 (SEQ ID NO: 108), and where the methods use a probe or a primer comprising a nucleotide sequence which binds under high stringency or moderately stringent conditions to the RECQL4 gene (encoded by SEQ ID NO: 108) and also uses probes or primers comprising nucleotide sequences which bind under high stringency or moderately stringent conditions to at least 3, or at least 4, or at least about 5 different nucleic acid sequences selected from the group of genes encoded by nucleic acid sequences selected from the list of 13 genes listed in Group V.

[0295] In some embodiments, the present invention provides diagnostic kits for determining the prognosis of a subject having cancer by gene expression analysis of at least 3, or at least 4 or at least about 5 gene transcripts of cancer biomarkers as listed in Group VI (7 gene set) as disclosed in Table 6, and where the methods use a probe or a primer comprising a nucleotide sequence which bind under high stringency or moderately stringent conditions to at least 3, or at least 4, or at least about 5 different nucleic acid sequences selected from the group of genes encoded by nucleic acid sequences selected from the list of 7 genes listed in Group VI.

[0296] In some embodiments, the present invention provides diagnostic kits for determining the prognosis of a subject having cancer by gene expression analysis of at least 3, or at least 4, or at least about 5 gene transcripts of cancer biomarkers as listed in Group VII (6 gene set) as disclosed in Table 7, and where the methods use a probe or a primer comprising a nucleotide sequence which bind under high stringency or moderately stringent conditions to at least 3, or at least 4, or at least about 5 different nucleic acid sequences selected from the group of genes encoded by nucleic acid sequences selected from the list of 6 genes listed in Group VII.

[0297] The kit can comprise at least 3, or at least 4, or at least 5 probes, and/or 6 primer-pairs which are capable of specifically hybridizing to at least 3, or at least 4, or at least 5 biomarkers selected from the group of core prognostic cancer biomarkers from Group I (151 gene set) as disclosed in Table 1 and instructions for use. Preferred kits amplify all or a portion of at least 5 gene transcripts selected from the group of core prognostic cancer biomarkers of Group I (151 gene set) as disclosed in Table 1. Such kits are suitable for detection of level of transcript expression by, for example, fluorescence detection, by electrochemical detection, by radioactive detection, by enzymatic detection or by other detection. Such kits can be prepared for example as multiplex beads (e.g. Luminex) or as arrays. Alternatively, such kits can be a SAGE or SuperSAGE kit.

[0298] The kit can comprise at least 3, or at least 4, or at least 5 probes and/or 6 primer-pairs which are capable of specifically hybridizing to at least 3, or at least 4, or at least, 5 genes selected from the group of core prognostic cancer biomarkers from Group I or Group II (136 gene subset or Group III (91 gene subset) or Group IV (51 gene subset) or Group V (14 gene subset) or Group VI (7 gene subgroup) or Group VII (6 gene subgroup) as disclosed in Tables 1-7 respectively and instructions for use. In an alternative embodiment, a kit of the present invention can comprise at least 3, or at least 4 or at least about 5 probes or 6 primer-pairs which are capable of specifically hybridizing to at least 3, or at least 4 or at least about 5 genes selected from the group of cancer biomarkers listed in Group VI (positive gene subset from Group II), or Group VII (negative gene subset from Group II), or Group X (positive

gene subset from Group III), or Group XI (negative gene subset from Group III), or Group XII (positive gene subset from Group IV) or Group XIII (negative gene subset from Group IV) and instructions for use.

[0299] In some embodiments, a preferred kit can amplify all or a portion of RECQL4 (SEQ ID NO: 108), and optionally and at least 1, or at least 2, or at least 3, or at least 4, or at least 5, or at least 6, or at least 7, or at least 8 or at least 9, or at least 10, 11, 12 or at least 13 gene transcripts selected from the group of genes listed in Group V (14 gene subset) as disclosed in Table 5. In some embodiments, probes or primer-pairs which are capable of specifically hybridizing to additional genes not listed in Group I can also be included in the kits. Such probes or primer-pairs typically specifically hybridize to genes for purposes of quality control (i.e. to assess the quality of the nucleic acid sample (RNA or DNA) sample, or alternatively serves as positive or negative controls. In one embodiment, the probes/primers recognize additional cancer biomarkers and serve to increase the sensitivity of the kit in identification of cancer in a subject. Any probe or primer-pair for a known cancer biomarker can be included in the kits of the present invention.

[0300] Such kits are suitable for detection of level of transcript expression by, for example, fluorescence detection, by electrochemical detection, by radioactive detection, by enzymatic detection or by other detection. Oligonucleotides, whether used as probes or primers, contained in a kit can be detectably labeled. Labels can be detected either directly (e.g., fluorescent labels), or indirectly. Indirect detection can include any detection method known to one of skill in the art, such as biotin-avidin interactions, antibody binding and the like. Fluorescently labeled oligonucleotides can also contain a quenching molecule. Alternative sequencing methods may be included in the kit, for example SAGE or SuperSAGE, to determine gene expression levels.

[0301] The kit can further comprise at least one reagent necessary to perform the assay. For example, the kit can comprise an enzyme. Alternatively the kit can comprise a buffer or any other necessary reagent.

[0302] Conditions for incubating a nucleic acid probe with a biological sample depend on the format employed in the assay, the detection methods used, and the type and nature of the nucleic acid probe used in the assay. One skilled in the art will recognize that any one of the commonly available hybridization, amplification or immunological assay formats can readily be adapted to employ the nucleic acid probes for use in the present invention.

[0303] In alternative embodiments, the present invention provides diagnostic kits for determining the prognosis of a subject having or developing cancer by protein expression

analysis of at least 5 proteins encoded by the cancer biomarkers which are listed in any of the Groups I-XVII. In some embodiments, the present invention provides diagnostic methods for detecting cancer in a subject by protein expression analysis of at least 5 proteins encoded by the cancer biomarkers as listed in Group I (and disclosed herein in Table 1) using a protein-binding assay, such as for example, but not limited to, ELISA or peptide-chip arrays, or antibody arrays. Any means commonly known to persons of ordinary skill in the art can be used to determine protein expression. A high-throughput format can be used to determine the protein expression of at least 5 proteins listed in Group I (Table 1). In some embodiments, such a protein-binding assay can detect at least 5, or about at least 10, or about at least 20, or about at least 30, or about at least 40, or about at least 50, or about at least 100 or about at least 150 or any number of genes between 5 and 151 listed in Group I (Table 1). In some embodiments, such a protein-binding assay can detect any number of genes listed in Group I (Table 1) between 5-151, and can further comprise an additional at least 10, or at least about 20, or at least about 30, or at least about 40, or at least about 50, or at least about 60, or at least about 70, or at least about 80, or at least about 90, or at least about 100 or more additional genes not listed in Group I (Table 1).

[0304] In some embodiments, the biological samples that can be used with the diagnostic kits include cells, protein or membrane extracts of cells, or biological fluids such as sputum, blood, serum, plasma, or urine. The biological samples used in the above described method will vary based on the assay format, nature of the detection method and the tissues, cells or extracts used as the sample to be assayed. Methods for preparing protein extracts or membrane extracts of cells are known in the art and can be readily adapted in order to obtain a sample which is compatible with the system utilized.

[0305] The kits can also include all or some of the reference biological samples as well as positive and negative controls, reagents, primers, sequencing markers, probes and antibodies described herein for determining the protein and/or gene transcript expression level of at least 5 core prognostic cancer biomarkers as disclosed herein, in order to determine a subject's likelihood of having or being at risk of having or developing cancer.

[0306] As amenable, these kit components can be packaged in a manner customary for use by those of skill in the art. For example, these suggested kit components can be provided in solution or as a liquid dispersion or the like.

[0307] The invention also provides diagnostic and experimental kits which include antibodies for determining the protein expression level encoded by at least 5 core prognostic cancer biomarkers as disclosed herein, in order to determine a subject's likelihood of having or being

at risk of developing cancer. In such kits, the antibodies can be provided with means for binding to detectable marker moieties or substrate surfaces. Alternatively, the kits can include the antibodies already bound to marker moieties or substrates. The kits can further include reference biological samples as well as positive and/or negative control reagents as well as other reagents for adapting the use of the antibodies to particular experimental and/or diagnostic techniques as desired. The kits can be prepared for in vivo or in vitro use, and can be particularly adapted for performance of any of the methods of the invention, such as ELISA. For example, kits containing antibody bound to multi-well microtiter plates can be manufactured.

[0308] In some embodiments, such a protein-binding probe kit can further comprises at least 1, or at least about 10, or at least about 100, or at least about 1000 additional set of protein-binding probes. In some embodiments, the kit does not comprise more than about 100, or more than about 1000, or more than about 5000 additional sets of protein-binding probes. In some embodiments, such a kit is an ELISA or other kits commonly known by persons of the art using protein-binding probes such as, for example, antibodies or antibody fragments.

[0309] In some embodiments, the kit contains control proteins or housekeeping genes for example but not limited to; genes or proteins such as: GAPDH (glyceraldehyde 3-phosphate dehydrogenase), SDHA (succinate dehydrogenase), HPRT (hypoxanthine phosphoribosyl transferase), HBS1L (HBS1-like protein), AHSP (alpha hemoglobin stabilizing protein), B2M (beta-2-microglobulin), beta-actin, beta-tubulin, large ribosomal protein, 18S ribosomal RNA, transferrin receptor, TATA-box-binding protein, beta-glucuronidase, cyclophilin A, porphobilinogen deaminase, ATP synthase 6 ; also see Trends in Genetics 19, 362-365 (2003). Other genes or proteins can be used as a normalizing gene, for example CALCR (calcitonin receptor), CRHR2 (Corticotropin-releasing hormone receptor 2), CYLC1 (Cylicin I), EDA (ectodysplasin A), GNRHR (GnRH receptor), HNF4A (hepatocyte nuclear factor 4, alpha), PAX8 (paired box 8), Mbd2 (methyl-CpG binding domain protein 2), ADAM22 (A disintegrin and metalloproteinase domain 22), Nf2 (neurofibromin 2), NPY2R (neuropeptide Y receptor Y2), TRIM15 (tripartite motif-containing 15), TRIM10 (tripartite motif-containing 10), RHAG (CD241 antigen), PDE10A (phosphodiesterase 10A) or ZBTB7A (Factor binding IST protein 1). These controls can be used at the DNA, RNA or protein level, such that they correspond to the measured biomarker species. For example, if mRNA of a cancer biomarker is measured, the mRNA for a control gene such as GAPDH is used for normalization of the expression level of the biomarker.

Methods of treatment of a subject identified to have cancer.

[0310] In some embodiments, where a subject is diagnosed with cancer by the methods described herein, the present invention further provides methods of treating such subjects identified to have, or be at risk of developing cancer. Accordingly, in one embodiment, the present invention relates to a method of directing treatment in a subject with cancer, wherein if the subject has a change in a prognostic cancer biomarker as disclosed in the methods herein, a clinician directs the subject to be treated with an appropriate anti-cancer treatment.

[0311] The term “anti-cancer agent” or “anti-cancer drug” is any agent, compound or entity that is capable of negatively affecting the cancer in the subject, for example killing cancer cells, inducing apoptosis in cancer cells, reducing the growth rate of cancer cells, reducing the number of metastatic cells, reducing tumor size, inhibiting tumor growth, reducing blood supply to a tumor or cancer cells, promoting an immune response against cancer cells or a tumor, preventing or inhibiting the progression of cancer, or increasing the lifespan of the subject with cancer. Anti-cancer therapy includes biological agents (biotherapy), chemotherapy agents, and radiotherapy agents. Accordingly, in one embodiment, the methods as disclosed herein permit an appropriate therapeutic regimen to be selected for the individual based on the prognosis of the detected cancer. For example, an aggressive anti-cancer therapy can be selected for a subject having an aggressive cancer prognosis. In alternative embodiments, a prophylactic anti-cancer therapeutic regimen can be pursued in a subject at risk of developing cancer, where the subject is administered a prophylactic dose or maintenance dose of an anti-cancer agent to prevent the development of cancer. In alternative embodiments, a subject can be monitored for cancer using the methods described herein, and the treatment regimen modified according to one of skill in the art of medicine. For example, upon initial testing a subject diagnosed with cancer is administered an appropriate anti-cancer therapy, and on a second (i.e. follow-up testing), if the subject is identified as not having cancer or having a change in expression of a prognostic cancer biomarker to a value closer to the reference sample than at the first testing, the therapeutic regimen can be modified as deemed appropriate by a skilled clinician (e.g., a reduction in dose, further aggressive therapy, or use of a different anti-cancer agent).

[0312] Anti-cancer agents or factors suitable for use in a combined therapy are any anti-angiogenic agent and/or any chemical compound or treatment method with anticancer activity. Such compounds or methods include alkylating agents, topoisomerase inhibitors, topoisomerase II inhibitors, RNA/DNA antimetabolites, DNA antimetabolites, antimetabolic agents, and DNA damaging agents, among others.

[0313] In specific embodiments, it is contemplated that an anti-cancer therapy, such as chemotherapy, radiotherapy, surgery, or immunotherapy, is employed in combination with the oncolytic adenovirus therapies, as described herein.

1. Chemotherapy

[0314] Anti-cancer agents that are contemplated for use in treating cancer in a subject include, but are not limited to, cisplatin (CDDP), carboplatin, procarbazine, mechlorethamine, cyclophosphamide, camptothecin, ifosfamide, melphalan, chlorambucil, busulfan, nitrosurea, dactinomycin, daunorubicin, doxorubicin, bleomycin, plicomycin, mitomycin, etoposide (VP16), tamoxifen, raloxifene, estrogen receptor binding agents, taxol, gemcitabine, navelbine, farnesyl-protein transferase inhibitors, transplatinum, 5-fluorouracil, vincristine, vinblastine and methotrexate, or any analog or derivative variant of the foregoing.

[0315] Further non-limiting examples of chemotherapy drugs and pro-drugs include, irinotecan (CPTII), temozolomide, platinum compounds and pro-drugs such as 5-FC. Examples of alkylating agents include, chloroambucil, cisplatin, cyclodisone, flurodopan, methyl CCNU (lomustine), piperazinedione, teroxirone. Topoisomerase inhibitors encompass compounds such as camptothecin and camptothecin derivatives, as well as morpholinodoxorubicin. Doxorubicin, pyrazoloacridine, mitoxantrone, and rubidazole are illustrations of topoisomerase II inhibitors. RNA/DNA antimetabolites include L-alanosine, 5-fluorouracil, aminopterin derivatives, methotrexate, and pyrazofurin, while the DNA antimetabolite group encompasses, for example, ara-C, guanazole, hydroxyurea, thiopurine. Typical antimitotic agents are colchicine, rhizoxin, taxol, and vinblastine sulfate.

[0316] Selection of an anti-cancer agent for a particular cancer and appropriate treatment methodologies using an anti-cancer agent are well within the abilities of one skilled in the art of medicine.

2. Radiotherapy

[0317] Other factors that cause DNA damage and have been used extensively for cancer treatment include γ -rays, X-rays, and/or the directed delivery of radioisotopes to tumor cells. Other forms of DNA damaging factors are also contemplated such as microwaves, proton beam irradiation (U.S. Patent 5,760,395 and U.S. Patent 4,870,287, which are incorporated herein in its entirety by reference) and UV- irradiation.

3. Immunotherapy

[0318] In the context of cancer treatment, immunotherapeutics, generally, rely on the use of immune effector cells and molecules (e.g. monoclonal antibodies) to target and destroy

cancer cells. Trastuzumab (Herceptin) is such an example. The immune effector may be, for example, an antibody specific for some marker on the surface of a tumor cell.

[0319] The antibody alone may serve as an effector of therapy or it may recruit other cells to actually effect cell killing. The antibody also may be conjugated to a drug or toxin (chemotherapeutic, radionuclide, ricin A chain, cholera toxin, pertussis toxin, etc.) and serve merely as a targeting agent. Alternatively, the effector may be a lymphocyte carrying a surface molecule that interacts, either directly or indirectly, with a tumor cell target. Various effector cells include cytotoxic T cells and NK cells. The combination of therapeutic modalities, i.e., direct cytotoxic activity and inhibition or reduction of ErbB2 would provide therapeutic benefit in the treatment of ErbB2 overexpressing cancers.

[0320] Some non-limiting examples of immunomodulatory anti-cancer agents include tumor necrosis factor; interferon alpha, beta, and gamma; IL-2 and other cytokines; F42K and other cytokine analogs; or MIP-1, MIP-1beta, MCP-1, RANTES, and other chemokines.

4. Surgery

[0321] Approximately 60% of persons with cancer will undergo surgery of some type, which includes preventative, diagnostic or staging, curative and palliative surgery. These types of surgery may be used in conjunction with other anti-cancer therapies.

[0322] Curative surgery includes resection in which all or part of cancerous tissue is physically removed, excised, and/or destroyed. Tumor resection refers to physical removal of at least part of a tumor. In addition to tumor resection, treatment by surgery includes laser surgery, cryosurgery, electrosurgery, and microscopically controlled surgery (Mobs surgery).

Disease Diagnosis and Prognosis

[0323] In addition to diagnosis and prognosis of cancer, the methods described herein are also useful in a variety of other diseases. In some embodiments, the methods as disclosed herein can be applied to other diseases with appropriate animal models to identify biomarkers for such diseases. Further, in some embodiments, to establish metagene score for other diseases a reference sample (i.e. a normal or control sample (i.e. a negative sample)) can be used to establish a reference metagene score and a diseased sample (i.e. a positive sample) can be used to establish the metagene score for the disease, disease prognosis, longevity etc.

[0233] In some embodiments, the methods as disclosed herein can be applied to all diseases with copy number variations, such as nervous system diseases (e. g. Alzheimer, Charcot-Marie-Tooth disease type 1A, neuropathies, psychiatric illnesses), autoimmune diseases (e.g. systemic lupus erythematosus (SLE), idiopathic thrombocytopenic purpura (ITP)), metabolic

diseases (e.g. diabetes, obesity), cardiovascular diseases, HIV and malaria susceptibility and trisomy 21, Turner syndrome, Klinefelter syndrome, triple X syndrome, fragile X syndrome, Wolf-Hirschhorn syndrome, Cri du chat, Angelman syndrome/Prader-Willi syndrome, Miller-Dieker syndrome/Smith-Magenis syndrome, 22q11.2 deletion syndrome or gonadal dysgenesis (Lee and Lupski, 2006; McCarroll and Altshuler, 2007; Gonzalez, E. et al. 2005; Aitman, T.J. et al. 2006; Yang, Y. et al. 2007; Fanciulli, M. et al. 2007; Sebat J. et al. 2007).

[0324] Other objects, features and advantages will become apparent from the following detailed description. It should be understood, however, that the detailed description and specific examples, while indicating specific embodiments of the invention, are given by way of illustration only, since various changes and modifications within the spirit and scope of the invention will become apparent to those skilled in the art from this detailed description.

[0325] The invention now being generally described, will be more readily understood by reference to the following examples, which are included merely for purposes of illustration of certain aspects and embodiments of the present invention, and are not intended to limit the invention

[0326] The following examples are provided to illustrate certain embodiments of the invention. They are not intended to limit in any way the remainder of the disclosure.

EXAMPLES

[0327] Using mouse models of cancer, the inventors have discovered a prognostic signature of cancer biomarker genes. The inventors have further identified a subset of these prognostic cancer biomarkers as a core prognostic biomarker signature. The inventors have discovered the expression of these cancer biomarkers using a novel method, herein termed Genome Association Analysis of Transcription and Comparative Oncogenomics (GAATCO), to identify genes with altered gene copy numbers and genes differentially expressed in a cancer sample, which are correlated using bioinformatic algorithms, e.g. metagene-based methodology. In some embodiments, the altered gene copy number is determined by comparing the copy number for the biological sample, i.e. a sample with a chromosomal abnormality or gene copy number variation such as a cancer sample, with a reference copy number for that gene, such as a reference copy number for that gene from a sample that does not have the chromosomal abnormality or gene copy number variation. The inventors then selected those genes with an altered copy number and gene expression level in the cancer

sample, and cross-compared using comparative methods with another species (orthologous methods), for example looking for human genes using standard bioinformatics tools to identify orthologous genes. Using such a method on mouse models of cancer, such as mouse models of leukemia and lymphoma, the inventors have identified a group of cancer biomarkers or a subgroup thereof as diagnostic and/or prognostic cancer biomarkers as disclosed in Table 1.

[0328] The examples presented herein relate to the methods of diagnosing a subject with cancer, by measuring expression of cancer biomarkers and subgroups thereof. Another aspect of the present invention relates to biomarker discovery, in particular a method to identify disease biomarkers, such as biomarkers of diseases with chromosomal abnormalities, the method comprising GAATCO. Throughout this application, various publications are referenced. The disclosures of all of the publications and those references cited within those publications in their entireties are hereby incorporated by reference into this application in order to more fully describe the state of the art to which this invention pertains. The following examples are not intended to limit the scope of the claims to the invention, but are rather intended to be exemplary of certain embodiments. Any variations in the exemplified methods which occur to the skilled artisan are intended to fall within the scope of the present invention.

METHODS

[0329] The inventors generated DNA copy-number and gene expression data from 18 progenitor B lymphomas in *Art^{NN}Trp53^{NN}* mice using array CGH and Affymetrix GeneChip MOE 430v2 methods as described by Woo et al. 2007. Raw aCGH and Affymetrix GeneChip data were normalized and processed into matched DNA copy-number and gene expression profiles for 15867 unique genes and 18 samples as described by Woo et al. 2007.

Human Cancer Data Sets

[0330] The inventors obtained expression data sets of seven human cohorts comprising two B lymphoma, three breast cancer, and two glioma cohorts, from the Gene Expression Omnibus(GEO) of the National Center for Biotechnology Information (GEO accession number: GSE4475, GSE3494, GSE7390 ,GSE4412, GSE4271, GSE2034 and the Molecular Diagnosis of Burkitt's Lymphoma resource available on the world wide web at lmpp.nih.gov/BL.

GAATCO Analysis

[0331] GAATCO uses multi-step gene identification algorithm. Here, the inventors used GAATCO to identify gene expression profiles associated with chromosome copy number aberrations (CNAs), as follows: 1) CNAs containing genes with concordant expression behavior were identified, by comparison of array CGH and Affymetrix GeneChip data. 2) Genome-wide (global) transcriptional signatures significantly associated with individual CNAs were identified. True global correlations were distinguished from spurious or coincidental correlations as described by Bussey et al. 2006 and Nigro et al. 2005 using the linear model approach. 3) Finally step genome-wide transcriptional signatures were winnowed to generate core, evolutionarily conserved signatures by comparing expression patterns across species as described by Stuart et al. 2003. Specific details of GAATCO algorithm construction and implemented are described in Methods above..

Step 1: Identification of CNAs

[0332] Genes showing copy-number aberrations (absolute ratios greater than 0.1) in at least one sample were selected. Next, by regression analysis, genes showing significant correlation between its copy-number and expression levels were selected (Holms-Bonferroni adjusted $p < 0.05$), and they were grouped into CNAs by the following criteria: genes consecutive by chromosomal order and showing correlated copy-numbers (Spearman correlation $r > 0.95$) were grouped as a CNA.

Step 2: Global Association Analysis

[0333] For all possible pairs between CNAs and genes, the inventors tested for copy-number (CNA) to expression (gene) association using the following linear models:

$$T_j = \beta_0 + \epsilon_j \quad (1)$$

$$T_j = \beta_0 + \beta_1 D_j + \epsilon_j \quad (2)$$

$$T_j = \beta_0 + \beta_2 C_j + \epsilon_j \quad (3)$$

$$T_j = \beta_0 + \beta_1 D_j + \beta_2 C_j + \epsilon_j \quad (4)$$

,where T_j denotes expression, D_j is copy-number, and ϵ_j is random error of the gene in sample j . C_j denotes copy-number of the CNA in sample j . β_1 and β_2 denote regression parameters measuring local and global association, respectively. β_0 is an intercept regression parameter. The inventors used a forward, step-wise model selection strategy to test for statistical significance of β_2 . First, the inventors tested for statistical significance of β_1 by comparing model 1 and 2. Second, based on the model selected from the first step, the inventors assessed statistical significance of β_2 , the parameter of interest, either by comparing model 1 and 3 or

by comparing model 2 and 4. Resulting p-values were adjusted for multiple-testing by false discovery rate calculation, separately for each CNA (Storey and Tibshirani, 2003).

Step 3: Comparative Analysis

[0334] A group of genes detected from the above step were then narrowed to core subsets showing conserved co-expression patterns by: (i) identifying mouse-to-human orthologs, (ii) constructing gene-by-gene correlation matrices for the mouse and human data sets, (iii) generating a conservation matrix which scores agreement between the correlation matrices, and (iv) k-means clustering the conservation matrix to identify a core subset.

Simulation

[0335] To evaluate the ability of GAATCO to distinguish bona fide global associations from spurious copy-number dependent associations, simulated copy number and gene expression data were generated, representing 100 idealized samples containing 24 genes (6 genes per chromosome x 4 chromosomes). Simulated copy-number gains of the second gene, with correlated global transcriptional changes throughout the genome, were added to 20 of the 100 samples. Random chromosomal gains and losses, along with corresponding expression changes of the same genes, affecting 10 randomly selected subsets of the remaining 80 samples were then superimposed on the dataset. To introduce additional noise, simulated chromosomes 1 and 4 were given identical chromosomal gain and loss patterns. This simulated data set was used to test both the sensitivity and the specificity of GAATCO, relative to conventional correlation analysis, in detecting true global associations between copy number aberrations and gene expression, in the context of experimental noise.

EXAMPLE 1

[0336] The inventors used a well-characterized mouse pro/pre-B cell cancer model to investigate the relationship between copy number abnormalities (CNA) and global transcriptional profiles in cancer cells. Mice doubly deficient in the DNA double-strand break repair gene *Dclre1c*/Art and the tumor suppressor gene *Trp53* (hereafter referred to as AP mice) rapidly succumb to aggressive pro-B cell neoplasms that closely resemble human pediatric B-ALL (Rooney et al. 2004b; Woo et al. 2007; Zhu et al. 2002). Lymphomas from AP mice (AP lymphomas) harbor highly characteristic complicons, that harbor result co-amplification of the immunoglobulin heavy chain (IgH) locus and either the c-myc or the N-myc proto-oncogene (Rooney et al. 2004). These tumors also contain a host of recurrent, but less penetrating lesions of unknown importance. The inventors applied the GAATCO technique to a panel of AP lymphomas, identifying an evolutionarily conserved, global gene

expression signature associated with deletions on mouse chromosome (Maser et al. 2007). The inventors demonstrate that this signature comprises genes in key cancer-related pathways and, importantly, that these can predict survival in the mouse model and in a range of human cancers. These results demonstrate that GAATCO is a robust method for detecting functionally relevant, predictive gene signatures associated with specific chromosome lesions. Notably, the inventors have demonstrated that discrete chromosome lesions may be selected on the bases of both local and global gene deregulation, highlighting the need to analyze CNA-associated transcriptional behavior genome-wide.

Mouse model to test the molecular etiology and the pathogenetic effects of chromosomal copy number abnormalities (CNA).

[0337] As mouse model for B-lymphoid malignancy *Lig4 Trp53* double null mutant (Frank et al., 2000). The *Lig 4* knockout (B6;129S6-Lig4tm1Fwa/Kvm) can be obtained from The Jackson Laboratory (JR6482) and crossed to the *Trp53* null mouse, which can be obtained from The Jackson Laboratory (JR2103), resulting in double null *Lig4 Trp53* mice. For the studies requiring *Lig4*^{-/-} *Trp53*^{-/-} mice, these can be generated by intercrossing *Lig4*^{+/-} *Trp53*^{+/-} mice. Because *Lig4* is an essential gene, with homozygous null mice exhibiting lethality by approximately day 18 of embryonic development (E18), mice containing a null allele of *Lig4* will be maintained as heterozygotes. While deficiency for p53, encoded by the *Trp53* gene, rescues viability in *Lig4*-null animals, the rapid onset of aggressive pro-B cell lymphoma precludes breeding of *Lig4 Trp53*-double null mice. Thus *Lig4 Trp53* double mutant animals can be maintained and bred as *Lig4*^{+/-} *Trp53*^{+/-} or *Lig4*^{+/-} *Trp53*^{-/-}. This breeding scheme will produce *Lig4 Trp53* double null animals at frequencies from 0.0625 to 0.25.

[0338] Five *Lig4 Trp53* pro-B cell tumors were isolated and processed for DNA and RNA isolation using standard molecular biology methods. Tumor tissue was soaked overnight in RNALater (Ambion), and stored at -20 degree Celsius, or further processed by TRIZOL® (Invitrogen) extraction, according to the manufacturer's protocol. Genomic DNA was prepared by phenol chloroform extraction and isopropanol precipitation according to standard molecular biology method as described in Molecular Cloning: A Laboratory Manual (3rd edition; 2001); by Joseph Sambrook, Peter MacCallum and David Russell.

[0339] The isolated DNA was used for comparative genomic hybridization to an array of defined genomic probes (array CGH) using a BAC-mediated array with a 3Mb average inter-probe resolution (Spectral Genomics Houston, TX). This method is also described by Zhu et

al. 2002. For each tumor sample, hybridizations were performed in dye swapped duplicates, and compared against reference non-tumor DNA from a normal mouse.

[0340] For analysis of copy number abnormalities in mouse tumors, labeled genomic DNA from the tumor sample was compared with non-tumor reference DNA. Copy number gains resulted in an increase of test-to-reference signal ratio while sequence losses result in a decrease of test-to-reference ratio. Data from dye-reversed duplicate experiments is plotted graphically as a function of the test/reference fluorescence at each probe position. The data is shown in Figure 5 showing amplification encompassing *Igh* and *c-myc* on chromosomes 12 and 15 (see Figure 5 B). Further a putative amplicon on chromosome 5 and a putative deletion on chromosome 9 were identified (see Figure 5 C). The chromosome 5 amplicon encompasses the Sequence-Tagged-Site (STS) marker D5Mit202 was identified in two out of five tumors. The deletion on chromosome 9 encompasses the STS marker D9MIT303 in four out of five tumors.

EXAMPLE 2

Correlation of mouse model to human disease.

[0341] The affected region in Lig4 Trp53 lymphomas of mouse chromosome 5, surrounding the array marker D5Mit202, shows conservation in the syntenic human chromosome 4q23 and the affected region of mouse chromosome 9 surrounding the array marker D9Mit330 shows conservation in the syntenic human chromosome 11q21. While the precise identity of all affected genes from each region is unknown, several candidates for involvement in pro-B lymphomagenesis map to these regions and are indicated in Figure 1.

EXAMPLE 3

Identification of affected genes by obtaining a higher resolution CNA.

[0342] Higher resolution CNA mapping using an oligonucleotide-mediated platform, for example based on a Compugen/Sigma-Genosys probe set, with about 22,000 oligonucleotides (65nt) corresponding to the 3' ends of mouse expressed sequences allows a higher CNA mapping resolution and a robust detection of copy number abnormalities, following elimination of poorly mapped or high-background probes. Probes that were ambiguously mapped relative to the NCBI mouse genome Build 34 and probes that exhibit outlier signals in normal versus normal control hybridizations were eliminated (Barrett et al., 2004; Brennan et al., 2004; Rauch et al., 2004).

[0343] After probe mapping, the inventors retained probes that aligned to one, and only one genomic site, and exhibited a 100% sequence match. These criteria, even more stringent than those previously used, excluded approximately 8000 suboptimal probes, and produced a high quality CGH collection of more than 14,000 probes eliminated (Barrett et al., 2004; Brennan et al., 2004; Rauch et al., 2004).

EXAMPLE 4

Calibration of an array for Copy Number Analysis (CNA) mapping.

[0344] To calibrate the array for CNA mapping, the inventors first performed array CGH analysis of normal C57BL/6J male versus female genomic DNA. The genomic DNA was isolated from kidneys using standard molecular genetics methods as described in Molecular Cloning: A Laboratory Manual (3rd edition); by Joseph Sambrook, Peter MacCallum and David Russell. The purified DNA was treated with DpnII restriction enzyme. DNA fragment were labeled with Cy3 or Cy5 by random octamer priming and extension, according to manufacturer's protocols (Invitrogen; Carlsbad, CA).

[0345] For copy number analysis, Mouse Exonic Evidence-Based Oligonucleotide (MEEBO), two-channel spotted microarrays were used (Microarrays Inc.; Nashville, TN). MEEBO contains 38,784 long oligonucleotide (70-mer) probes containing 25,000 genes of the mouse genome. 4ug of genomic DNA was labeled with either Cy3 or Cy5 by random octamer priming and extension, according to manufacturer's protocols (Invitrogen; Carlsbad, CA). The MEEBO microarrays were simultaneously hybridized with 100 pmoles each of labeled tumor and control DNA. Images were obtained using an Axon 4000B Scanner, and gridded and flagged by the spot finding feature of GenePix Software, version 5.1. For all tumors test-versus-reference hybridizations were performed in dye-swapped duplicates (Kerr and Churchill, 2001). In a previous study by the inventors using this platform, copy number changes as detected by the MEEBO array CGHs showed a high agreement with copy number changes; as detected by an independent QPCR experiment.

[0346] Microarray hybridization was performed according to manufacturer description using compugen microarrays and image collection were performed as described above. Both X and Y chromosome copy number differences are readily detectable (Fig. 6).

EXAMPLE 5***Detection of copy number abnormalities by quantitative PCR.***

[0347] Copy number abnormalities can be detected by quantitative PCR, which can be done in a high-throughput 384-well real time PCR platform together with software analysis, for example Global Pattern Recognition (GPR) (Akilesh et al., 2003; Brown et al., 2005; Hart et al., 2004). DNA isolated from a pro-B cell lymphoma with verified Igh and c-myc copy number imbalances determined by FISH, was analyzed by quantitative PCR and GPR analysis. Triplicate tumor DNA samples were compared with reference DNA. Cycle threshold (Ct) values were normalized using quantile normalizations, and copy numbers were estimated by ANOVA using R/Maanova (Irizarry et al., 2003; Wu et al., 2003; Churchill, 2004).

[0348] Table 9 shows the ten highest scoring copy number gains identified. Five of the top ten hits returned by this approach represent the Igh and c-myc loci, and six of the top ten hits reflect chromosome 12 amplification at a distal position, these results also indicate that Igh and c-myc are amplified to the same level (~4 fold), which is expected due to their co-amplification by complicon formation. This verifies the sensitivity and reliability of this method analysis, and demonstrates its general utility as an independent and complementary approach to array CGH.

Table 9: Quantitative PCR/GPR analysis of the top ten copy numbers detected in the Art Trp53 tumor (AP546)

Top ten copy number increases detected							
Rank	Gene	Chr	Fold Increase (norm)	Control CT (avg)	Test CT (avg)	Δ CT	$2^{-\Delta\Delta\text{CT}}$
1	<i>Traf3</i>	12	5.41	29.97	28.64	-1.33	2.52
2	<i>IgG2a</i>	12	4.66	29.78	28.74	-1.04	2.05
3	<i>IgG1a</i>	12	4.40	29.82	28.86	-0.96	1.94
4	<i>Nras</i>	3	4.17	29.48	28.63	-0.85	1.80
5	<i>IgG2c</i>	12	4.02	30.46	29.58	-0.88	1.85
6	<i>Xrcc3</i>	12	4.00	31.29	30.52	-0.77	1.71
7	<i>Myc</i>	15	3.99	29.15	28.37	-0.78	1.72
8	<i>IgG1m</i>	12	3.93	30.10	29.25	-0.85	1.80
9	<i>Itgal</i>	7	3.49	32.29	31.55	-0.74	1.67
10	<i>Hmgbl3</i>	X	3.47	26.38	25.72	-0.66	1.58

EXAMPLE 6***Mouse model deficient in DNA repair (*Art^{NN}*) and cell cycle checkpoint (*Trp53^{NN}*) pathways.***

[0349] Mice heterozygous for a null allele of the gene encoding Artemis (Dclre1c/*Art*; Rooney et al., 2004b) were crossed with Trp53 heterozygous mice to produce doubly heterozygous

129Sv/C57BL/6 $Art^{N/+}Trp53^{N/+}$ mice. These were backcrossed to C57BL/6 $Art^{N/N}$ mice, and progeny were crossed to generate $Art^{N/N}Trp53^{N/N}$ and $Art^{N/N}Trp53^{N/+}$ mice.

[0350] Differences in survival for $Art^{N/N}Trp53^{N/+}$ and $Trp53^{N/+}$ cohorts were tested using a log-rank test of survival analysis in JMP6.0 (SAS Institute Inc., Cary, NC, USA). Mice deficient in DNA repair ($Art^{N/N}$) and cell cycle checkpoint ($Trp53^{N/N}$) pathways develop aggressive B-lineage lymphomas harboring numerous genetic alterations (Rooney et al., 2003; Woo et al., 2007). B lineage lymphoid tumors were isolated from $Art^{N/N}Trp53^{N/N}$ mice and processed for genomic DNA and total RNA. Genomic DNA was prepared by phenol chloroform extraction and isopropanol precipitation. Tumor tissue was soaked overnight in RNALater (Ambion), and stored at -20 °C, or further processed by Trizol (Invitrogen) extraction, according to the manufacturer's protocol for total RNA isolation. For a control DNA sample, genomic DNA was extracted from C57BL/6J (B6) mouse kidneys. DNA and RNA have been used in gene expression and gene copy number analysis.

EXAMPLE 7

Comparative Genomic Hybridization (CGH) and Affymetrix GeneChips to interrogate genome-wide DNA copy number and gene expression profiles of lymphomas.

[0351] From 19 lymphomas from ($Art^{N/N}Trp53^{N/N}$) mice total DNA and RNA was isolated and used for genome-wide DNA copy number and gene expression profiles using Comparative Genomic Hybridization (CGH) array, MEEBO aCGH and Affymetrix GeneChips (e.g. GeneChip MOE 430v2).

[0352] For copy number analysis, Mouse Exonic Evidence-Based Oligonucleotide (MEEBO), two-channel spotted microarrays were used (Microarrays Inc.; Nashville, TN). MEEBO contains 38,784 long oligonucleotide (70-mer) probes covering 25,000 mouse genes. 4 ug of genomic DNA was labeled with either Cy3 or Cy5 by random octamer priming and extension, according to manufacturer's protocols (Invitrogen; Carlsbad, CA). The MEEBO microarrays were simultaneously hybridized with 100 pmoles each of labeled tumor and control DNA. Images were obtained using an Axon 4000B Scanner, and gridded and flagged by the spot finding feature of GenePix Software, version 5.1. For all tumors test-versus-reference hybridizations were performed in dye-swapped duplicates (Kerr and Churchill, 2001). In a previous study using this platform, copy number changes as detected by the MEEBO array CGHs showed a high agreement with copy number changes as detected by an independent QPCR experiment (Woo et al., 2007). For expression studies, the inventors used GeneChip MOE 430v2 from Affymetrix, Inc: one-channel, oligonucleotide microarray

platform. DNA and RNA samples from the 27 tumors were assayed for copy number and expression changes using 38 (19 °—2) MEEBO array CGH and 19 (19 °— 1) Affymetrix GeneChips. See Figure 9 for experimental design. Each tumor vs. control comparison in DNA copy number analysis was replicated twice in a dye-swap fashion to account for dye-specific biases.

EXAMPLE 8

GAATCO - Derivation of Core prognostic cancer biomarkers Genome Association Analysis of Transcription and Comparative Oncogenomics (GAATCO).

[0353] The inventors applied GAATCO to genome-wide copy-number and expression profiles of B lymphoma samples from 19 Art^{N/N}Trp53^{N/+} mice (see Example 6). First, 19 candidates for phenotypically significant copy number lesions were identified. The inventors applied regression analysis to identify 120 genes showing statistically significant correlation between their copy number and expression levels. Comparison of copy-number and transcriptional levels is important for prioritizing candidate genetic lesions since all copy number lesions, for example chromosome 3 gains, show corresponding expression changes. The 120 genes were further categorized into 19 genetic lesions showing distinct copy number patterns (Figure 11A). They comprised both novel and known genetic lesions in B cell lymphomas.

[0354] Recurrent amplification of *c-myc* and *N-myc* loci and their transcriptional up-regulations is consistent with previously published cytogenetic and Northern blot findings (Rooney et al., 2004b; Zhu et al., 2002). The inventors also observed recurrent chromosome 6 and 19 deletions with corresponding transcriptional down-regulations. Therefore, the inventors screened for gene expression signatures associated with the 19 genetic lesions on a genome-wide basis (Figure 11 B, C). While most genetic lesions show linkage to few genes, chromosome 6 deletions were associated with expression changes of 370 genes, significantly more than expected by permutation analysis ($p < 0.05$). Chromosome 6 deletions were also associated with changes in tumor-free survival ($p < 0.001$, Figure 11 D). These data show that the large-scale transcriptional change associated with the chromosome 6 deletion can play a role in determining clinical outcomes. The gene signature, here forth called “the prognostic signature”, was subject to further analyses.

[0355] By comparative coexpression analysis, the inventors filtered the prognostic signature to a core conserved signature comprising 136 genes. Out of the 464 genes, the inventors identified 321 *M.musculus*-*H.sapiens* orthologous genes and constructed gene-by-gene correlation matrix for one mouse and three human B expression datasets (Basso et al., 2005;

Dave et al., 2006; Hummel et al., 2006). According to Figure 2 A, many genes, but not all, show conserved co-expression patterns across species. Thus, the inventors compacted the gene set by filtered for genes with high degree of co-expression conservation. The inventors applied hierarchical clustering to select a core set of 136 genes, here forth called “the core prognostic cancer biomarker s” (Lee et al., 2004) (Figure 2 B, C).

EXAMPLE 9

Testing Human Cohorts

[0356] Eleven publicly available human cancer datasets have been obtained and analyzed for the 51, 91 and 136 genes. Data sets were selected such to contain tumor types with at least two independent cohorts and with at least 50 patients (Bhattacharjee et al. 2001; Bild et al. 2006; Dave et al 2006, Desmedt et al 2007, Freije et al. 2004, Miller et al. 2005, Nutt et al. 2003, Pawitan et al. 2005, Philips et al. 2006, Wang et al. 2005, Daily et al. 2007). The inventors tested for prognostic value of the core conserved set in various cancer cohorts. After summarizing expression of the 51 genes, 91 genes and 136 genes into a metagene score by singular value decomposition technique, the inventors tested by Cox regression analysis whether the metagene score are significantly correlated with tumor-free survival. The inventors observed that the metagene scores were significant predictors of survival in 9 out of the 12 cohorts representing several tumor types: B lymphoma, glioma, and breast cancer. To test if the inventors can stratify the cohorts into patients with different survival outcomes, the inventors divided each cohort into two equal-sized groups by their metagene scores and detected significant survival difference in 8 out of 12 cohorts. These results demonstrate the prognostic ability of the core signature set and its potential as a clinical diagnostic tool for predicting cancer prognosis (Figure 3).

Survival Analysis

[0357] For continuous score, the inventors performed survival analysis by cox proportional hazard regression analysis using R/SURVIVAL:

$$H_0 : \log S(t) = \alpha(t) + \varepsilon \quad H_a : \log S(t) = \alpha(t) + sDC_i + \varepsilon \quad (5)$$

[0358] $S(t)$ denotes the duration of tumor-free survival, $\alpha(t)$ the baseline hazard function, sDC_i the slope of survival data over copy number of genetic lesion i , m - stratifying variable, and ε residual. For discrete score, the inventors used a log-rank test to test for group-wide difference in survival between the two groups.

EXAMPLE 10***GAATCO links segmental chromosome 6 deletions to a conserved genome-wide signature***

[0359] Many recurrent chromosomal lesions observed in AP lymphomas are likely to critically influence tumor phenotypes, but have not yet been characterized in detail (Woo et al. 2007; Rooney et al. 2004). AP refers to mice doubly deficient in the DNA double-strand break repair gene *Dclre1c/Art* and the tumor suppressor gene *Trp53*. To investigate the function of various lymphoma/leukemia-associated chromosomal lesions, the inventors analyzed a panel of 18 AP lymphomas using the novel GAATCO algorithm. As discussed herein, GAATCO is a novel, linear model-based method for uncovering genome-wide expression patterns associated with specific copy number aberrations (CNAs). The key distinguishing features of GAATCO are 1) its use of multi-step filtering to winnow large gene lists to those significantly associated with chromosome lesions of interest; and 2) the use of cross-species comparison, to identify genes with functional significance based on evolutionary conservation in related tumor types (Figure 10A). To assess the sensitivity and specificity of GAATCO to detect bona fide global signatures, simulated, idealized genomic and transcriptomic data were generated, and analyzed by either conventional correlation or by GAATCO. Even after removing copy-number dependent correlations, GAATCO was able to detect global associations at least as effectively as correlation coefficients, but could also discern true associations, despite experimental noise.

[0360] As a first step, all CNAs in the lymphoma panel were identified by array comparative genomic hybridization (aCGH). This approach readily detected the previously described compicons, and several deletions or amplifications of unknown functional importance (Figure 10A) (Rooney et al. 2004; Woo et al 2007). Transcriptional profiles were obtained for the same panel of AP lymphoma panel, by Affymetrix GeneChip analysis. A total of 15 CNAs were found that showed significant correlation between DNA copy number and expression level of the encompassed genes (local correlation). Survival analysis revealed that, of the 15 correlated CNAs, deletions on chromosome 6 (6C1-G3) were most significantly associated with tumor-free survival (Fig.3B). Notably, this region of mouse chromosome 6 shares blocks of conserved synteny with several human genome regions implicated in B-cell neoplasms (Figure 11D).

[0361] Next, association analysis was carried out to detect genome-wide (global) transcriptional patterns associated with each CNA. Owing to their association with survival time, the inventors focused further analyses on chromosome 6C1-G3 deletions. After adjusting for

multiple testing, 353 genes, throughout the genome, showed expression levels significantly (FDR $q < 0.005$) correlated with deletions on chromosome 6 (Fig.11C).

[0362] To narrow the gene list to those with probable functional roles in lymphomagenesis, cross-species comparative analysis was performed. After identifying 321 genes with unique mouse-to-human orthology, the expression pattern of each ortholog was compared in the mouse lymphoma and two human B-cell lymphoma datasets (Wheeler et al. 2008; Dave et al. 2006; Hummel et al. 2006). Correlation coefficients were determined between all possible gene pairs in each dataset, and then evaluated for agreement of correlation coefficients between the datasets (Fig. 2B). K-means clustering next identified a set of 136 genes (herein referred to group II and as disclosed in Table 2) with correlation coefficients that agreed significantly in sign (i.e. positive or negative correlation). Remarkably, 96% of the comparisons between correlation coefficients showed sign agreement in all three datasets, demonstrating that the 136-gene set contains an evolutionarily conserved gene expression profile for B-lineage neoplasms.

The core conserved signature contains genes implicated in tumor critical pathways

[0363] To identify the relevant biological pathways, the 136-gene (Group II) set was subdivided into those genes with expression that positively correlated or negatively correlated with chromosome 6 deletions in AP lymphomas (Fig.2C-D). These subgroups were independently analyzed by PANTHER (Protein ANalysis THrough Evolutionary Relationships) ontology analysis (Thomas et al 2003; Thomas et al. 2006). The positively correlated subset was significantly enriched for genes with cell cycle and genome stability related annotations including DNA metabolism, cell cycle, DNA repair, and DNA replication ($p < 0.05$), while the negatively correlated subset was enriched with genes implicated in extracellular signaling pathways, cell-based immunity, and cell motility ($p < 0.05$). Given the large body of evidence showing deregulation of these pathways in tumor initiation or progression, their significant enrichment here validated the ability of GAATCO to uncover functionally important pathways (Whitfield et al. 2006). The inventors have demonstrated that the possibility that many such changes can occur coordinately as a consequence of specific chromosomal lesions.

The conserved signature can predict cancer prognosis in multiple tumor types

[0364] Given the enrichment for tumor-related pathways in the core signature, the inventors assessed use of the biomarker gene expression signature as a cancer prognostic indicator. To test this, the inventors obtained gene expression data and patient outcome information from 7 publicly available human cancer studies of lymphoma, breast cancer, and glioma (Dave et al.

2006; Hummel et al. 2006; Desmedt et al. 2007; Freije et al. 2004; Miller et al. 2005; Phillips et al. 2006; Wang et al. 2005). For each sample, a risk score was calculated based on the 136-gene signature (Group II, Table 2), and used to bifurcate each dataset into high- and low-risk subsets. These were compared for survival by Kaplan-Meier analysis, identifying statistically significant differences ($p < 0.05$) in survival the AP lymphoma set and in 5/7 human studies. These data indicate that the core, conserved signature contains functionally important genes that modulate tumor phenotypes, and could be used as a molecular prognostic for a range of human cancers.

EXAMPLE 11: FOXM1 and RECQL4 as prognostic markers

[0365] The prognostic role of FOXM1 (Forkhead box M1) and RECQL4 transcription factor in terms of tumor-free survival was identified in seven human cohorts (Dave et al., 2006; Desmedt et al., 2007; Freije et al., 2004; Hummel et al., 2006; Miller et al., 2005; Phillips et al., 2006; Wang et al., 2005). Cox proportional hazard regression calculation was used to assess statistical significance of the prognostic value (see Table 10).

Table: 10

Human cohort	p value for FOXM1	p value for RECQL4
B lymphoma 1 (Dave06)	0.0484100	0.002330
B lymphoma 2 (Hummel06)	0.5638000	0.001360
Breast cancer 1 (Desmedt07)	0.0012690	0.090894
Breast cancer 2 (Miller05)	0.0051340	0.005950
Breast cancer 3 (Wang05)	0.0114600	0.000024
Glioma 1 (Freije04)	0.0002655	0.003880
Glioma 2 (Phillips06)	0.0027990	0.002461

Table 11. Primer Sequences for Exemplary Prognostic Genes

Types	haEntrez	homID	hgsg	Primer ID	Primer Sequence	RefSeq Transcript ID	RefSeq Transcript GI ID
prognostic sig	51205	41128	AC6	NM_016361.3,forward.1, 2	GGACAGATGGGCGATTGACA	NM_016361.3	151301212
prognostic sig	51205	41128	AC6	NM_016361.3,forward.3	GGAGCTTGAGCGAGAAC	NM_016361.3	151301212
prognostic sig	51205	41128	AC6	NM_016361.3,reverse.1	GCAGCTTGAGAGGTTGTG	NM_016361.3	151301212
prognostic sig	51205	41128	AC6	NM_016361.3,reverse.2	CAGCTTGAGAGGTTGTG	NM_016361.3	151301212
prognostic sig	51205	41128	AC6	NM_016361.3,reverse.3	GCCCATCTGCTTCACT	NM_016361.3	151301212
prognostic sig	50650	41329	ARHGEF3	NM_019555.2,forward.1	ACTGCAGCTGAGCTTACC	NM_019555.2	190885489
prognostic sig	50650	41329	ARHGEF3	NM_019555.2,forward.2	CCACATGAGCAGCTTGC	NM_019555.2	190885489
prognostic sig	50650	41329	ARHGEF3	NM_019555.2,forward.3	TAGCCAGTGAATCGGAA	NM_019555.2	190885489
prognostic sig	50650	41329	ARHGEF3	NM_019555.2,reverse.1	TAGCCAGTGAATCGGAA	NM_019555.2	190885489
prognostic sig	50650	41329	ARHGEF3	NM_019555.2,reverse.2	TCCTGCTGCTGAGGATCGT	NM_019555.2	190885489
prognostic sig	50650	41329	ARHGEF3	NM_019555.2,reverse.3	TCCTGCTGCTGAGGATCGT	NM_019555.2	190885489
prognostic sig	79915	32611	ATAD5	NM_024857.3,forward.1	TGCTGCTGCTGAGGATCGT	NM_024857.3	26080430
prognostic sig	79915	32611	ATAD5	NM_024857.3,forward.2	GGCCACAGAGTGGGAAA	NM_024857.3	26080430
prognostic sig	79915	32611	ATAD5	NM_024857.3,forward.3	AGGACCTGCTGATACCTG	NM_024857.3	26080430
prognostic sig	79915	32611	ATAD5	NM_024857.3,reverse.1	TGCTGCTGCTGAGGATCGT	NM_024857.3	26080430
prognostic sig	79915	32611	ATAD5	NM_024857.3,reverse.2	TACCACTGCTGAGGAGAG	NM_024857.3	26080430
prognostic sig	79915	32611	ATAD5	NM_024857.3,reverse.3	GATCAGCCGAATCGAACACG	NM_024857.3	26080430
prognostic sig	535	3795	ATP6V0A1	NM_005177.3,forward.1	CGAGGCTGACAGTATGCT	NM_005177.3	77539781
prognostic sig	535	3795	ATP6V0A1	NM_005177.3,forward.2	GCGACAGGCTGAAATCGAGA	NM_005177.3	77539781
prognostic sig	535	3795	ATP6V0A1	NM_005177.3,forward.3	CCGACAGAGAGATGCTGA	NM_005177.3	77539781
prognostic sig	535	3795	ATP6V0A1	NM_005177.3,reverse.1	AAAGAGACCCCACTCCACA	NM_005177.3	77539781
prognostic sig	535	3795	ATP6V0A1	NM_005177.3,reverse.2	CAGCTGATCGCTTGAAAGA	NM_005177.3	77539781
prognostic sig	535	3795	ATP6V0A1	NM_005177.3,reverse.3	CCATGCTGCCCAAGTCA	NM_005177.3	77539781
prognostic sig	83990	32766	BRP1	NM_032043.1,forward.1	CCAGGCTCGCTTATTG	NM_032043.1	14042977
prognostic sig	83990	32766	BRP1	NM_032043.1,forward.2	TCCTCAGGCTCCGCTTA	NM_032043.1	14042977
prognostic sig	83990	32766	BRP1	NM_032043.1,forward.3	CAGGCTCCGCTTATTG	NM_032043.1	14042977
prognostic sig	83990	32766	BRP1	NM_032043.1,reverse.1, 2, 3	GGAAACCGAAGCAACGCTCT	NM_032043.1	14042977
prognostic sig	55108	32386	BSDC1	NM_018045.5,forward.1	AGGCTCCCTGCTCTG	NM_018045.5	141802627
prognostic sig	55108	32386	BSDC1	NM_018045.5,forward.2	TGGAGACCTGCTGCTTAC	NM_018045.5	141802627
prognostic sig	55108	32386	BSDC1	NM_018045.5,forward.3	GAGAGCTGCTGCTTACGA	NM_018045.5	141802627
prognostic sig	55108	32386	BSDC1	NM_018045.5,reverse.1	AGTGCAGCTGCTGATGACC	NM_018045.5	141802627
prognostic sig	55108	32386	BSDC1	NM_018045.5,reverse.2, 3	CCTGACCCAGCATCCCTTC	NM_018045.5	141802627
prognostic sig	55172	10026	C14orf104	NM_001083908.1,forward.1;	CTCTGCTCCGTACCACT	NM_001083908.1	145580585
prognostic sig	55172	10026	C14orf104	NM_001083908.1,forward.2;	CTCTCTGCTGCTGAGGACT	NM_001083908.1	145580585
prognostic sig	55172	10026	C14orf104	NM_001083908.1,forward.3;	CTTGACCTGAGCGAGAGG	NM_001083908.1	145580585

prognostic sig	55172	10026	C14orf104	NM_001083908.1,reverse.1; NM_018139.2,reverse.1; NM_001083908.1,reverse.2; NM_018139.2,reverse.2; NM_001083908.1,reverse.3; NM_018139.2,reverse.3; NM_001039535.2,forward.3	AGCACCACTGGCAGCGTAC	NM_001083908.1	145580585
prognostic sig	55172	10026	C14orf104	NM_001083908.1,reverse.1; NM_018139.2,reverse.2; NM_001083908.1,reverse.3; NM_018139.2,reverse.3; NM_001039535.2,forward.3	GGAGAACATTCGCCGGA	NM_001083908.1	145580585
prognostic sig	55172	10026	C14orf104	NM_001083908.1,reverse.1; NM_018139.2,reverse.2; NM_001083908.1,reverse.3; NM_018139.2,reverse.3; NM_001039535.2,forward.3	GGTGAGCTCTCGGCGTACT	NM_001083908.1	145580585
prognostic sig	220134	11975	C18orf24	NM_001039535.2,reverse.3 NM_145060.3,forward.1,2, 3;NM_001039535.2,forward.1,2 NM_145060.3,reverse.1; NM_001039535.2,reverse.1 NM_145060.3,reverse.2; NM_001039535.2,reverse.2	TTCCCAACACACAGACTC	NM_001039535.2	154689876
prognostic sig	220134	11975	C18orf24	NM_001039535.2,reverse.3 NM_145060.3,forward.1,2, 3;NM_001039535.2,forward.1,2 NM_145060.3,reverse.1; NM_001039535.2,reverse.1 NM_145060.3,reverse.2; NM_001039535.2,reverse.2	CTTCGGAAGCATGATGTGC	NM_145060.3	154689618
prognostic sig	220134	11975	C18orf24	NM_001039535.2,reverse.3 NM_145060.3,forward.1,2, 3;NM_001039535.2,forward.1,2 NM_145060.3,reverse.1; NM_001039535.2,reverse.1 NM_145060.3,reverse.2; NM_001039535.2,reverse.2	GGAAGCCAGGATGTGAGA	NM_145060.3	154689618
prognostic sig	220134	11975	C18orf24	NM_001039535.2,reverse.3 NM_145060.3,forward.1,2, 3;NM_001039535.2,forward.1,2 NM_145060.3,reverse.1; NM_001039535.2,reverse.1 NM_145060.3,reverse.2; NM_001039535.2,reverse.2	AAGCCAGATGTGAGACG	NM_145060.3	154689618
prognostic sig	220134	11975	C18orf24	NM_001039535.2,reverse.3 NM_145060.3,forward.1,2, 3;NM_001039535.2,forward.1,2 NM_145060.3,reverse.1; NM_001039535.2,reverse.1 NM_145060.3,reverse.2; NM_001039535.2,reverse.2	TTGGAAGCCAGATGGTG	NM_145060.3	154689618
prognostic sig	54069	41294	C21orf45	NM_018944.2,forward.1 NM_018944.2,forward.2 NM_018944.2,forward.3 NM_018944.2,reverse.1 NM_018944.2,reverse.2 NM_018944.2,reverse.3	TAGGTATGGCAGCGTTCG	NM_018944.2	45827699
prognostic sig	54069	41294	C21orf45	NM_018944.2,forward.1 NM_018944.2,forward.2 NM_018944.2,forward.3 NM_018944.2,reverse.1 NM_018944.2,reverse.2 NM_018944.2,reverse.3	AGAGGCCGCTGCTTCT	NM_018944.2	45827699
prognostic sig	54069	41294	C21orf45	NM_018944.2,forward.1 NM_018944.2,forward.2 NM_018944.2,forward.3 NM_018944.2,reverse.1 NM_018944.2,reverse.2 NM_018944.2,reverse.3	GCACTCTGCTTCTGTT	NM_018944.2	45827699
prognostic sig	54069	41294	C21orf45	NM_018944.2,forward.1 NM_018944.2,forward.2 NM_018944.2,forward.3 NM_018944.2,reverse.1 NM_018944.2,reverse.2 NM_018944.2,reverse.3	AGCGAGGATGCTGCATTT	NM_018944.2	45827699
prognostic sig	54069	41294	C21orf45	NM_018944.2,forward.1 NM_018944.2,forward.2 NM_018944.2,forward.3 NM_018944.2,reverse.1 NM_018944.2,reverse.2 NM_018944.2,reverse.3	GCAGGATGCAGTTGTCTC	NM_018944.2	45827699
prognostic sig	54069	41294	C21orf45	NM_018944.2,forward.1 NM_018944.2,forward.2 NM_018944.2,forward.3 NM_018944.2,reverse.1 NM_018944.2,reverse.2 NM_018944.2,reverse.3	TCAGGACGCAACCATTTCC	NM_018944.2	45827699
prognostic sig	56951	10620	C5orf15	NM_020199.2,forward.1,3 NM_020199.2,forward.2 NM_020199.2,reverse.1 NM_020199.2,reverse.2 NM_020199.2,reverse.3 NM_020199.2,reverse.3	TCTCAGGCTTGCAGTCA	NM_020199.2	208431821
prognostic sig	56951	10620	C5orf15	NM_020199.2,forward.1 NM_020199.2,forward.2 NM_020199.2,reverse.1 NM_020199.2,reverse.2 NM_020199.2,reverse.3 NM_020199.2,reverse.3	CGAGCAAGCAAACTGCT	NM_020199.2	208431821
prognostic sig	56951	10620	C5orf15	NM_020199.2,forward.1 NM_020199.2,forward.2 NM_020199.2,reverse.1 NM_020199.2,reverse.2 NM_020199.2,reverse.3 NM_020199.2,reverse.3	CTGCGCTGTGCTAGGCTCT	NM_020199.2	208431821
prognostic sig	56951	10620	C5orf15	NM_020199.2,forward.1 NM_020199.2,forward.2 NM_020199.2,reverse.1 NM_020199.2,reverse.2 NM_020199.2,reverse.3 NM_020199.2,reverse.3	ATAGAGCGCGGACACAAGC	NM_020199.2	208431821
prognostic sig	56951	10620	C5orf15	NM_020199.2,forward.1 NM_020199.2,forward.2 NM_020199.2,reverse.1 NM_020199.2,reverse.2 NM_020199.2,reverse.3 NM_020199.2,reverse.3	AGCCTGCGCTGTGCTAGG	NM_020199.2	208431821
prognostic sig	890	55562	CCNA2	NM_001237.3,forward.1 NM_001237.3,forward.2 NM_001237.3,forward.3 NM_001237.3,reverse.1 NM_001237.3,reverse.2 NM_001237.3,reverse.3	GGTTAGCGCAGCAGCAGAG	NM_001237.3	166197663
prognostic sig	890	55562	CCNA2	NM_001237.3,forward.1 NM_001237.3,forward.2 NM_001237.3,forward.3 NM_001237.3,reverse.1 NM_001237.3,reverse.2 NM_001237.3,reverse.3	GTCTAGCGCAGCAGCAGAG	NM_001237.3	166197663
prognostic sig	890	55562	CCNA2	NM_001237.3,forward.1 NM_001237.3,forward.2 NM_001237.3,forward.3 NM_001237.3,reverse.1 NM_001237.3,reverse.2 NM_001237.3,reverse.3	TGGGTTACCATTCATGTGG	NM_001237.3	166197663
prognostic sig	890	55562	CCNA2	NM_001237.3,forward.1 NM_001237.3,forward.2 NM_001237.3,forward.3 NM_001237.3,reverse.1 NM_001237.3,reverse.2 NM_001237.3,reverse.3	TCCAAGAGGAACGGTGACA	NM_001237.3	166197663
prognostic sig	890	55562	CCNA2	NM_001237.3,forward.1 NM_001237.3,forward.2 NM_001237.3,forward.3 NM_001237.3,reverse.1 NM_001237.3,reverse.2 NM_001237.3,reverse.3	TCCAAGAGGAACGGTGACA	NM_001237.3	166197663
prognostic sig	890	55562	CCNA2	NM_001237.3,forward.1 NM_001237.3,forward.2 NM_001237.3,forward.3 NM_001237.3,reverse.1 NM_001237.3,reverse.2 NM_001237.3,reverse.3	AAAGCCAGGCACTTTCACG	NM_001237.3	166197663
prognostic sig	993	1355	CD25A	NM_201567.1,forward.1; NM_201567.1,forward.2; NM_201567.1,forward.3 NM_201567.1,reverse.1; NM_201567.1,reverse.2; NM_201567.1,reverse.3	GCAACGCTGTCAGCTGCTCT	NM_001789.2	42490757
prognostic sig	993	1355	CD25A	NM_201567.1,forward.1; NM_201567.1,forward.2; NM_201567.1,forward.3 NM_201567.1,reverse.1; NM_201567.1,reverse.2; NM_201567.1,reverse.3	CGGCGATTGACACAGATTCC	NM_001789.2	42490757
prognostic sig	993	1355	CD25A	NM_201567.1,forward.1; NM_201567.1,forward.2; NM_201567.1,forward.3 NM_201567.1,reverse.1; NM_201567.1,reverse.2; NM_201567.1,reverse.3	GTCGCTGCTCTCCCTCA	NM_001789.2	42490757
prognostic sig	993	1355	CD25A	NM_201567.1,forward.1; NM_201567.1,forward.2; NM_201567.1,forward.3 NM_201567.1,reverse.1; NM_201567.1,reverse.2; NM_201567.1,reverse.3	CTCGCTGCTCTCCCTCA	NM_001789.2	42490757
prognostic sig	993	1355	CD25A	NM_201567.1,forward.1; NM_201567.1,forward.2; NM_201567.1,forward.3 NM_201567.1,reverse.1; NM_201567.1,reverse.2; NM_201567.1,reverse.3	TTCTGAAGGCATCCACCTT	NM_001789.2	42490757
prognostic sig	993	1355	CD25A	NM_201567.1,forward.1; NM_201567.1,forward.2; NM_201567.1,forward.3 NM_201567.1,reverse.1; NM_201567.1,reverse.2; NM_201567.1,reverse.3	TGAGGAAGCATCCGAGCTG	NM_001789.2	42490757
prognostic sig	993	1355	CD25A	NM_201567.1,forward.1; NM_201567.1,forward.2; NM_201567.1,forward.3 NM_201567.1,reverse.1; NM_201567.1,reverse.2; NM_201567.1,reverse.3	CAGGAGGAGCTTGATGAT	NM_201567.1	42490759
prognostic sig	993	1355	CD25A	NM_201567.1,forward.1; NM_201567.1,forward.2; NM_201567.1,forward.3 NM_201567.1,reverse.1; NM_201567.1,reverse.2; NM_201567.1,reverse.3	GGGACACCGGTGATGAT	NM_201567.1	42490759
prognostic sig	80321	11387	CE70	NM_024491.2,forward.1	GAGGAGCCGAGGAGACAGA	NM_024491.2	50658081

prognostic sig	80321	11387	CEP70	NM_024491.2,forward,2,3	AGCCGAGGAGACAGATTGC	NM_024491.2	50658081
prognostic sig	80321	11387	CEP70	NM_024491.2,reverse,1,2	TGGCATTCCCTTTGGTGG	NM_024491.2	50658081
prognostic sig	80321	11387	CEP70	NM_024491.2,reverse,3	AGCCCTTGGCATTCTTTTG	NM_024491.2	50658081
prognostic sig	11111	975	CHEK1	NM_001274.4,forward,1	GAACCCCGCACAGGCTTTC	NM_001274.4	166295191
prognostic sig	11111	975	CHEK1	NM_001274.4,forward,2	CCACCTCTCCCTCTCTT	NM_001274.4	166295191
prognostic sig	11111	975	CHEK1	NM_001274.4,forward,3	GTTCCCGGAAAGCATTTGT	NM_001274.4	166295191
prognostic sig	11111	975	CHEK1	NM_001274.4,reverse,1	ACATGTGGCTGGGAAAGC	NM_001274.4	166295191
prognostic sig	11111	975	CHEK1	NM_001274.4,reverse,2	CCTAGGCCCCAGAGGAAT	NM_001274.4	166295191
prognostic sig	11111	975	CHEK1	NM_001274.4,reverse,3	CCAGCACACCGAAGTTGA	NM_001274.4	166295191
prognostic sig	10462	7836	CLEC10A	NM_006344.2,forward,1	ACACCTGATGGGCTCACT	NM_006344.2	53832015
prognostic sig	10462	7836	CLEC10A	NM_006344.2,forward,2,3	GCGCTCCCAACTGTGATGA	NM_006344.2	53832015
prognostic sig	10462	7836	CLEC10A	NM_182906.2,reverse,1	CTGGCTTGCTTCCAGTTCT	NM_006344.2	53832015
prognostic sig	10462	7836	CLEC10A	NM_006344.2,reverse,2	GGTGGAGCTGTGGAGGTCT	NM_006344.2	53832015
prognostic sig	10462	7836	CLEC10A	NM_182906.2,reverse,2	CAGTCTGGGGTGGAGCTGTG	NM_006344.2	53832015
prognostic sig	10462	7836	CLEC10A	NM_006344.2,reverse,3	GGCTTTGCCAAGTCTTCTT	NM_015041.1	52345625
prognostic sig	23059	14831	CLUAP1	NM_024793.1,reverse,2	GAAGCGCTGCTCAAGAGTG	NM_024793.1	13435144
prognostic sig	23059	14831	CLUAP1	NM_024793.1,forward,1	GGAGGACGATGAATCCGACA	NM_024793.1	13435144
prognostic sig	23059	14831	CLUAP1	NM_015041.1,forward,2,3	CAAGCCGCTTTCTTCAACT	NM_024793.1	13435144
prognostic sig	23059	14831	CLUAP1	NM_024793.1,reverse,1	GCCAGTCTTCTTGATGA	NM_024793.1	13435144
prognostic sig	23059	14831	CLUAP1	NM_015041.1,reverse,3	GAACCCAGAGGCTGGAGA	NM_005211.3	195947380
prognostic sig	23059	14831	CLUAP1	NM_005211.3,forward,1	ATGGGTGGCAGGAAGTGAT	NM_005211.3	195947380
prognostic sig	1436	3817	CSF1R	NM_005211.3,forward,2	CCACGAGCTGCTTACAAAC	NM_005211.3	195947380
prognostic sig	1436	3817	CSF1R	NM_005211.3,forward,3	TGCCAGAGCCGTTGTAAT	NM_005211.3	195947380
prognostic sig	1436	3817	CSF1R	NM_005211.3,reverse,1	TCTGCAAGGACCAAGTCA	NM_005211.3	195947380
prognostic sig	1436	3817	CSF1R	NM_005211.3,reverse,2	CCTCACACCACTTCCAAAG	NM_005211.3	195947380
prognostic sig	1436	3817	CSF1R	NM_005211.3,reverse,3	TTGGCTCCCAAGTCTGATTC	NM_001334.2	38327639
prognostic sig	1519	1020	CTSO	NM_001334.2,forward,1	CCATCCCAATGTCCTTGC	NM_001334.2	38327639
prognostic sig	1519	1020	CTSO	NM_001334.2,forward,2	TGGCTCCAGCTTGATTC	NM_001334.2	38327639
prognostic sig	1519	1020	CTSO	NM_001334.2,forward,3	CAGCTGTGGACGAGGCTA	NM_001334.2	38327639
prognostic sig	1519	1020	CTSO	NM_001334.2,reverse,1,3	CCAGATCTCTCACACATC	NM_001334.2	38327639
prognostic sig	1519	1020	CTSO	NM_001334.2,reverse,2	ACCTCTGAACCTGGCCTTG	NM_001337.3	68348719
prognostic sig	1524	20350	CX3CR1	NM_001337.3,forward,1	CCATATGTCCTCCAAACCTG	NM_001337.3	68348719
prognostic sig	1524	20350	CX3CR1	NM_001337.3,forward,2	GGCTCTACTCCCTGCT	NM_001337.3	68348719
prognostic sig	1524	20350	CX3CR1	NM_001337.3,forward,3	GCAATGTGAGGCCCTTTTC	NM_001337.3	68348719
prognostic sig	1524	20350	CX3CR1	NM_001337.3,reverse,1	GGATCCCATGTGTGACGTAT	NM_001337.3	68348719
prognostic sig	1524	20350	CX3CR1	NM_001337.3,reverse,2			

prognostic sig	1524	20350	CXCR1	NM_001337.3.reverse.3	TGCTTTGGCTTCTTG	NM_001337.3	68348719
prognostic sig	57007	22419	CXCR7	NM_020311.2.forward.1	ATGCCCAACAAAGCGTCT	NM_020311.2	114155149
prognostic sig	57007	22419	CXCR7	NM_020311.2.forward.2	CTGGCTCTGCTGCTGACA	NM_020311.2	114155149
prognostic sig	57007	22419	CXCR7	NM_020311.2.forward.3	CCCCCTGCATTCATTCTCTC	NM_020311.2	114155149
prognostic sig	57007	22419	CXCR7	NM_020311.2.reverse.1	ACCCAGCACACGAGAGTT	NM_020311.2	114155149
prognostic sig	57007	22419	CXCR7	NM_020311.2.reverse.2	TGCTGTCTCGGGGTAGAA	NM_020311.2	114155149
prognostic sig	57007	22419	CXCR7	NM_020311.2.reverse.3	GGACACAGCTCTGCTGTT	NM_020311.2	114155149
prognostic sig	10212	68487	DDX39	NM_005804.2.forward.1	ATGACGTCCAGGACCGGTTT	NM_005804.2	21040370
prognostic sig	10212	68487	DDX39	NM_005804.2.forward.2, 3	ACAGGTGACGCTCTGCTGCA	NM_005804.2	21040370
prognostic sig	10212	68487	DDX39	NM_005804.2.reverse.1	ACGTGTGTTACCGGCTCT	NM_005804.2	21040370
prognostic sig	10212	68487	DDX39	NM_005804.2.reverse.2	ACGGTGGCATGACTTGA	NM_005804.2	21040370
prognostic sig	10212	68487	DDX39	NM_005804.2.reverse.3	ACCTTGACGCTGGGCATGTA	NM_005804.2	21040370
prognostic sig	1763	6124	DNA2	NM_001080449.1.forward.1	CGCAGTCAATTTGGGACATC	NM_001080449.1	122937298
prognostic sig	1763	6124	DNA2	NM_001080449.1.forward.2	TCTGTGTACCCCGGTTCT	NM_001080449.1	122937298
prognostic sig	1763	6124	DNA2	NM_001080449.1.forward.3	CGGGCTTTTCTCTTTCT	NM_001080449.1	122937298
prognostic sig	1763	6124	DNA2	NM_001080449.1.reverse.1	CCGGGGGTAAACAGAAAGC	NM_001080449.1	122937298
prognostic sig	1763	6124	DNA2	NM_001080449.1.reverse.2	GCAGCTCAGTCTGTTGAC	NM_001080449.1	122937298
prognostic sig	1763	6124	DNA2	NM_001080449.1.reverse.3	TCACTGCTCCATCTCTGAC	NM_001080449.1	122937298
prognostic sig	29940	8354	DSE	NM_001080976.1.forward.1, 2;	TGCCACGATGAGGACTCACA	NM_001080976.1	124494230
prognostic sig	29940	8354	DSE	NM_013352.2.forward.1, 2	AGTCCCGCTGGAAATGAAT	NM_001080976.1	124494230
prognostic sig	29940	8354	DSE	NM_001080976.1.forward.3;	CTGGGTTCTCTGCTGATG	NM_001080976.1	124494230
prognostic sig	29940	8354	DSE	NM_013352.2.reverse.1;	TGGGTTCTCTGCTGATGT	NM_001080976.1	124494230
prognostic sig	29940	8354	DSE	NM_001080976.1.reverse.2;			
prognostic sig	29940	8354	DSE	NM_001080976.1.reverse.3;			
prognostic sig	29940	8354	DSE	NM_013352.2.reverse.3	CCATGTCGGGCTTCAATG	NM_001080976.1	124494230
prognostic sig	1869	3828	E2F1	NM_005225.2.forward.1	GCTCCAAAGCCGTGACTCTT	NM_005225.2	168480109
prognostic sig	1869	3828	E2F1	NM_005225.2.forward.2	TCAAGCCCTCTGAGAC	NM_005225.2	168480109
prognostic sig	1869	3828	E2F1	NM_005225.2.forward.3	GGTATCAAGCCCTCTCTG	NM_005225.2	168480109
prognostic sig	1869	3828	E2F1	NM_005225.2.reverse.1	TACGGTCTCTCAAGGCACA	NM_005225.2	168480109
prognostic sig	1869	3828	E2F1	NM_005225.2.reverse.2	AAACATGATCGGGCTTGT	NM_005225.2	168480109
prognostic sig	1869	3828	E2F1	NM_005225.2.reverse.3	ACATGATCGGGCTTGT	NM_005225.2	168480109
prognostic sig	79733	11654	E2F8	NM_024680.2.forward.1, 3	GGCTCAAGGCGCAAGCAT	NM_024680.2	38505225
prognostic sig	79733	11654	E2F8	NM_024680.2.forward.2	AATGAGCACTCAAGCAAG	NM_024680.2	38505225
prognostic sig	79733	11654	E2F8	NM_024680.2.reverse.1	GGGAACAAGTTGGGAGAC	NM_024680.2	38505225
prognostic sig	79733	11654	E2F8	NM_024680.2.reverse.2	GGCCCTGTCAGCAACT	NM_024680.2	38505225
prognostic sig	79733	11654	E2F8	NM_024680.2.reverse.3	GGAACAAGTTGGGAGACA	NM_024680.2	38505225
prognostic sig	30844	26430	EHD4	NM_139265.2.forward.1	TAAAGCCGACCAAGTGACA	NM_139265.2	34147619
prognostic sig	30844	26430	EHD4	NM_139265.2.forward.2	TGCTGAATAAGCCGACCA	NM_139265.2	34147619
prognostic sig	30844	26430	EHD4	NM_139265.2.forward.3	GGAGCCACCAAGACTCT	NM_139265.2	34147619
prognostic sig	30844	26430	EHD4	NM_139265.2.reverse.1, 2	GTAACCTGGGCGTTGATG	NM_139265.2	34147619
prognostic sig	30844	26430	EHD4	NM_139265.2.reverse.3	GCTTTTGGGGTCCAGACT	NM_139265.2	34147619
prognostic sig	54465	10369	ETAA1	NM_019002.3.forward.1, 2, 3	AGAAGAAGCGGCTGTGAG	NM_019002.3	89242129
prognostic sig	54465	10369	ETAA1	NM_019002.3.reverse.1	CTGTTTGTGGGGCTTTTC	NM_019002.3	89242129
prognostic sig	54465	10369	ETAA1	NM_019002.3.reverse.2	TGTTTTGTGGGGCTTTCT	NM_019002.3	89242129

prognostic sig	54465	10369	ETAA1	NM_019002.3.reverse.3	CACGTGTTTGTGGCGGTTT	NM_019002.3	89242129
prognostic sig	116496	62170	FAM129A	NM_052966.2.forward.1	ACCTGGGACGCCCTTCTT	NM_052966.2	93277091
prognostic sig	116496	62170	FAM129A	NM_052966.2.forward.2	ATGACCCCGTGTGAATCCT	NM_052966.2	93277091
prognostic sig	116496	62170	FAM129A	NM_052966.2.forward.3	CGTTGGCCAAAGGAGAAAGC	NM_052966.2	93277091
prognostic sig	116496	62170	FAM129A	NM_052966.2.reverse.1	TGCTGACGCGATCCTCAG	NM_052966.2	93277091
prognostic sig	116496	62170	FAM129A	NM_052966.2.reverse.2	TCCTCAAGGCTCCTCCAAA	NM_052966.2	93277091
prognostic sig	116496	62170	FAM129A	NM_052966.2.reverse.3	GGATTCAGCCCAAGATGACC	NM_052966.2	93277091
prognostic sig	2314	11092	FLI1	NM_002018.2.forward.1	CAATGGGATGACCTTGCT	NM_002018.2	22547155
prognostic sig	2314	11092	FLI1	NM_002018.2.forward.2	CGACTTCAAGGGCGGCTACT	NM_002018.2	22547155
prognostic sig	2314	11092	FLI1	NM_002018.2.reverse.1	ATCTGGGCTCTCTCTGCT	NM_002018.2	22547155
prognostic sig	2314	11092	FLI1	NM_002018.2.reverse.2	GGAAGAGTAGGGGGCTTCA	NM_002018.2	22547155
prognostic sig	2314	11092	FLI1	NM_002018.2.reverse.3	GGCAGGTAGCAGAGGCCAGT	NM_002018.2	22547155
prognostic sig	2305	7318	FOXN1	NM_021953.2.forward.1;NM_202002.1.forward.1	CAGGAGACTGGCATTGACG	NM_202003.1	42544160
prognostic sig	2305	7318	FOXN1	NM_202003.1.forward.2;NM_021953.2.forward.2	AACTCATCCGCCACAACCT	NM_202003.1	42544160
prognostic sig	2305	7318	FOXN1	NM_021953.2.forward.3;NM_202003.1.forward.3	CCTCCCGTGTTCCAAGTC	NM_202003.1	42544160
prognostic sig	2305	7318	FOXN1	NM_021953.2.forward.1;NM_021953.2.reverse.1	GCTGAGGGGCAAGTAAAGGA	NM_202003.1	42544160
prognostic sig	2305	7318	FOXN1	NM_021953.2.reverse.2;NM_202003.1.reverse.2	CACCTGGGGTGAATGTCCAG	NM_202003.1	42544160
prognostic sig	2305	7318	FOXN1	NM_021953.2.reverse.3;NM_202003.1.reverse.3	CCCTCCACCCCAATTCAAC	NM_202003.1	42544160
prognostic sig	11337	10508	GABARAP	NM_007278.1.forward.1	GGTTCGCGGAGATGAAGT	NM_007278.1	6005763
prognostic sig	11337	10508	GABARAP	NM_007278.1.forward.2	GGTTCGCGGAGATGAAG	NM_007278.1	6005763
prognostic sig	11337	10508	GABARAP	NM_007278.1.forward.3	GTCCTCGGAGATTCCTTGG	NM_007278.1	6005763
prognostic sig	11337	10508	GABARAP	NM_007278.1.reverse.1;2,3	CCCTCCGCGGATTCCTTGG	NM_013296.4	164519121
prognostic sig	29899	56584	GPSM2	NM_013296.4.forward.1	ATCCCCCAACGATGAGT	NM_013296.4	164519121
prognostic sig	29899	56584	GPSM2	NM_013296.4.forward.2	GAGGAGGGCGGTGTGAGAC	NM_013296.4	164519121
prognostic sig	29899	56584	GPSM2	NM_013296.4.forward.3	CGAAGCCCTGGCAATTACTG	NM_013296.4	164519121
prognostic sig	29899	56584	GPSM2	NM_013296.4.reverse.3	ACGTGGGCCAGAGCTGTA	NM_013296.4	164519121
prognostic sig	26959	8171	HBP1	NM_012257.3.forward.1	GGCATTGCTGAGTACCA	NM_012257.3	47834345
prognostic sig	26959	8171	HBP1	NM_012257.3.forward.2	TACCAAGCATGGGCGAGGT	NM_012257.3	47834345
prognostic sig	26959	8171	HBP1	NM_012257.3.forward.3	CTTGCTGGGCTAGGCTTC	NM_012257.3	47834345
prognostic sig	26959	8171	HBP1	NM_012257.3.reverse.1	ACACCAACACGACCACTG	NM_012257.3	47834345
prognostic sig	26959	8171	HBP1	NM_012257.3.reverse.2	GCCACCTGGTGATCACTC	NM_012257.3	47834345

prognostic sig	3055	20489	HCK	NM_002110.2,forward.1	CGAGCTCTGAGCGTGATA	NM_002110.2	30795228
prognostic sig	3055	20489	HCK	NM_002110.2,forward.2	TCACCATCCACCATCAAGC	NM_002110.2	30795228
prognostic sig	3055	20489	HCK	NM_002110.2,forward.3	TTCCACATCCACCATCAAGC	NM_002110.2	30795228
prognostic sig	3055	20489	HCK	NM_002110.2,reverse.1	TTTTTCAGCAGCGCATCAT	NM_002110.2	30795228
prognostic sig	3055	20489	HCK	NM_002110.2,reverse.2	CAGGGCAACCAAGATGATG	NM_002110.2	30795228
prognostic sig	3055	20489	HCK	NM_002110.2,reverse.3	ACAGGGCAACCAAGATGATG	NM_002110.2	30795228
prognostic sig	51020	5431	HDDC2	NM_016063.2,forward.1	CGGTCCCTACTGCAATTCTTG	NM_016063.2	116875825
prognostic sig	51020	5431	HDDC2	NM_016063.2,forward.2	CCGGAGAGCGTTTCAGATCA	NM_016063.2	116875825
prognostic sig	51020	5431	HDDC2	NM_016063.2,forward.3	TCGGTCCCTACTGCAATTCC	NM_016063.2	116875825
prognostic sig	51020	5431	HDDC2	NM_016063.2,reverse.1, 3	CTCTCGGGCTCTGACATT	NM_016063.2	116875825
prognostic sig	51020	5431	HDDC2	NM_016063.2,reverse.2	CCAGGGCTAGGCGTACACATC	NM_016063.2	116875825
prognostic sig	3689	20092	ITGB2	NM_000211.3,forward.1	CTCCCCGACACCTGAAAGT	NM_000211.3	188595673
prognostic sig	3689	20092	ITGB2	NM_000211.3,forward.2	GTGACTGTGATGGCGTGAG	NM_000211.3	188595673
prognostic sig	3689	20092	ITGB2	NM_000211.3,forward.3	AAGGCTACCCCATCGACTT	NM_000211.3	188595673
prognostic sig	3689	20092	ITGB2	NM_000211.3,reverse.1	CTGCACGCCATCACAAGTCA	NM_000211.3	188595673
prognostic sig	3689	20092	ITGB2	NM_000211.3,reverse.2	CCGGATGACAAAGCACTGCT	NM_000211.3	188595673
prognostic sig	3689	20092	ITGB2	NM_000211.3,reverse.3	GCAAGTGCACCACTAGCTTC	NM_000211.3	188595673
prognostic sig	3812	3812	MAP3K8	NM_005204.2,forward.1	AGGCTGTCTGCTGCTCT	NM_005204.2	22035597
prognostic sig	3812	3812	MAP3K8	NM_005204.2,forward.2	CCCAAGAGCCGACAGCTTAC	NM_005204.2	22035597
prognostic sig	3812	3812	MAP3K8	NM_005204.2,forward.3	CCCAAGCAAGCACTTCCAC	NM_005204.2	22035597
prognostic sig	3812	3812	MAP3K8	NM_005204.2,reverse.1	GGAGGCACATTCATGGGTCA	NM_005204.2	22035597
prognostic sig	3812	3812	MAP3K8	NM_005204.2,reverse.2	GCTCCAAGAGGCGAGAGTCC	NM_005204.2	22035597
prognostic sig	3812	3812	MAP3K8	NM_005204.2,reverse.3	GCGGTATTGGGGTTTCTCT	NM_005204.2	22035597
prognostic sig	4176	4323	MCM7	NM_005916.3,forward.1;	ACCAGGGTGTGTGCTGATT	NM_182776.1	33469921
prognostic sig	4176	4323	MCM7	NM_005916.3,forward.2, 3;	AGATGAGGCGAGAGGCTTGG	NM_182776.1	33469921
prognostic sig	4176	4323	MCM7	NM_005916.3,forward.2, 3	GGAGATGGTGTGCTGCTCCA	NM_182776.1	33469921
prognostic sig	4176	4323	MCM7	NM_005916.3,reverse.1;	ACGTGCCAGACAGTGGAAA	NM_182776.1	33469921
prognostic sig	4176	4323	MCM7	NM_005916.3,reverse.2;	TCAGACGTGCCAGAGCAAGT	NM_182776.1	33469921
prognostic sig	4176	4323	MCM7	NM_005916.3,reverse.3;	GGGGGTCTCAGAGGCAAGT	NM_005590.3	56550106
prognostic sig	4361	4083	MRE11A	NM_005590.3,forward.1	CCCGCGGAATTCAGGTTTAC	NM_005590.3	56550106
prognostic sig	4361	4083	MRE11A	NM_005590.3,forward.2	TCCCGCGGAATTCAGGTTTAC	NM_005590.3	56550106
prognostic sig	4361	4083	MRE11A	NM_005590.3,forward.3	GCGACCCAAAGCTGTCTTCT	NM_005590.3	56550106
prognostic sig	4361	4083	MRE11A	NM_005590.3,reverse.1	AAAGGGCAGGATCCGTGAAA	NM_005590.3	56550106
prognostic sig	4361	4083	MRE11A	NM_005590.3,reverse.2	TCAGAGCAACCAAAAGGA	NM_005591.3	56550105
prognostic sig	4361	4083	MRE11A	NM_005591.3,reverse.1	TTGAGACCTCTCTCTCGATGC	NM_005591.3	56550105
prognostic sig	4361	4083	MRE11A	NM_005591.3,reverse.2	TGAGACCTCTCTCTCGATGC	NM_005591.3	56550105
prognostic sig	4361	4083	MRE11A	NM_005591.3,reverse.3	GAGACCTCTCTCTCGATGCTG	NM_005591.3	56550105

133

134

prognostic sig	25788	8240	RAD54B	NM_012415.2,forward.1	TGCTTCAAGGGTTGTGGA	NM_012415.2	20143928
prognostic sig	25788	8240	RAD54B	NM_012415.2,forward.2	CTTTGACGGCCAGTGTGTT	NM_012415.2	20143928
prognostic sig	25788	8240	RAD54B	NM_012415.2,forward.3	GGAGGCCAGTGTGTTCTGT	NM_012415.2	20143928
prognostic sig	25788	8240	RAD54B	NM_012415.2,reverse.1	TGAACAAAAGGACAGGGGTGA	NM_012415.2	20143928
prognostic sig	25788	8240	RAD54B	NM_012415.2,reverse.2;	CCCTTGCACTGACTTGTTG	NM_012415.2	20143928
prognostic sig	9401	3144	RECQL4	NM_004260.2,forward.1	TCGGGTAGAGCTGGGGTTGT	NM_004260.2	116812615
prognostic sig	9401	3144	RECQL4	NM_004260.2,forward.2	GTCGGGTAGAGCTGGGGTTG	NM_004260.2	116812615
prognostic sig	9401	3144	RECQL4	NM_004260.2,forward.3	CGGGTAGAGCTGGGGTTGTC	NM_004260.2	116812615
prognostic sig	9401	3144	RECQL4	NM_004260.2,reverse.1, 2, 3	GCAATTTGGAGCTTCTCTGT	NM_004260.2	116812615
prognostic sig	5982	6885	RFC2	NM_002914.3,forward.3	CCACCCACTGCTGTAAAGG	NM_002914.3	31563535
prognostic sig	5982	6885	RFC2	NM_002914.3,reverse.1;	GGCAATTTCCCCATCAGAA	NM_002914.3	31563535
prognostic sig	5982	6885	RFC2	NM_181471.1,reverse.1	AGCCCAAGATGCCACAAGTGA	NM_002914.3	31563535
prognostic sig	5982	6885	RFC2	NM_002914.3,reverse.3	CGGTGGCCAGTTAGAGCAGA	NM_181471.1	31563533
prognostic sig	5982	6885	RFC2	NM_181471.1,forward.1;	CGGTGGCCAGTTAGAGCAGA	NM_181471.1	31563533
prognostic sig	5982	6885	RFC2	NM_002914.3,forward.2;	GCCCCGCAAGTGTGAGTT	NM_181471.1	31563533
prognostic sig	5982	6885	RFC2	NM_181471.1,forward.2	CGGCAAGACCAAGCATTC	NM_181471.1	31563533
prognostic sig	5982	6885	RFC2	NM_181471.1,forward.3	CAGATCAGAGACCGCCCTCA	NM_181471.1	31563533
prognostic sig	5982	6885	RFC2	NM_002914.3,reverse.2	TGCCCTGTCAATTTGAAGCA	NM_181471.1	31563533
prognostic sig	5982	6885	RFC2	NM_181471.1,reverse.3			
prognostic sig	5982	6885	RFC2	NM_005785.2,forward.1;			
prognostic sig	5982	6885	RFC2	NM_194358.1,forward.1;			
prognostic sig	5982	6885	RFC2	NM_194359.1,forward.1	CCTGATCACCCAGTGGTTC	NM_005785.2	37588857
prognostic sig	5982	6885	RFC2	NM_005785.2,forward.2;			
prognostic sig	5982	6885	RFC2	NM_194358.1,forward.2;			
prognostic sig	5982	6885	RFC2	NM_194359.1,forward.2	GATGCTGTGCTCAGGCTGT	NM_005785.2	37588857
prognostic sig	5982	6885	RFC2	NM_005785.2,forward.3;			
prognostic sig	5982	6885	RFC2	NM_194358.1,forward.3;			
prognostic sig	5982	6885	RFC2	NM_194359.1,forward.3	AGGCTCTCCACCACCATGTA	NM_005785.2	37588857
prognostic sig	5982	6885	RFC2	NM_005785.2,reverse.1;			
prognostic sig	5982	6885	RFC2	NM_194358.1,reverse.1;			
prognostic sig	5982	6885	RFC2	NM_194359.1,reverse.1	CCGCATGATCCGAGTACTG	NM_005785.2	37588857
prognostic sig	5982	6885	RFC2	NM_005785.2,reverse.2;			
prognostic sig	5982	6885	RFC2	NM_194358.1,reverse.2;			
prognostic sig	5982	6885	RFC2	NM_194359.1,reverse.2	CGTGGGCAATTTCAATCAGC	NM_005785.2	37588857
prognostic sig	5982	6885	RFC2	NM_005785.2,reverse.3;			
prognostic sig	5982	6885	RFC2	NM_194358.1,reverse.3;			
prognostic sig	5982	6885	RFC2	NM_194359.1,reverse.3	CAATCCAGTGGAGCCATGA	NM_005785.2	37588857
prognostic sig	5982	6885	RFC2	NM_001033.3,forward.1	TTTGATTTGTCGCTCTG	NM_001033.3	157266329
prognostic sig	5982	6885	RFC2	NM_001033.3,forward.2, 3	CGTCTGGGGCATTTGCTAT	NM_001033.3	157266329
prognostic sig	5982	6885	RFC2	NM_001033.3,reverse.1	GTTAGTTCCAGGCGCTGCT	NM_001033.3	157266329
prognostic sig	5982	6885	RFC2	NM_001033.3,reverse.2	GCAAGCTGCTCTTCTTTC	NM_001033.3	157266329
prognostic sig	5982	6885	RFC2	NM_001033.3,reverse.3	GGCAGCTGCTCTTCTTTC	NM_001033.3	157266329
prognostic sig	5982	6885	RFC2	NM_005500.1,forward.1	GCTCCCTGAGTATGAGCA	NM_005500.1	4885584
prognostic sig	5982	6885	RFC2	NM_005500.1,forward.2	TTCCCTGTCCCTTCTTCA	NM_005500.1	4885584

136

137

prognostic sig	10024	28697	TROAP	NM_001100620.1.reverse.3	ACGGCCACGATGTAACAGT	NM_001100620.1	154800454
prognostic sig	10024	28697	TROAP	NM_005480.3.forward.1	TGTGGCTCTTGTCTGG	NM_005480.3	154800452
prognostic sig	10024	28697	TROAP	NM_005480.3.forward.2	CCGATCCAGCTGCTTCT	NM_005480.3	154800452
prognostic sig	10024	28697	TROAP	NM_005480.3.forward.3	CCGATCCAGCTGCTTCT	NM_005480.3	154800452
prognostic sig	10024	28697	TROAP	NM_005480.3.reverse.1	GCATGAGTGGAGTCATGG	NM_005480.3	154800452
prognostic sig	10024	28697	TROAP	NM_005480.3.reverse.2	CCGACCAAGAGAGCCACA	NM_005480.3	154800452
prognostic sig	10024	28697	TROAP	NM_005480.3.reverse.3	CTGGCCAGACCAAGAGAGG	NM_005480.3	154800452
prognostic sig	51393	7993	TRPV2	NM_016113.3.forward.1	TTCTGGCCATCGAGTGTA	NM_016113.3	22547178
prognostic sig	51393	7993	TRPV2	NM_016113.3.forward.2	AGCGCAATGGGATCTGCTC	NM_016113.3	22547178
prognostic sig	51393	7993	TRPV2	NM_016113.3.forward.3	GTGCTGATGAAGCTGTGC	NM_016113.3	22547178
prognostic sig	51393	7993	TRPV2	NM_016113.3.reverse.1	TGTGCTGGAAGCCAGTGTA	NM_016113.3	22547178
prognostic sig	51393	7993	TRPV2	NM_016113.3.reverse.2	ATGTAGGCAACAGCGGTGA	NM_016113.3	22547178
prognostic sig	51393	7993	TRPV2	NM_016113.3.reverse.3	GGGCTAGGATTTGCCAGAGT	NM_016113.3	22547178
prognostic sig	10054	4018	UBA2	NM_005499.2.forward.1	TTGCATCGTTTGGCAAGT	NM_005499.2	50592990
prognostic sig	10054	4018	UBA2	NM_005499.2.forward.2	GTGATGCCCGGAAAAAGTG	NM_005499.2	50592990
prognostic sig	10054	4018	UBA2	NM_005499.2.reverse.1	GGCAGCTTCAGGGTCACTC	NM_005499.2	50592990
prognostic sig	10054	4018	UBA2	NM_005499.2.reverse.2	GGAGGGCTGAGCTCCATCAT	NM_005499.2	50592990
prognostic sig	7374	6585	UNG	NM_003362.2.forward.1;	CCGATCCAGAGAACAAAGG	NM_003362.2	19718744
prognostic sig	7374	6585	UNG	NM_003362.2.forward.2;	CTCCGCCAGTTGGAGAAC	NM_003362.2	19718744
prognostic sig	7374	6585	UNG	NM_080911.1.forward.2	GCCTGCTGCTCTTGTTTT	NM_003362.2	19718744
prognostic sig	7374	6585	UNG	NM_003362.2.forward.3;	GCCTGAGGTGCTTCTCCAG	NM_003362.2	19718744
prognostic sig	7374	6585	UNG	NM_080911.1.reverse.1;	CGCTGAGGTGCTTCTCCAG	NM_003362.2	19718744
prognostic sig	7374	6585	UNG	NM_080911.1.reverse.2;	GCTTGGCCACCCAGATAAA	NM_003362.2	19718744
prognostic sig	7374	6585	UNG	NM_080911.1.reverse.3;	CCCTGCAAGGCTGAGACAAG	NM_003362.2	19718744
prognostic sig	7374	6585	UNG	NM_080911.1.reverse.3;	CGCCTGAGGAACACTCCGTA	NM_001017415.1	63053521
prognostic sig	7398	2528	USP1	NM_001017415.1.forward.1	CCACCCGTCGAGGATTACA	NM_001017415.1	63053521
prognostic sig	7398	2528	USP1	NM_001017415.1.forward.2	AATTAGATGGGGGGAGCAG	NM_001017415.1	63053521
prognostic sig	7398	2528	USP1	NM_001017415.1.forward.3;			
prognostic sig	7398	2528	USP1	NM_001017415.1.reverse.1;	CCCTCTTCCGCTCTCATTTG	NM_001017415.1	63053521
prognostic sig	7398	2528	USP1	NM_001017415.1.reverse.2;	CCTCTTCCGCTCTCATTTG	NM_001017415.1	63053521
prognostic sig	7398	2528	USP1	NM_001017415.1.reverse.3;			
prognostic sig	7398	2528	USP1	NM_001017416.1.forward.1;	CCACCCGTCGAGGATTACA	NM_001017416.1	63053523
prognostic sig	7398	2528	USP1	NM_003368.4.forward.1, 2	GGAAACGGCAATTTCAAAGG	NM_003368.4	63053520
prognostic sig	7398	2528	USP1	NM_003368.4.forward.3	CTCTGAGCGTGAGACTGG	NM_003368.4	63053520
prognostic sig	7398	2528	USP1	NM_003368.4.reverse.1	GGCGTTTCTCTCAGTCTCA	NM_003368.4	63053520
prognostic sig	7398	2528	USP1	NM_003368.4.reverse.2	GGCGCTTTTCTCTCAGTCTC	NM_003368.4	63053520
prognostic sig	7398	2528	USP1	NM_003368.4.reverse.3	TCTCAGGACCCCGGTGTTT	NM_003368.4	63053520

prognostic sig	10352	5673	WARS2	NM_015836.3,forward,1; NM_201263.2,forward,1	GCTCCGCTCTCCAGAAAGA	NM_015836.3	126273540
prognostic sig	10352	5673	WARS2	NM_015836.3,forward,2; NM_201263.2,forward,2	TCCCGCTCTCCAGAAAGACA	NM_015836.3	126273540
prognostic sig	10352	5673	WARS2	NM_015836.3,forward,3; NM_015836.3,forward,3	GAAATTCGCGAAGGCTGTGA	NM_201263.2	126273538
prognostic sig	10352	5673	WARS2	NM_201263.2,reverse,1,2; NM_015836.3,reverse,1,2	GCTTCATGATGGCTCCAGGT	NM_201263.2	126273538
prognostic sig	10352	5673	WARS2	NM_201263.2,reverse,3; NM_015836.3,reverse,3	GCATGACCCGCCACTATGTT	NM_201263.2	126273538
prognostic sig	23558	32160	WBP2	NM_012478.3,forward,1	CGCACTGAGCTTGATTGG	NM_012478.3	95089415
prognostic sig	23558	32160	WBP2	NM_012478.3,forward,2	ATGCTCAGGTGGCATCTCA	NM_012478.3	95089415
prognostic sig	23558	32160	WBP2	NM_012478.3,forward,3	CGGCGAGTGATCGTCAATA	NM_012478.3	95089415
prognostic sig	23558	32160	WBP2	NM_012478.3,reverse,1	GGGTCCGAGACTCAAAAG	NM_012478.3	95089415
prognostic sig	23558	32160	WBP2	NM_012478.3,reverse,2	TGGCGGGGATAGACATAGG	NM_012478.3	95089415
prognostic sig	23558	32160	WBP2	NM_012478.3,reverse,3	GGCTTCTGGCACGTTCTTCA	NM_012478.3	95089415
prognostic sig	11169	56019	WDHD1	NM_001008396.1,forward,1,2,3; NM_007086.2,forward,1,2,3	GCCGCTACTAGTGCCCTGCT	NM_001008396.1	56550085
prognostic sig	11169	56019	WDHD1	NM_001008396.1,reverse,1; NM_007086.2,reverse,1	TGACACACAGGTCACGCAA	NM_001008396.1	56550085
prognostic sig	11169	56019	WDHD1	NM_001008396.1,reverse,2; NM_007086.2,reverse,2	TCCTGCCATTGACACACAG	NM_001008396.1	56550085
prognostic sig	11169	56019	WDHD1	NM_001008396.1,reverse,3; NM_007086.2,reverse,3	CCTGCCATTGACACACAGG	NM_001008396.1	56550085
prognostic sig	79968	38573	WDR76	NM_024908.2,forward,1	AGAACGGCAGATGCGATGAA	NM_024908.2	142354445
prognostic sig	79968	38573	WDR76	NM_024908.2,forward,2	AGAACCGCAGATGCGATG	NM_024908.2	142354445
prognostic sig	79968	38573	WDR76	NM_024908.2,forward,3	CAACCCAGAGAACGGCAGA	NM_024908.2	142354445
prognostic sig	79968	38573	WDR76	NM_024908.2,reverse,1,2,3	AAATTCAGGGGCTGGTGGT	NM_024908.2	142354445
prognostic sig	7465	31152	WEE1	NM_003390.2,forward,1	ATCGGTAGCTGCTCTTCC	NM_003390.2	19718775
prognostic sig	7465	31152	WEE1	NM_003390.2,forward,2,3	ATTGTTGAGCTGGGCACTCA	NM_003390.2	19718775
prognostic sig	7465	31152	WEE1	NM_003390.2,reverse,1	TGAGTGCCAGTCCACCAAT	NM_003390.2	19718775
prognostic sig	7465	31152	WEE1	NM_003390.2,reverse,2	CACCGAGGCTGAGGCTCTCT	NM_003390.2	19718775
prognostic sig	7465	31152	WEE1	NM_003390.2,reverse,3	GAGCACCGAGGCTGAGGCTCT	NM_003390.2	19718775
prognostic sig	23503	9102	ZFVE26	NM_015346.3,forward,1	CCAGAGCCTCACTTGCTGGA	NM_015346.3	198041764
prognostic sig	23503	9102	ZFVE26	NM_015346.3,forward,2	ACGCAGGAAAGCCCTGGAC	NM_015346.3	198041764
prognostic sig	23503	9102	ZFVE26	NM_015346.3,forward,3	TGCTGAGGGCTCTTGATGG	NM_015346.3	198041764
prognostic sig	23503	9102	ZFVE26	NM_015346.3,reverse,1	TCCAAGAGGCGCTCACACTC	NM_015346.3	198041764
prognostic sig	23503	9102	ZFVE26	NM_015346.3,reverse,2	ATCTGCAAGCTGTGTGACAA	NM_015346.3	198041764

Table 12. Primer Sequences for Exemplary Normalizing Genes

Types	heEntrez	homID	haga	Primer ID	Primer Sequence	RefSeq Transcript ID	RefSeq Transcript GI ID
normalizer	53616	37898	ADAM22	NM_016351.3, forward.1, 2, 3	GCTCTGGCTGGTGGACCT	NM_016351.3	21536382
normalizer	53616	37898	ADAM22	NM_016351.3, reverse.1	CACCAAGCGTTTGCTTCC	NM_016351.3	21536382
normalizer	53616	37898	ADAM22	NM_016351.3, reverse.2	CCACGAAGCGGTTTCTTCC	NM_016351.3	21536382
normalizer	53616	37898	ADAM22	NM_016351.3, reverse.3	TCACGAAGCGGTTTCTTCC	NM_016351.3	21536382
normalizer	799	1320	CALCR	NM_001742.2, forward.1 and 3	TCACATGATGCGGGTGCT	NM_001742.2	46361988
normalizer	799	1320	CALCR	NM_001742.2, forward.2	CTCAACATGTTCCGGTGCT	NM_001742.2	46361988
normalizer	799	1320	CALCR	NM_001742.2, reverse.1 and 2	GGATCATGTTGGCTTACAA	NM_001742.2	46361988
normalizer	799	1320	CALCR	NM_001742.2, reverse.3	ATATGATGCTGCTTACAGC	NM_001742.2	46361988
normalizer	1395	55612	CHRR2	NM_001883.3, forward.1	TCACACGACGGATTCAGTT	NM_001883.3	155523258
normalizer	1395	55612	CHRR2	NM_001883.3, reverse.1	AGAGACGACAGGAGAGCTGT	NM_001883.3	155523258
normalizer	1395	55612	CHRR2	NM_001883.3, forward.2	TCAGCCGTAGAGGAAGAGTG	NM_001883.3	155523258
normalizer	1395	55612	CHRR2	NM_001883.3, reverse.2	TCGCTGGGTGATGTAGAGG	NM_001883.3	155523258
normalizer	1395	55612	CHRR2	NM_001883.3, forward.3	TCCTTCTCGAGTGTGTCA	NM_001883.3	155523258
normalizer	1395	55612	CHRR2	NM_001883.3, reverse.3	CACCTTCTCTACGGCTGA	NM_001883.3	155523258
normalizer	1538	79613	CYLC1	NM_021118.1, forward.1	ACACACACTCTCTGTGTG	NM_021118.1	117935037
normalizer	1538	79613	CYLC1	NM_021118.1, reverse.1	TGGAGGTGGAGAGGACATC	NM_021118.1	117935037
normalizer	1538	79613	CYLC1	NM_021118.1, forward.2	GAAGGCCACGAACACAGTCA	NM_021118.1	117935037
normalizer	1538	79613	CYLC1	NM_021118.1, forward.3	AGAGGCCAGAAACACAGTCA	NM_021118.1	117935037
normalizer	1538	79613	CYLC1	NM_021118.1, reverse.2 and 3	CTGTTCCTCTGGACGCTGT	NM_021118.1	117935037
normalizer	1538	79613	CYLC1	NM_00105610.1, forward.1	AGCCATCCATCCGACTCATCT	NM_00105610.1	54112102
normalizer	1896	68180	EDA	NM_00105610.1, forward.2	CCGATCCGACGATCAGAGT	NM_00105610.1	54112102
normalizer	1896	68180	EDA	NM_00105610.1, forward.3	CCCATTCACAGTCACTTGG	NM_00105610.1	54112102
normalizer	1896	68180	EDA	NM_00105610.1, forward.2	CCGCTCTCTGGGAGTCAAGT	NM_00105610.1	54112102
normalizer	1896	68180	EDA	NM_00105610.1, reverse.1, 2 and 3	TTGGGGCAAGGAGATGAGCA	NM_00105610.1	54112099
normalizer	1896	68180	EDA	NM_001399.4, reverse.1	TTTGTGGCAGCTGGCTGT	NM_001399.4	54112099
normalizer	1896	68180	EDA	NM_001399.4, reverse.2	TTTGTGGCAGCTGGCTGT	NM_001399.4	54112099
normalizer	1896	68180	EDA	NM_001399.4, forward.2	AGCCCATCCGACTCATCT	NM_001399.4	54112099
normalizer	1896	68180	EDA	NM_001399.4, forward.3	CCGCTCTGGGAGTCAAGT	NM_001399.4	54112099
normalizer	1896	68180	EDA	NM_000406.2, forward.1	CATCAGACCCCAAGCACT	NM_000406.2	61676182
normalizer	1896	68180	EDA	NM_000406.2, forward.2	CAGGACCCCAAGCACTACA	NM_000406.2	61676182
normalizer	1896	68180	EDA	NM_000406.2, forward.3	GTGGCAATGCACTGCAT	NM_000406.2	61676182
normalizer	1896	68180	EDA	NM_000406.2, reverse.2	TGGCAACAGTGCCTCTCT	NM_000406.2	61676182
normalizer	1896	68180	EDA	NM_000406.2, reverse.1	GGTCAAGTGGGAGGTTGC	NM_000406.2	61676182
normalizer	1896	68180	EDA	NM_000457.3, forward.1	AACTCTGTGCATGTAGACA	NM_000457.3	31077204
normalizer	1896	68180	EDA	NM_000457.3, forward.2	CCATGTCACTGCACTGACA	NM_000457.3	31077204
normalizer	1896	68180	EDA	NM_000457.3, reverse.1	CGATGTCACTGCACTGACA	NM_000457.3	31077204
normalizer	1896	68180	EDA	NM_000457.3, forward.2	TGCTCTCTCTGCTGCTGT	NM_000457.3	31077204
normalizer	1896	68180	EDA	NM_000457.3, forward.2	GGCTCTCTGAGAGTGCTGT	NM_000457.3	31077204
normalizer	1896	68180	EDA	NM_000457.3, reverse.2	CGTCTCTCTGAGAGTGCTGT	NM_000457.3	31077208
normalizer	1896	68180	EDA	NM_000457.3, forward.2 and 3	CACCAAGCTGCAAGATG	NM_000457.3	31077208
normalizer	1896	68180	EDA	NM_000457.3, reverse.2 and 3	CACCAAGCTGCAAGATG	NM_000457.3	31077208
normalizer	1896	68180	EDA	NM_178850.1, reverse.2	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3	GCCTCTCTGTCATCACT	NM_178850.1	31077208
normalizer	1896	68180	EDA	NM_178850.1, forward.3			

normalizer	4771	2180	NP2	NM_181825.2,reverse,3	TCAGTCTTGCCATT	NM_181825.2	163644291
normalizer	4771	2180	NP2	NM_181833.2,reverse,2	GACGTTCACCGTGAGATCG	NM_181833.2	163644290
normalizer	4771	2180	NP2	NM_181833.2,reverse,2	GTCCGGCACCAATCAAA	NM_181833.2	163644290
normalizer	4887	701	NP2R	NM_000910.2,forward,1	TGCTGAGACGGTGGGAAA	NM_000910.2	27552771
normalizer	4887	701	NP2R	NM_000910.2,reverse,1	CATAAGCAAGCGGAGCA	NM_000910.2	27552771
normalizer	4887	701	NP2R	NM_000910.2,forward,2	GAGGTGTGACCTGCTGCT	NM_000910.2	27552771
normalizer	4887	701	NP2R	NM_000910.2,reverse,2	TGCGATGAAGAGACCTGT	NM_000910.2	27552771
normalizer	4887	701	NP2R	NM_000910.2,reverse,2	ATCGACAGACGACGTGCAC	NM_000910.2	27552771
normalizer	4887	701	NP2R	NM_000910.2,reverse,3	GCTTCAACCAAGCAAGAG	NM_000910.2	27552771
normalizer	4887	701	NP2R	NM_003466.3,forward,1 and 3	CTGGGAGGTGAGTGTGCTG	NM_003466.3	81295799
normalizer	7849	2589	PAX8	NM_003466.3,reverse,1	CTTGGCAGAGCTGCAGGA	NM_003466.3	81295799
normalizer	7849	2589	PAX8	NM_003466.3,reverse,3	TGGCAGAGGTGTCAAGAG	NM_003466.3	81295799
normalizer	7849	2589	PAX8	NM_003466.3,forward,2	CGATCCCTTGCTTACAGC	NM_003466.3	81295799
normalizer	7849	2589	PAX8	NM_003466.3,reverse,2	CAGCCCAAGCCCTTATTG	NM_003466.3	81295799
normalizer	10846	4852	PDE10A	NM_006661.2,forward,1	CGCAAGGAGCTGTGATTG	NM_006661.2	194595489
normalizer	10846	4852	PDE10A	NM_006661.2,forward,2	GCAGAAAGGAGCTGTGAGCA	NM_006661.2	194595489
normalizer	10846	4852	PDE10A	NM_006661.2,reverse,1 and 2	GCCAGAGGTGTGGAATT	NM_006661.2	194595489
normalizer	10846	4852	PDE10A	NM_006661.2,forward,3	AACTGGAAGCATGCGTAC	NM_006661.2	194595489
normalizer	10846	4852	PDE10A	NM_006661.2,reverse,3	TCCTTTGCGTCAAGGCTG	NM_006661.2	194595489
normalizer	6005	68045	RHAG	NM_000324.2,forward,1	GCCATTGCTGAACTGGAG	NM_000324.2	156627564
normalizer	6005	68045	RHAG	NM_000324.2,forward,3	AACTGGCCATTGCTGAAC	NM_000324.2	156627564
normalizer	6005	68045	RHAG	NM_000324.2,reverse,1 and 3	GAAAGCAAGGCTGTGAGCA	NM_000324.2	156627564
normalizer	6005	68045	RHAG	NM_000324.2,reverse,2	TCCTCAGACGCTTGCTTC	NM_000324.2	156627564
normalizer	6005	68045	RHAG	NM_000324.2,reverse,2	CAGCAACTCTCCAGCAAG	NM_000324.2	156627564
normalizer	6005	68045	RHAG	NM_052828.2,forward,1	ACTGGCAGCTGGCTAACGTG	NM_052828.2	157426895
normalizer	6005	68045	RHAG	NM_052828.2,forward,2	CAACTGGCAGCTGGCTAACG	NM_052828.2	157426895
normalizer	6005	68045	RHAG	NM_052828.2,reverse,1 and 2	TCTCGGTGCTTGGCAGAC	NM_052828.2	157426895
normalizer	6005	68045	RHAG	NM_052828.2,forward,3	GCTCTGCTGCTGTGTGAC	NM_052828.2	157426895
normalizer	10107	4932	TRIM10	NM_006778.3,forward,1	ACTGGCAGCTGGCTAACGTG	NM_006778.3	157426897
normalizer	10107	4932	TRIM10	NM_052828.2,forward,2	TTGTGGCCGAGTCGATAGT	NM_052828.2	157426895
normalizer	10107	4932	TRIM10	NM_052828.2,reverse,1 and 2	GGCTCCCACTGCTCAATCT	NM_052828.2	157426895
normalizer	10107	4932	TRIM10	NM_052828.2,forward,1	TATGGCTCCCACTGCTCAA	NM_052828.2	157426895
normalizer	10107	4932	TRIM10	NM_033229.2,forward,1	GTTCAGGAGGAGTGGTTTC	NM_033229.2	149193330
normalizer	89870	23819	TRIM15	NM_033229.2,forward,2	TTATGGCTCCCACTGCTCA	NM_033229.2	149193330
normalizer	89870	23819	TRIM15	NM_033229.2,forward,3	GTTCAGGAGGAGTGGTTTC	NM_033229.2	149193330
normalizer	89870	23819	TRIM15	NM_033229.2,reverse,1-2-3	AACTCTGGCTCCCAAGG	NM_033229.2	149193330
normalizer	89870	23819	TRIM15	NM_015898.2,reverse,1	CCTGGCATCTGGCGCATAG	NM_015898.2	40549452
normalizer	89870	23819	TRIM15	NM_015898.2,reverse,1	AGCAGGAGCTGGGTGTTT	NM_015898.2	40549452
normalizer	89870	23819	TRIM15	NM_015898.2,forward,2	ACAGAGGAGGCGCTAAGAC	NM_015898.2	40549452
normalizer	89870	23819	TRIM15	NM_015898.2,forward,2	CCTTCAGAAAGTGCCTCATC	NM_015898.2	40549452
normalizer	89870	23819	TRIM15	NM_015898.2,forward,3	TGCACTGTAGGGCTCTCTG	NM_015898.2	40549452
normalizer	89870	23819	TRIM15	NM_015898.2,reverse,3		NM_015898.2	40549452

[0366] The expression level of FOXM1 was significantly associated with tumor-free survival in multiple human cancer cohorts which confirms that reduced FOXM1 level lead longer tumor-free survival (Table 10). The B cell interaction network data provide an independent evidence that FOXM1 is a highly connected “hub” gene in the pathway represented by the 136 signature. The inventors show that FOXM1 acts as a regulator of the 136 gene signature. The transcriptional changes of the 136 genes associated with the 6C1-G3 deletions in the mouse lymphomas were in part mediated by down-regulation of FOXM1. Furthermore, findings that the expression level of FOXM1 was significantly associated with tumor-free survival in multiple human cancer cohorts suggest that reduced FOXM1 level lead to delayed proliferation, and thus longer tumor-free survival (Table 10). Here, we provide evidence linking discrete chromosome lesions with global transcriptional deregulation targeting key pathways in cancer. Using a well-defined mouse lymphoma model to perform integrative oncogenomic analyses, we found recurrent Chromosome 6 deletions, corresponding to major lymphoma/leukemia related regions of the human genome that significantly correlated with tumor-free survival. This suggests that FOXM1 is useful as prognostic marker for tumor-free survival.

[0367] The expression level of RECQL4 is significantly associated with tumor-free survival in multiple human cancer cohorts which confirms that reduced RECQL4 level is correlated with longer tumor-free survival (Table 10). RECQL4 is useful as prognostic marker for tumor-free survival.

Example 12: Cancer Typing

[0368] For each human B lymphoma cohort (comprising both Diffuse Large B-Cell Lymphoma (DLBCL) and Burkitt’s Lymphoma (BL)) (Dave et al, 2006, Hummel et al, 2006), the gene expression patterns of 136 genes were grouped based on unsupervised hierarchical clustering. The clustering results effectively distinguished BL from DLBCL, and performed better than conventional molecular diagnostic methods. This demonstrates the diagnostic utility of the 136 gene signature and is depicted herein in Figure 12.

REFERENCES

[0134] The references cited herein and throughout the application are incorporated herein by reference.

1. Adler, AS et al. (2006) *Nat Genet* 38:421–430.
2. Aitman T.J., Dong R., Vyse T.J., Norsworthy P.J., Johnson M.D., Smith J., Mangion J., Robertson-Lowe C., Marshall A.J., Petretto E., et al. *Nature* (2006) 439:851–855.
3. Akilesh, S., Shaffer, D.J., and Roopenian, D. (2003). *Genome Res* 13, 1719-1727.
4. Armstrong, S.A., Golub, T.R., and Korsmeyer, S.J. (2003). *Seminars in Hematology* 40, 268-273.
5. Artandi, S.E., and DePinho, R.A. (2000) *Nature Medicine* 6, 852-855.
6. Albertson, DG, Collins, C, McCormick, F, Gray, JW (2003) *Nat Genet* 34:369–376.
7. Bahassi, E.M., Penner, C.G., Robbins, S.B., Tichy, E., Feliciano, E., Yin, M., Liang, L., Deng, L., Tischfield, J.A., and Stambrook, P.J. (2007). *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 616, 201-209.
8. Barnes, D.E., Stamp, G., Rosewell, I., Denzel, A., and Lindahl, T. (1998). *Current Biology* 8, 1395-1398.
9. Barrett, M.T., Scheffer, A., Ben-Dor, A., Sampas, N., Lipson, D., Kincaid, R., Tsang, P., Curry, B., Baird, K., Meltzer, P.S., et al. (2004). *Proc Natl Acad Sci U S A* 101, 17765-17770.
10. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. (2007). *Nucleic Acids Res* 35: D760–765.
11. Bassing, C.H., Swat, W., and Alt, F.W. (2002). *Cell* 109.
12. Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Fa vera, R., and Califano, A. (2005). *Nat Genet* 37, 382–390.
13. Beckman, R.A., and Loeb, L.A. (2005). *Seminars in Cancer Biology* 15, 423-435.
14. Bennett, L.M., McAllister, K.A., Blackshear, P.E., Malphurs, J., Goulding, G., Collins, N.K., Ward, T., Bunch, D.O., Eddy, E.M., Davis, B.J., et al. (2000). *Molecular Carcinogenesis* 28, 174-183.
15. Bhattacharjee, A, Richards, W. G, Staunton, J, Li, C, Monti, S, Vasa, P, Ladd, C, Beheshti, J, Bueno, R, Gillette, M, Loda, M, Weber, G, Mark, E. J, Lander, E. S, Wong, W, Johnson, B. E, Golub, T. R, Sugarbaker, D. J, & Meyerson, M. (2001) *Proc Natl Acad Sci U S A* 98, 13790–13795.

16. Bild, A. H, Yao, G, Chang, J. T, Wang, Q, Potti, A, Chasse, D, Joshi, M.-B, Harpole, D, Lancaster, J. M, Berchuck, A, Olson, J. A, Marks, J. R, Dressman, H. K, West, M, & Nevins, J. R. (2006) *Nature* 439, 353–357.
17. Biondi, A., Cimino, G., Pieters, R., and Pui, C.H. (2000). *Blood* 96, 24-33.
18. Bosma, G.C., Fried, M., Custer, R.P., Carroll, A., Gibson, D.M., and Bosma, M.J. (1988). *J Exp Med* 167, 1016-1033.
19. Brennan, C., Zhang, Y., Leo, C., Feng, B., Cauwels, C., Aguirre, A.J., Kim, M., Protopopov, A., and Chin, L. (2004). *Cancer Research* 64, 4744-4748.
20. Brown, A.C., Olver, W.I., Donnelly, C.J., May, M.E., Naggert, J.K., Shaffer, D.J., and Roopenian, D.C. (2005).. *BMC Genet* 6, 12.
21. Bussey, K.J., Chin, K., Lababidi, S., Reimers, M., Reinhold, W.C., Kuo, W.L., Gwadry, F., Ajay, Kouroos-Mehr, H., Fridlyand, J., et al. (2006). *Molecular Cancer Therapeutics* 5, 853-867.
22. Carroll, W.L., Bhojwani, D., Min, D.J., Raetz, E., Relling, M., Davies, S., Downing, J.R., Willman, C.L., and Reed, J.C. (2003). *Hematology (Am Soc Hematol Educ Program)*, 102-131.
23. Castillo, A., Morse Iii, H.C., Godfrey, V.L., Naeem, R., and Justice, M.J. (2007). *Cancer Research* 67, 10138-10147.
24. Celeste, A., Difilippantonio, S., Difilippantonio, M.J., Fernandez-Capetillo, O., Pilch, D.R., Sedelnikova, O.A., Eckhaus, M., Ried, T., Bonner, W.M., and Nussenzweig, A. (2003). *Cell* 114, 371-383.
25. Chu, S et al. (1998) *Science* 282:699–705.
26. Cui, X., and Churchill, G.A. (2003). *Genome Biology* 4.
27. Daily, J. P, Scanfeld, D, Pochet, N, Roch, K. L, Plouffe, D, Kamal, M, Sarr, O, Mboup, S, Ndir, O, Wypij, D, Levasseur, K, Thomas, E, Tamayo, P, Dong, C, Zhou, Y, Lander, E. S, Ndiaye, D, Wirth, D, Winzeler, E. A, Mesirov, J. P, & Regev, A. (2007) *Nature* 450, 1091–1095.
28. Dave, S.S., Fu, K., Wright, G.W., Lam, L.T., Kluin, P., Boerma, E.-J., Greiner, T.C., Weisenburger, D.D., Rosenwald, A., Ott, G., et al. (2006). *N Engl J Med* 354, 2431–2442.
29. Desmedt, C et al. (2007) *Clin Cancer Res* 13:3207–3214
30. Donehower, L.A., Harvey, M., Slagle, B.L., McArthur, M.J., Montgomery Jr, C.A., Butel, J.S., and Bradley, A. (1992). *Nature* 356, 215-221.
31. Evers, B., and Jonkers, J. (2006). *Oncogene* 25, 5885-5897.

32. Fanciulli, M. et al. *Nat. Genet.* 39, 721–723 (2007).
33. Ferber, M.J., Eilers, P., Schuurin, E., Fenton, J.A.L., Fleuren, G.J., Kenter, G., Szuhai, K., Smith, D.I., Raap, A.K., and Brink, A.A.T.P. (2004). *Cancer Genetics and Cytogenetics* 154, 1-9.
34. Frank, K.M., Sharpless, N.E., Gao, Y., Sekiguchi, J.M., Ferguson, D.O., Zhu, C., Manis, J.P., Horner, J., DePinho, R.A., and Alt, F.W. (2000). *Molecular Cell* 5, 993-1002.
35. Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. (2006). *Genome Research* 16, 949-961.
36. Freije, W. A, Castro-Vargas, F. E, Fang, Z, Horvath, S, Cloughesy, T, Liao, L. M, Mischel, P. S, & Nelson, S. F. (2004) *Cancer Res* 64, 6503–6510.
37. Friedberg, E.C., and Meira, L.B. (2004). *DNA Repair* 3, 1617-1638.
38. Gao, Y., Chaudhuri, J., Zhu, C., Davidson, L., Weaver, D.T., and Alt, F.W. (1998a). *A Immunity* 9, 367-376.
39. Gao, Y., Ferguson, D.O., Xie, W., Manis, J.P., Sekiguchi, J., Frank, K.M., Chaudhuri, J., Horner, J., DePinho, R.A., and Alt, F.W. (2000). *Nature* 404, 897-900.
40. Gao, Y., Sun, Y., Frank, K.M., Dikkes, P., Fujiwara, Y., Seidl, K.J., Sekiguchi, J.M., Rathbun, G.A., Swat, W., Wang, J., et al. (1998b). *Cell* 95, 891-902.
41. Garraway, LA et al. (2005) *Nature* 436:117–122.
42. Gilfillan, S., Dierich, A., Lemeur, M., Benoist, C., and Mathis, D. (1993). *Science* 261, 1175-1178.
43. Gladdy, R.A., Taylor, M.D., Williams, C.J., Grandal, I., Karaskova, J., Squire, J.A., Rutka, J.T., Guidos, C.J., and Danska, J.S. (2003). *Cancer Cell* 3, 37-50.
44. Glinsky, GV, Krones-Herzig, A, Glinskii, AB (2003) *Neoplasia* 5:218–228.
45. Gonzalez E., Kulkarni H., Bolivar H., Mangano A., Sanchez R., Catano G., Nibbs R.J., Freedman B.I., Quinones M.P., Bamshad M.J., et al. *Science* (2005) 307:1434–1440
46. Grabher, C., and Look, A.T. (2006). *Nat Biotech* 24, 45-46.
47. Gu, Y., Seidl, K.J., Rathbun, G.A., Zhu, C., Manis, J.P., van der Stoep, N., Davidson, L., Cheng, H.-L., Sekiguchi, J.M., Frank, K., et al. (1997). *Immunity* 7, 653-665.
48. Hanahan, D & Weinberg, R. A. (2000) *Cell* 100, 57–70.
49. Hart, G.T., Shaffer, D., Akilesh, S., Brown, A.C., Moran, L., and Roopenian, D.C. (2004). *Infect Immun* 72, 4471–4479.

50. Hartigan, JA, Wong., MA (1979) *Applied Statistics* 28(1):100–108.
51. Heidenblad, M et al. (2005) *Oncogene* 24:1794–1801.
52. Hill, A., Harada, Y., Takahashi, E., Hou, J., Wagner, M.J., and Wells, D.E. (1997). *Cytogenetics and Cell Genetics* 78, 56-57.
53. Hollander, M.C., Sheikh, M.S., Bulavin, D.V., Lundgren, K., Augeri-Henmueller, L., Shehee, R., Molinaro, T.A., Kim, K.E., Tolosa, E., Ashwell, J.D., et al. (1999). *Nature Genetics* 23, 176-184.
54. Hori, T.A., Seki, N., Ohira, M., Saito, T., Yamauchi, M., Sagara, M., Hayashi, A., Tsuji, S., Ito, H., and Imai, T. (1998) *Cancer Genetics and Cytogenetics* 101, 24-34.
55. Houghtaling, S., Timmers, C., Noll, M., Finegold, M.J., Jones, S.N., Stephen Meyn, M., and Grompe, M. (2003). *Genes and Development* 17, 2021-2035.
56. Hummel, M., Bentink, S., H. Berger, W. Klapper, S. Wessendorf, T. F. E. Barth, H.-W. Bernd, Cogliatti, S.B., J. Dierlamm, A. C. Feller, et al. (2006) *N Engl J Med* 354, 2419–2430.
57. Jacks, T., Remington, L., Williams, B.O., Schmitt, E.M., Halachmi, S., Bronson, R.T., and Weinberg, R.A. (1994). *Current Biology* 4, 1-7.
58. Kent, W.J. (2002). *Genome Research* 12, 656-664.
59. Kerr, M.K., and Churchill, G.A. (2001). *Biostatistics* 2, 183–201.
60. Kim, M et al. (2006) *Cell* 125:1269–1281.
61. Kurimasa, A., Ouyang, H., Dong, L.-j., Wang, S., Li, X., Cordon-Cardo, C., Chen, D.J., and Li, G.C. (1999). *Proceedings of the National Academy of Sciences* 96, 1403-1408.
62. Lee, H, Kong, SW, Park, PJ (2008) *Bioinformatics* 24:889–896.
63. Lee, HK, Hsu, AK, Sajdak, J, Qin, J, Pavlidis, P (2004) *Genome Res* 14:1085–1094.
64. Lee, J.A., and Lupski, J.R. (2006). *Neuron* 52, 103-121.
65. Lehner, B, Crombie, C, Tischler, J, Fortunato, A, Fraser, AG (2006) *Nat Genet* 38:896–903.
66. Lim, D.S., and Hasty, P. (1996). *Mol Cell Biol* 16, 7133-7143.
67. Liu, X., Holstege, H., Van Der Gulden, H., Treur-Mulder, M., Zevenhoven, J., Velds, A., Kerkhoven, R.M., Van Vliet, M.H., Wessels, L.F.A., Petersen, J.L., et al. (2007). *Proceedings of the National Academy of Sciences of the United States of America* 104, 12111-12116.
68. Lu, Y et al. (2007) *Genome Biol* 8:R146.

69. Luo, G., Yao, M.S., Bender, C.F., Mills, M., Bladl, A.R., Bradley, A., and Petrini, J.H.J. (1999). *Proceedings of the National Academy of Sciences of the United States of America* 96, 7376-7381.
70. Kim, M et al. (2006) *Cell* 125:1269–1281.
71. M. Scott Lucia, D.G.B.M.B.A.T.K.C.D.W.K.I.L.M.P.C.R.-S.T.S.B.A.W. (1998). *The Prostate* 36, 49-55.
72. Mak, T.W., Hakem, A., McPherson, J.P., Shehabeldin, A., Zablocki, E., Migon, E., Duncan, G.S., Bouchard, D., Wakeham, A., Cheung, A., et al. (2000). *Nature Immunology* 1, 77-82.
73. Maser, RS et al. (2007) *Nature* 447:966–971.
74. Mayhew, C.N., Bosco, E.E., Fox, S.R., Okaya, T., Tarapore, P., Schwemberger, S.J., Babcock, G.F., Lentsch, A.B., Fukasawa, K., and Knudsen, E.S. (2005). *Cancer Res* 65, 4568-4577.
75. Mayhew, C.N., Carter, S.L., Fox, S.R., Sexton, C.R., Reed, C.A., Srinivasan, S.V., Liu, X., Wikenheiser-Brokamp, K., Boivin, G.P., Lee, J., et al. (2007). *Gastroenterology* 133, 976-984.
76. McCarroll, S.A., and Altshuler, D.M. (2007). *Nature Genetics* 39.
77. McCarroll, SA et al. (2004). *Nat Genet* 36:197–204.
78. McCarthy, E.E., Celebi, J.T., Baer, R., and Ludwig, T. (2003). *Mol Cell Biol* 23, 5056-5063.
79. Miller, L. D, Smeds, J, George, J, Vega, V. B, Vergara, L, Ploner, A, Pawitan, Y, Hall, P, Klaar, S, Liu, E. T, & Bergh, J. (2005). *Proc Natl Acad Sci U S A* 102, 13550–13555.
80. Mills, K.D., Ferguson, D.O., and Alt, F.W. (2003). *Immunological Reviews* 194, 77-95.
81. Moorman, A.V., Raimondi, S.C., Pui, C.H., Baruchel, A., Biondi, A., Carroll, A.J., Forestier, E., Gaynon, P.S., Harbott, J., Harms, D.O., et al. (2005). *Leukemia* 19, 557-563.
82. Morales, J.C., Franco, S., Murphy, M.M., Bassing, C.H., Mills, K.D., Adams, M.M., Walsh, N.C., Manis, J.P., Rassidakis, G.Z., Alt, F.W., et al. (2006). *Proceedings of the National Academy of Sciences of the United States of America* 103, 3310-3315.
83. Moshous, D., Pannetier, C., De Chasseval, R., Le Deist, F., Cavazzana-Calvo, M., Romana, S., Macintyre, E., Canioni, D., Brousse, N., Fischer, A., et al. (2003). *Journal of Clinical Investigation* 111, 381-387.

84. Nigro, J.M., Misra, A., Zhang, L., Smirnov, I., Colman, H., Griffin, C., Ozburn, N., Chen, M., Pan, E., Koul, D., et al. (2005). *Cancer Research* 65, 1678-1686.
85. Niida, H., Matsumoto, T., Satoh, H., Shiwa, M., Tokutake, Y., Furuichi, Y., and Shinkai, Y. (1998). *Nature Genetics* 19, 203-206.
86. Niimi, N., Sugo, N., Aratani, Y., and Koyama, H. (2005). *Cell Death Differ* 12, 184-191.
87. Noll, M., Battaile, K.P., Bateman, R., Lax, T.P., Rathbun, K., Reifsteck, C., Bagby, G., Finegold, M., Olson, S., and Grompe, M. (2002). *Experimental Hematology* 30, 679-688.
88. Nussenzweig, A., Chen, C., da Costa Soares, V., Sanchez, M., Sokol, K., Nussenzweig, M.C., and Li, G.C. (1996). *Nature* 382, 551-555.
89. Nutt, C. L, Mani, D. R, Betensky, R. A, Tamayo, P, Cairncross, J. G, Ladd, C, Pohl, U, Hartmann, C, McLaughlin, M. E, Batchelor, T. T, Black, P. M, von Deimling, A, Pomeroy, S. L, Golub, T. R, & Louis, D. N. (2003) *Cancer Res* 63, 1602–1607.
90. Okazaki, I.M., Hiai, H., Kakazu, N., Yamada, S., Muramatsu, M., Kinoshita, K., and Honjo, T. (2003). *CJournal of Experimental Medicine* 197, 1173-1181.
91. Ouyang, H., Nussenzweig, A., Kurimasa, A., Soares, V.d.C., Li, X., Cordon-Cardo, C., Li, W.-h., Cheong, N., Nussenzweig, M., Iliakis, G., et al. (1997). *J Exp Med* 186, 921-929.
92. Pamidi, A., Cardoso, R., Hakem, A., Matysiak-Zablocki, E., Poonepalli, A., Tamblyn, L., Perez-Ordenez, B., Hande, M.P., Sanchez, O., and Hakem, R. (2007). *Cancer Research* 67, 8527-8535.
93. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A. (2007). *Nucleic Acids Res* 35: D747–D750.
94. Pawitan, Y, Bjöhle, J, Amler, L, Borg, A.-L, Egyhazi, S, Hall, P, Han, X, Holmberg, L, Huang, F, Klaar, S, Liu, E. T, Miller, L, Nordgren, H, Ploner, A, Sandelin, K, Shaw, P. M, Smeds, J, Skoog, L, Wedrén, S, & Bergh, J. (2005) *Breast Cancer Res* 7, R953–R964.
95. Payne, S.R., and Kemp, C.J. (2003). *Toxicologic Pathology* 31, 355-363.
96. Payne, S.R., Zhang, S., Tsuchiya, K., Moser, R., Gurley, K.E., Longton, G., DeBoer, J., and Kemp, C.J. (2008). *Molecular and Cellular Biology* 28, 258-268.

97. Phillips, H. S, Kharbanda, S, Chen, R, Forrest, W. F, Soriano, R. H, Wu, T. D, Misra, A, Nigro, J. M, Colman, H, Soroceanu, L, Williams, P. M, Modrusan, Z, Feuerstein, B. G, & Aldape, K. (2006). *Cancer Cell* 9, 157–173.
98. Popescu, N.C., and Zimonjic, D.B. (1997). *Cancer Genetics and Cytogenetics* 93, 10-21.
99. Pui, C.H., Chessells, J.M., Camitta, B., Baruchel, A., Biondi, A., Boyett, J.M., Carroll, A., Eden, O.B., Evans, W.E., Gadner, H., et al. (2003). *Leukemia* 17, 700-706.
100. Pui, C.H., Schrappe, M., Ribeiro, R.C., and Niemeyer, C.M. (2004). *American Society of Hematology American Society of Hematology Education Program*, 118-145.
101. Rauch, A., Ruschendorf, F., Huang, J., Trautmann, U., Becker, C., Thiel, C., Jones, K.W., Reis, A., and Nurnberg, P. (2004). *Journal of Medical Genetics* 41, 916-922.
102. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). *Nature* 444, 444-454.
103. Risinger, M.A., and Groden, J. (2004). *Cancer Cell* 6, 539-545.
104. Robak, T. (2004). *Hematology* 9, 387-400.
105. Rooney, S., Alt, F.W., Lombard, D., Whitlow, S., Eckersdorff, M., Fleming, J., Fugmann, S., Ferguson, D.O., Schatz, D.G., and Sekiguchi, J. (2003). *Journal of Experimental Medicine* 197, 553-565.
106. Rooney, S., Chaudhuri, J., and Alt, F.W. (2004a). *Immunological Reviews* 200, 115-131.
107. Rooney, S., Sekiguchi, J., Whitlow, S., Eckersdorff, M., Manis, J.P., Lee, C., Ferguson, D.O., and Alt, F.W. (2004b). *Proceedings of the National Academy of Sciences of the United States of America* 101, 2410-2415.
108. Rooney, S., Sekiguchi, J., Zhu, C., Cheng, H.-L., Manis, J., Whitlow, S., DeVido, J., Foy, D., Chaudhuri, J., Lombard, D., et al. (2002). *Molecular Cell* 10, 1379-1390.
109. Sebat J., Lakshmi B., Malhotra D., Troge J., Lese-Martin C., Walsh T., Yamrom B., Yoon S., Krasnitz A., Kendall J., et al. *Science* (2007) 316:445–449
110. Shima, N., Alcaraz, A., Liachko, I., Buske, T.R., Andrews, C.A., Munroe, R.J., Hartford, S.A., Tye, B.K., and Schimenti, J.C. (2007). *Nature Genetics* 39, 93-98.
111. Shivji, M.K.K., and Venkitaraman, A.R. (2004). *DNA Repair* 3, 835-843.
112. Snouwaert, J.N., Gowen, L.C., Latour, A.M., Mohn, A.R., Xiao, A., Dibiase, L., and Koller, B.H. (1999). *Oncogene* 18, 7900-7907.

113. Soulas-Sprauel, P., Guyader, G.L., Rivera-Munoz, P., Abramowski, V., Olivier-Martin, C., Goujet-Zalc, C., Charneau, P., and de Villartay, J.-P. (2007). *J Exp Med*, jem.20070255.
114. Storey, JD, Tibshirani, R (2003) *Proc Natl Acad Sci U S A* 100:9440–9445.
115. Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). *Science* 302, 249–255.
116. Sweet-Cordero, A et al. (2005) *Nat Genet* 37:48–55.
117. Taccioli, G.E., Amatucci, A.G., Beamish, H.J., Gell, D., Xiang, X.H., Arzayus, M.I.T., Priestley, A., Jackson, S.P., Rothstein, A.M., Jeggo, P.A., et al. (1998). *Immunity* 9, 355-366.
118. Takahashi, E., Hori, T., O'Connell, P., Leppert, M., and White, R. (1991). *Cytogenetics and Cell Genetics* 57, 109-111.
119. Taniguchi, Y., Takeda, S., Furutani-Seiki, M., Kamei, Y., Todo, T., Sasado, T., Deguchi, T., Kondoh, H., Mudde, J., Yamazoe, M., et al. (2006). *Genome Biology* 7, R116.
120. The Autism Genome Project Consortium. *Nat. Genet.* (2007) 39:319–328.
121. Thomas, PD et al. (2003) *Genome Res* 13:2129–2141.
122. Thomas, PD et al. (2006) *Nucleic Acids Res* 34:W645–W650.
123. Thornton, J.D., Silva, R.L.A., Martin, A.C., and Skapek, S.X. (2005). *Cell Cycle* 4, 1316-1319.
124. Tischfield, J.A., and Shao, C. (2003). *Nature Genetics* 33, 5-6.
125. Tsafrir, D., Bacolod, M., Selvanayagam, Z., Tsafrir, I., Shia, J., Zeng, Z., Liu, H., Krier, C., Stengel, R.F., Barany, F., et al. (2006). *Cancer Research* 66, 2129-2137.
126. Tsimberidou, A.M., and Keating, M.J. (2005). *Cancer* 103, 216-228.
127. Vidal, M., and Cagan, R.L. (2006). *Current Opinion in Genetics & Development* 16, 10-16.
128. Walker, B.A., Leone, P.E., Jenner, M.W., Li, C., Gonzalez, D., Johnson, D.C., Ross, F.M., Davies, F.E., and Morgan, G.J. (2006). *IBlood* 108, 1733-1743.
129. Wang, Y et al. (2005) *Lancet* 365:671–679.
130. Ward, I.M., Reina-San-Martin, B., Olaru, A., Minn, K., Tamada, K., Lau, J.S., Cascalho, M., Chen, L., Nussenzweig, A., Livak, F., et al. (2004). *Journal of Cell Biology* 165, 459-464.
131. Wei, K.a.C.A.B.d.W.E.a.K.M.F.e.M.D.J.e.P.T.a.K.N.K.b.R. (2003). *Genes and Development* 17, 5, 603-614.
132. Wheeler, DL et al. (2008) *Nucleic Acids Res* 36:D13–D21.

133. Whitfield, ML, George, LK, Grant, GD, Perou, CM (2006) *Nat Rev Cancer* 6:99–106.
134. Williams, C.J., Grandal, I., Vesprini, D.J., Wojtyra, U., Danska, J.S., and Guidos, C.J. (2001). *Molecular and Cellular Biology* 21, 400–413.
135. Wolf, M., Mousses, S., Hautaniemi, S., Karhu, R., Huusko, P., Allinen, M., Elkahloun, A., Monni, O., Chen, Y., Kallioniemi, A., et al. (2004). *Neoplasia* 6, 240–247.
136. Wong, J.C.Y., and Buchwald, M. (2002). 8, 139–142.
137. Woo, Y et al. (2007). *Oncogene* 26:6010–6020.
138. Xu, Y., Ashley, T., Brainerd, E.E., Bronson, R.T., Meyn, M.S., and Baltimore, D. (1996). *Genes and Development* 10, 2411–2422.
139. Yan, C.T., Kaushal, D., Murphy, M., Zhang, Y., Datta, A., Chen, C., Monroe, B., Mostoslavsky, G., Coakley, K., Gao, Y., et al. (2006). *Proceedings of the National Academy of Sciences of the United States of America* 103, 7378–7383.
140. Yang, Y. et al. *Am. J. Hum. Genet.* 80, 1037–1054 (2007).
141. Yee, K.W.L., O'Brien, S.M., and Giles, F.J. (2005) *Cancer Journal* 11, 161–174.
142. Zha, S., Alt, F.W., Cheng, H.-L., Brush, J.W., and Li, G. (2007). *Proceedings of the National Academy of Sciences* 104, 4518–4523.
143. Zender, L et al. (2006) *Cell* 125:1253–1267.
144. Zhu, C et al. (2002) *Cell* 109:811–821.

CLAIMS

1. A method for determining a prognosis for an individual having cancer, the method comprising:

(a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group I in a biological sample from the individual, wherein Group I consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); DUT (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); FOXM1 (SEQ ID NO: 57); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LSM6 (SEQ ID NO: 72); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ

ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SETD6 (SEQ ID NO: 120); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150) and ZFYVE26 (SEQ ID NO: 151); and

(b) determining a metagene score for said patient based on the expression of said at least 5 nucleic acid sequences;

(c) comparing said metagene score determined in step (b) with a set of reference metagene scores based on the expression of said at least 5 nucleic acid sequences in a population of individuals having cancer;

wherein a metagene score from said patient greater than the median of said set of reference metagene scores indicates a poor prognosis; and wherein a metagene score from said patient less than the median of said set of reference metagene scores

indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

2. A method for sub-typing or staging a cancer in a subject, the method comprising;
 - (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group I in a biological sample from the subject, wherein Group I consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); DUT (includes EG:1854) (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); FOXM1 (SEQ ID NO: 57); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LSM6 (SEQ ID NO: 72); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74);

MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SETD6 (SEQ ID NO: 120); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150) and ZFYVE26 (SEQ ID NO: 151); and

- (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level;
- (c) comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or

prognosis or outcome, wherein said comparing identifies the reference expression level in said database closest in value to the normalized expression level for said patient tumor, whereby the identification of the closest reference expression value indicates the tumor in said reference database that has the closest reference expression score is most similar to said patient tumor and therefore most similar in sub-type or stage to said patient tumor.

3. A method for determining a prognosis for an individual having cancer, the method comprising:

(a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group II in a biological sample from the individual, wherein Group II consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24) CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CEP70 (SEQ ID NO: 28) CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31) CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO:

74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MCM7 (SEQ ID NO: 79) MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151); and

(b) determining a metagene score for said patient based on the expression of said at least 5 nucleic acid sequences;

(c) comparing said metagene score determined in step (b) with a set of reference metagene scores based on the expression of said at least 5 nucleic acid sequences in a population of individuals having cancer;

wherein a metagene score from said patient greater than the median of said set of reference metagene scores indicates a poor prognosis; and wherein a metagene score

from said patient less than the median of said set of reference metagene scores indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

4. The method of claim 1 or 3, wherein the metagene score is designated on a scale of -1 to +1, with the median metagene score set at zero.
5. A method for sub-typing or staging a cancer in a subject, the method comprising:
 - (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group II in a biological sample from the subject, wherein Group II consists of gene sequences: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24) CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CEP70 (SEQ ID NO: 28) CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31) CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); ID3 (SEQ ID NO: 66); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO:

74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MCM7 (SEQ ID NO: 79) MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150); ZFYVE26 (SEQ ID NO: 151); and

(b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level;

(c) comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis or outcome, wherein said comparing identifies the reference expression level in said database closest in value to the normalized expression level for said

patient tumor, whereby the identification of the closest reference expression value indicates the tumor in said reference database that has the closest reference expression score is most similar to said patient tumor and therefore most similar in sub-type or stage to said patient tumor.

6. A method for determining a prognosis for an individual having cancer, the method comprising:

- (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group V in a biological sample from the individual, wherein Group V consists of gene sequences: RECQL4 (SEQ ID NO: 108) and at least one gene selected from the following: BRIP1 (SEQ ID NO: 10); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); MCM7 (SEQ ID NO: 79); MRE11A (SEQ ID NO: 81); NAP1L4 (SEQ ID NO: 83); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135); or UNG (SEQ ID NO: 143); and
- (b) determining a metagene score for said patient based on the expression of said at least 5 nucleic acid sequences;
- (c) comparing said metagene score determined in step (b) with a set of reference metagene scores based on the expression of said at least 5 nucleic acid sequences in a population of individuals having cancer;

wherein a metagene score from said patient greater than the median of said set of reference metagene scores indicates a poor prognosis; and wherein a metagene score from said patient less than the median of said set of reference metagene scores indicates a good prognosis; and wherein the degree to which the score is greater than the median reference metagene score correlates with the relative aggressiveness of the tumor.

7. A method for sub-typing or staging a cancer in a subject, the method comprising;

- (a) measuring the level of expression of at least 5 nucleic acid sequences encoding proteins selected from Group V in a biological sample from the subject, wherein Group V consists of gene sequences: RECQL4 (SEQ ID NO: 108) and at least one gene selected from the following: BRIP1 (SEQ ID NO: 10); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); MCM7

(SEQ ID NO:79); MRE11A (SEQ ID NO: 81); NAP1L4 (SEQ ID NO: 83); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135); or UNG (SEQ ID NO: 143); and

- (b) normalizing the level of expression of all nucleic acid sequences measured in (a) using a reference expression level;
- (c) comparing the normalized gene expression level for each nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known prognosis or outcome, wherein said comparing identifies the reference expression level in said database closest in value to the normalized expression level for said patient tumor, whereby the identification of the closest reference expression value indicates the tumor in said reference database that has the closest reference expression score is most similar to said patient tumor and therefore most similar in sub-type or stage to said patient tumor.

8. A method for determining a prognosis for an individual having cancer, the method comprising:

- (a) measuring the level of expression of a nucleic acid sequence encoding a protein selected from FOXM1 (SEQ ID NO: 57) or RECQL4 (SEQ ID NO: 108) in a biological sample from the subject; and
- (b) normalizing the level of expression of the nucleic acid sequence measured in (a) using a reference expression level;
- (c) comparing the normalized gene expression level for the nucleic acid sequence measured to a plurality of reference expression levels in a database determined for a plurality of reference tumors of known diagnosis and/or prognosis or outcome, wherein said comparing identifies the reference expression level in said database closest in value to the normalized expression level for said patient tumor, whereby the identification of the closest reference expression value indicates the tumor in said reference database that has the closest reference expression score is most similar to said patient tumor and therefore most similar in sub-type or stage to said patient tumor.

9. The method of any of claims 1 to 8, wherein the level of expression is the level of gene transcript expression or the level of protein expression.

10. The method of any of claims 1 to 7, wherein at least 10 said nucleic acid sequences are measured.

11. The method of any of claims 1, 2, 3 or 5, wherein said at least 5 nucleic acid sequences encoding the proteins are selected from the group of positively correlated genes of Group X, wherein Group X consists of: ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); ATAD5 (SEQ ID NO: 7); BRIP1 (SEQ ID NO: 10); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); CCNA2 (SEQ ID NO: 19); CDC25A (SEQ ID NO: 26); CEP70 (SEQ ID NO: 28); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLUAP1 (SEQ ID NO: 34); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); FAM128B (SEQ ID NO: 53); FBXO21 (SEQ ID NO: 55); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HDDC2 (SEQ ID NO: 64); ID3 (SEQ ID NO: 66); LRDD (SEQ ID NO: 71); MAZ (SEQ ID NO: 77); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NEK2 (SEQ ID NO: 86); NSMCE4A (SEQ ID NO: 89); ORC5L (SEQ ID NO: 91); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); PPP1R8 (SEQ ID NO: 97); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2 (SEQ ID NO: 113); SMARCD2 (SEQ ID NO: 123); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TOPBP1 (SEQ ID NO: 135); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); WARS2 (SEQ ID NO: 146); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149) and WEE1 (SEQ ID NO: 150).

12. The method of any of claims 1, 2, 3 or 5, wherein said at least 5 nucleic acid sequences encoding the proteins are selected from the group of negatively correlated

genes of Group XI, wherein Group XI consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BSDC1 (SEQ ID NO: 11); C5ORF15 (SEQ ID NO: 18); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CLEC10A (SEQ ID NO: 33); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DSE (SEQ ID NO: 44); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EVI2A (SEQ ID NO: 52); FAM129A (SEQ ID NO: 54); FLII (SEQ ID NO: 56); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); NCF2 (SEQ ID NO: 85); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); OPTN (SEQ ID NO: 90); PBXIP1 (SEQ ID NO: 92); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); RAB31 (SEQ ID NO: 103); RASGRP3 (SEQ ID NO: 107); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TRPV2 (SEQ ID NO: 140); UGCG (SEQ ID NO: 142); USP32 (SEQ ID NO: 145); WBP2 (SEQ ID NO: 147); ZFYVE26 (SEQ ID NO: 151).

13. The method of any of claims 1 to 8, wherein the cancer is selected from the group consisting of: a leukemia, a lymphoma, a neoplasm, a breast cancer, a colon cancer, a glioma, a lung cancer, a hyperproliferative disorder, a carcinoma, a sarcoma, a metastatic cancer, a lymphatic metastases, a blood cell malignancy, a multiple myeloma, an acute leukemia, a chronic leukemia, a lymphoma, a head and neck cancer, a mouth cancer, a larynx cancer, a thyroid cancer, a respiratory tract cancer, a lung cancer, a small cell carcinoma, a non-small cell cancer, a ductal carcinoma, a gastrointestinal cancer, an esophageal cancer, a stomach cancer, a colorectal cancer, a pancreatic cancer, a liver cancer, a genitourinary cancer, an urologic cancer, a bladder cancer, a prostate cancer, a ovarian carcinoma, an uterine cancer, an endometrial cancer, a kidney cancer, a renal cell

carcinoma, a brain cancer, a neuroblastoma, an astrocytic brain tumors, a metastatic tumor cell invasion in the central nervous system, a bone cancer, osteomas, a skin cancer, a malignant melanoma, a squamous cell carcinoma, a basal cell carcinoma, a hemangiopericytoma or a Kaposi's sarcoma or combinations thereof.

14. The method of any of claims 1 to 8, wherein the reference expression level for each nucleic acid measured is the level of the nucleic acid from a biological sample obtained from the subject at a first time point.

15. The method of any of claims 1 to 8, wherein the biological sample is from the subject at a second time point, and wherein the reference expression level is the level of the nucleic acid from a biological sample obtained from the subject at a first time point.

16. The method of any of claims 1 to 8, wherein the expression level of a biological sample from the subject is compared to the reference expression level from a plurality of biological samples obtained from other subjects.

17. The method of claim 1 to 8, wherein said method indicates that said cancer is an aggressive cancer, and wherein said individual is directed to be treated with an appropriate anti-cancer therapy for an aggressive cancer.

18. The method of claim 1 to 8, wherein said method indicates that said cancer is a non-aggressive cancer, and wherein said individual is directed to be treated with an appropriate anti-cancer therapy for a non-aggressive cancer.

19. A method to identify a prognostic disease biomarker, the method comprising;
- (a) identifying a copy-number variation in a genome of a subject, and measuring the copy number of the genes affected by the copy number variation in a biological sample from a subject with a gene copy-number variation;
 - (b) comparing the copy number of the genes measured in (a) with a reference copy number for each gene measured, and identifying genes where the copy number in the subject with the copy-number variation is different as compared with the reference sample;

- (c) measuring gene expression levels in a biological sample from the subject with a gene copy number variation;
- (d) performing a copy number-gene expression correlation analysis for each gene identified in (b); and
- (e) performing a copy number-gene expression correlation analysis for each gene identified in (c);

wherein genes with an expression level dependent on copy number variation are identified as a prognostic disease biomarker in a subject

20. A method to identify a prognostic disease biomarker, the method comprising;
- (a) identifying a chromosomal abnormality in the genome of a subject, and measuring the copy number of the genes affected by the chromosomal abnormality in a biological sample obtained from the subject with a chromosomal abnormality;
 - (b) comparing the copy number of the genes measured in (a) with a reference copy number for each gene measured, and identifying genes where the copy number in the subject with the chromosomal abnormality is different as compared to the reference sample;
 - (c) measuring the gene expression levels in a biological sample from the subject;
 - (d) performing a copy number-gene expression correlation analysis for each gene identified in (c); and

wherein a gene with whose expression level is changed proportionally to the change in copy number identifies the gene as a prognostic disease biomarker.

21. The method of claims 19 or 20, wherein the copy number-gene expression correlation analysis is determined using the algorithm:

$$T_j = \beta_0 + \epsilon_j \quad (1)$$

$$T_j = \beta_0 + \beta_1 D_j + \epsilon_j \quad (2)$$

$$T_j = \beta_0 + \beta_2 C_j + \epsilon_j \quad (3)$$

$$T_j = \beta_0 + \beta_1 D_j + \beta_2 C_j + \epsilon_j \quad (4)$$

wherein T_j is expression, D_j is the copy-number, ϵ_j is the random error of the gene in sample j ; C_j is the copy-number of the CNA in sample j ; β_1 is the regression

parameters measuring local association; β_2 is the regression parameters measuring global association; β_0 is an intercept regression parameter.

22. The method of claim 19 or 20, wherein the reference copy number is the copy number from a subject without the chromosomal abnormality.
23. The method of claim 19 or 20, wherein the difference in copy number is an increase in copy number.
24. The method of claim 19 or 20, wherein the difference in copy number is a decrease in copy number.
25. The method of claim 19 or 20, further comprising performing a cross-species gene ontology comparison of the genes identified as a prognostic disease biomarker, wherein if at least a 1:1 ratio of the prognostic disease biomarker to human or subject to be treated (non-human animal) ortholog is present, then the prognostic disease biomarker is identified as a core prognostic disease biomarker.
26. The method of claim 19 or 20, wherein the copy number-gene expression correlation analysis is determined by step 2 of the Genome Association Analysis of Transcription and Comparative Oncogenomics method.
27. The method of claim 19, wherein the gene copy number variation is a chromosomal abnormality.
28. The method of claim 19 or 20, wherein the gene copy number variation is present in a disease.
29. The method of any of claims 19, 20 or 28, wherein the gene copy number variation is present in cancer.

30. The method of any of claims 19, 20 or 28, wherein the gene copy number variation is present in a nervous system disorder, metabolic disorder or a cardiovascular disorder.
31. The method of claim 19 or 20, wherein the disease is cancer.
32. The method of claim 19 or 20, wherein the biological sample is cancer tissue.
33. The method of claim 32, wherein the cancer tissue is a biopsy cancer tissue.
34. The method of claim 19 or 20, wherein the subject is a human.
35. The method of claim 19 or 20, further comprising performing a gene ontology analysis of the genes identified as a prognostic disease biomarker, and using such to identify novel disease biomarkers.
36. The method of any of claims 19 or 20, wherein said method identifies therapeutic targets for disease.
37. The method of any of claims 19 or 20, wherein said method identifies a therapeutic pathway or strategy.
38. A kit comprising primers which specifically hybridize to the nucleic acid sequences of at least 5 genes selected from the group of genes encoding proteins of Group I, wherein Group I consists of gene sequences; ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7

(SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7 (SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); DUT (includes EG:1854) (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); FOXM1 (SEQ ID NO: 57); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LSM6 (SEQ ID NO: 72); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBP1 (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SETD6 (SEQ ID NO: 120); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682)

(SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAIIP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO: 141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150) and ZFYVE26 (SEQ ID NO: 151).

39. The kit of claim 37, further comprising primers which specifically hybridize to the nucleic acid sequences of at least one normalizing gene selected from the group consisting of: GAPDH (glyceraldehyde 3-phosphate dehydrogenase), SDHA (succinate dehydrogenase), HPRT (hypoxanthine phosphoribosyl transferase), HBS1L (HBS1-like protein), AHSP (alpha hemoglobin stabilizing protein), B2M (beta-2-microglobulin), beta-actin, beta-tubulin, large ribosomal protein, 18S ribosomal RNA, transferrin receptor, TATA-box-binding protein, beta-glucuronidase, cyclophilin A, porphobilinogen deaminase, ATP synthase; CALCR (calcitonin receptor), CRHR2 (Corticotropin-releasing hormone receptor 2), CYLC1 (Cylicin I), EDA (ectodysplasin A), GNRHR (GnRH receptor), HNF4A (hepatocyte nuclear factor 4, alpha), PAX8 (paired box 8), Mbd2 (methyl-CpG binding domain protein 2), ADAM22 (A disintegrin and metalloproteinase domain 22), Nf2 (neurofibromin 2), NPY2R (neuropeptide Y receptor Y2), TRIM15 (tripartite motif-containing 15), TRIM10 (tripartite motif-containing 10), RHAG (CD241 antigen), PDE10A (phosphodiesterase 10A) or ZBTB7A (Factor binding IST protein 1).

40. The kit of claim 38, which comprises primers which specifically hybridize to nucleic acid sequences of at least five normalizing genes selected from said group of normalizing genes.

41. A kit comprising primers which specifically hybridize to the nucleic acid sequences of at least 5 genes selected from the group of genes encoding proteins of Group V, wherein Group V consists of gene sequences; RECQL4 (SEQ ID NO: 108) and at least one gene selected from the following: BRIP1 (SEQ ID NO: 10); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); MCM7 (SEQ ID NO: 79); MRE11A (SEQ ID NO: 81); NAP1L4 (SEQ ID NO: 83); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95);

RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135); or UNG (SEQ ID NO: 143).

42. The kit of any of claims 38 to 41, wherein the kit comprises primers which specifically hybridize to the nucleic acid sequences of at least 10 genes.

43. The kit of any of the claims 38 to 41, wherein the kit comprises at least 6 but no more than 100 additional sets of primers.

44. The kit of any of the claims 38 to 41, wherein the kit further comprises at least 10 additional sets of primers.

45. The kit of any of the claims 38 to 41, wherein the kit comprises at least one primer which specifically hybridizes to the nucleic acid encoding the protein RECQL4 (SEQ ID NO: 108).

46. The kit of any of the claims 38 to 41, wherein the kit comprises at least one primer which specifically hybridizes to the nucleic acid encoding the protein FOXM1 (SEQ ID NO: 57).

47. A kit comprising protein-binding probes which specifically bind to the protein of at least 5 proteins selected from Group I, wherein Group I consists of: ABCA3 (SEQ ID NO: 1); ACADVL (SEQ ID NO: 2); ACP6 (SEQ ID NO: 3); ALG6 (SEQ ID NO: 4); APH1B (SEQ ID NO: 5); ARHGEF3 (SEQ ID NO: 6); ATAD5 (SEQ ID NO: 7); ATF6 (SEQ ID NO: 8); ATP6V0A1 (SEQ ID NO: 9); BRIP1 (SEQ ID NO: 10); BSDC1 (SEQ ID NO: 11); C12ORF29 (SEQ ID NO: 12); C14ORF104 (SEQ ID NO: 13); C14ORF124 (SEQ ID NO: 14); C16ORF80 (SEQ ID NO: 15); C18ORF24 (SEQ ID NO: 16); C21ORF45 (SEQ ID NO: 17); C5ORF15 (SEQ ID NO: 18); CCNA2 (SEQ ID NO: 19); CCNB2 (SEQ ID NO: 20); CCPG1 (SEQ ID NO: 21); CCR1 (SEQ ID NO: 22); CD47 (SEQ ID NO: 23); CD80 (SEQ ID NO: 24); CD83 (SEQ ID NO: 25); CDC25A (SEQ ID NO: 26); CDT1 (SEQ ID NO: 27); CEP70 (SEQ ID NO: 28); CHAF1B (SEQ ID NO: 29); CHCHD7 (SEQ ID NO: 30); CHEK1 (SEQ ID NO: 31); CISD1 (SEQ ID NO: 32); CLEC10A (SEQ ID NO: 33); CLUAP1 (SEQ ID NO: 34); CSF1R (SEQ ID NO: 35); CST3 (SEQ ID NO: 36); CTSO (SEQ ID NO: 37); CX3CR1 (SEQ ID NO: 38); CXCR7

(SEQ ID NO: 39); DAB2 (SEQ ID NO: 40); DDX39 (SEQ ID NO: 41); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); DSE (SEQ ID NO: 44); DUT (includes EG:1854) (SEQ ID NO: 45); E2F1 (SEQ ID NO: 46); E2F8 (SEQ ID NO: 47); EBI2 (SEQ ID NO: 48); EHD4 (SEQ ID NO: 49); EIF5A2 (SEQ ID NO: 50); ETAA1 (SEQ ID NO: 51); EVI2A (SEQ ID NO: 52); FAM128B (SEQ ID NO: 53); FAM129A (SEQ ID NO: 54); FBXO21 (SEQ ID NO: 55); FLII (SEQ ID NO: 56); FOXM1 (SEQ ID NO: 57); GABARAP (SEQ ID NO: 58); GALNT11 (SEQ ID NO: 59); GPSM2 (SEQ ID NO: 60); HADH (SEQ ID NO: 61); HBP1 (SEQ ID NO: 62); HCK (SEQ ID NO: 63); HDDC2 (SEQ ID NO: 64); HRAS (SEQ ID NO: 65); ID3 (SEQ ID NO: 66); IDH3A (SEQ ID NO: 67); IL2RA (SEQ ID NO: 68); ITGB2 (SEQ ID NO: 69); LAMB3 (SEQ ID NO: 70); LRDD (SEQ ID NO: 71); LSM6 (SEQ ID NO: 72); LY6E (SEQ ID NO: 73); MAF (SEQ ID NO: 74); MAN2A1 (SEQ ID NO: 75); MAP3K8 (SEQ ID NO: 76); MAZ (SEQ ID NO: 77); MBD3 (includes EG:53615) (SEQ ID NO: 78); MCM7 (SEQ ID NO: 79); MPHOSPH9 (SEQ ID NO: 80); MRE11A (SEQ ID NO: 81); MUM1 (SEQ ID NO: 82); NAP1L4 (SEQ ID NO: 83); NCAPH (SEQ ID NO: 84); NCF2 (SEQ ID NO: 85); NEK2 (SEQ ID NO: 86); NEU1 (SEQ ID NO: 87); NRBPI (SEQ ID NO: 88); NSMCE4A (SEQ ID NO: 89); OPTN (SEQ ID NO: 90); ORC5L (SEQ ID NO: 91); PBXIP1 (SEQ ID NO: 92); PFKM (SEQ ID NO: 93); PKMYT1 (SEQ ID NO: 94); POLD1 (SEQ ID NO: 95); POLE (SEQ ID NO: 96); PPP1R8 (SEQ ID NO: 97); PRDM1 (SEQ ID NO: 98); PRR13 (SEQ ID NO: 99); PSCDBP (SEQ ID NO: 100); QDPR (SEQ ID NO: 101); R3HCC1 (SEQ ID NO: 102); RAB31 (SEQ ID NO: 103); RACGAP1 (SEQ ID NO: 104); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RASGRP3 (SEQ ID NO: 107); RECQL4 (SEQ ID NO: 108); RFC2 (SEQ ID NO: 109); RNF41 (SEQ ID NO: 110); RRM1 (SEQ ID NO: 111); SAE1 (SEQ ID NO: 112); SAE2, also called UBA2 (SEQ ID NO: 113); SAMHD1 (SEQ ID NO: 114); SCAMP2 (SEQ ID NO: 115); SCARB2 (SEQ ID NO: 116); SCPEP1 (SEQ ID NO: 117); SEMA4A (SEQ ID NO: 118); SERINC1 (SEQ ID NO: 119); SETD6 (SEQ ID NO: 120); SFXN3 (SEQ ID NO: 121); SLAMF1 (SEQ ID NO: 122); SMARCD2 (SEQ ID NO: 123); SQSTM1 (SEQ ID NO: 124); STAB1 (SEQ ID NO: 125); STAT4 (SEQ ID NO: 126); STX11 (SEQ ID NO: 127); TAF4 (SEQ ID NO: 128); TDRD7 (SEQ ID NO: 129); TENC1 (SEQ ID NO: 130); TIPIN (SEQ ID NO: 131); TMEM186 (SEQ ID NO: 132); TMEM214 (also called FLJ20254; FLJ39682) (SEQ ID NO: 133); TNIP1 (SEQ ID NO: 134); TOPBP1 (SEQ ID NO: 135); TRAP (SEQ ID NO: 136); TRIB2 (SEQ ID NO: 137); TRIM28 (SEQ ID NO: 138); TROAP (SEQ ID NO: 139); TRPV2 (SEQ ID NO: 140); TUBB6 (SEQ ID NO:

141); UGCG (SEQ ID NO: 142); UNG (SEQ ID NO: 143); USP1 (SEQ ID NO: 144); USP32 (SEQ ID NO: 145); WARS2 (SEQ ID NO: 146); WBP2 (SEQ ID NO: 147); WDHD1 (SEQ ID NO: 148); WDR76 (SEQ ID NO: 149); WEE1 (SEQ ID NO: 150) and ZFYVE26 (SEQ ID NO: 151).

48. A kit comprising protein-binding probes which specifically bind to the protein of at least 5 proteins selected from Group V, wherein Group V consists of: RECQL4 (SEQ ID NO: 108) and at least one gene selected from the following: BRIP1 (SEQ ID NO: 10); DNA2 (SEQ ID NO: 42); DNMT1 (SEQ ID NO: 43); MCM7 (SEQ ID NO: 79); MRE11A (SEQ ID NO: 81); NAP1L4 (SEQ ID NO: 83); ORC5L (SEQ ID NO: 91); POLD1 (SEQ ID NO: 95); RAD51 (SEQ ID NO: 105); RAD54B (SEQ ID NO: 106); RFC2 (SEQ ID NO: 109); TOPBP1 (SEQ ID NO: 135); or UNG (SEQ ID NO: 143).

49. The kit of claim 47 or 48, further comprising a protein-binding probe that specifically binds to a normalizing protein of at least one normalizing gene selected from the group consisting of: GAPDH (glyceraldehyde 3-phosphate dehydrogenase), SDHA (succinate dehydrogenase), HPRT (hypoxanthine phosphoribosyl transferase), HBS1L (HBS1-like protein), AHSP (alpha hemoglobin stabilizing protein), B2M (beta-2-microglobulin), beta-actin, beta-tubulin, large ribosomal protein, 18S ribosomal RNA, transferrin receptor, TATA-box-binding protein, beta-glucuronidase, cyclophilin A, porphobilinogen deaminase, ATP synthase; CALCR (calcitonin receptor), CRHR2 (Corticotropin-releasing hormone receptor 2), CYLC1 (Cyclicin I), EDA (ectodysplasin A), GNRHR (GnRH receptor), HNF4A (hepatocyte nuclear factor 4, alpha), PAX8 (paired box 8), Mbd2 (methyl-CpG binding domain protein 2), ADAM22 (A disintegrin and metalloproteinase domain 22), Nf2 (neurofibromin 2), NPY2R (neuropeptide Y receptor Y2), TRIM15 (tripartite motif-containing 15), TRIM10 (tripartite motif-containing 10), RHAG (CD241 antigen), PDE10A (phosphodiesterase 10A) or ZBTB7A (Factor binding IST protein 1).

50. The kit of claims 47 or 48, wherein the kit comprises protein-binding probes which specifically bind to at least 10 proteins.

51. The kit of the claims 47 or 48, wherein the kit comprises at least 6 but no more than 100 additional sets of protein-binding probes.

52. The kit of claims 47 or 48, wherein the kit further comprises at least 10 additional sets of protein-binding probes.
53. The kit of claims 47 or 48, wherein the kit comprises at least one protein binding probe which specifically binds to the protein RECQL4 (SEQ ID NO: 108).
54. The kit of claims 47 or 48, wherein the kit comprises at least one protein binding probe which specifically binds to the protein FOXM1 (SEQ ID NO: 57).
55. The kit of claims 47 or 48, wherein the kit is an ELISA kit.
56. The kit of claims 38 to 41, wherein the kit is a quantitative PCR kit.
57. A computer system for predicting the prognosis of a cancer patient, the system comprising:
- (a) a computer processor,
 - (b) a computer readable physical storage medium having instructions recorded thereon sufficient to implement a process, employing the computer processor, for predicting the prognosis of a cancer patient, the instructions comprising:
 - (i) instructions for accepting a plurality of input values characterizing gene expression in a tumor from said patient, wherein said values include values characterizing a normalizing gene and a prognostic gene set;
 - (ii) instructions for normalizing values characterizing gene expression of genes in said prognostic gene set relative to said normalizing gene,
 - (iii) instructions for determining, using singular value decomposition, a metagene score for the patient;
 - (iv) instructions for comparing said metagene score of said patient with a reference metagene score normalized in the same manner as step (ii) and (iii), and
 - (v) instructions for transmitting to a user interface the results of said comparison,
- wherein a patient metagene score larger than said reference metagene score indicates a poor prognosis; and wherein a patient metagene score less than said reference metagene score indicates a good prognosis; and wherein the degree to which the score is

greater than or less than said reference metagene score correlates with the relative aggressiveness of the tumor.

58. A computer readable physical storage medium having instructions recorded thereon sufficient to implement a process, employing a computer processor, for predicting the prognosis of a cancer patient, the instructions comprising:

- (i) instructions for accepting a plurality of input values characterizing gene expression in a tumor from said patient, wherein said values include values characterizing a normalizing gene and a prognostic gene set;
- (ii) instructions for normalizing values characterizing gene expression of genes in said prognostic gene set relative to said normalizing gene,
- (iii) instructions for determining, using singular value decomposition, a metagene score for the patient;
- (iv) instructions for comparing said metagene score of said patient with a reference metagene score normalized in the same manner as in step (ii) and (iii), and
- (v) instructions for transmitting to a user interface the results of said comparison,

wherein a patient metagene score larger than said reference metagene score indicates a poor prognosis; and wherein a patient metagene score less than said reference metagene score indicates a good prognosis; and wherein the degree to which the score is greater than or less than said reference metagene score correlates with the relative aggressiveness of the tumor.

59. A computer system for predicting the prognosis of a cancer patient, the system comprising:

- (a) a computer processor,
- (b) a computer readable physical storage medium having instructions recorded thereon sufficient to implement a process, employing the computer, for predicting the prognosis of a cancer patient, the instructions comprising:
 - (i) instructions for accepting a plurality of input values characterizing gene expression in a tumor from said patient, wherein said values include values characterizing a normalizing gene and a prognostic gene set;
 - (ii) instructions for normalizing values characterizing gene expression of genes in said prognostic gene set relative to said normalizing gene,

(iii) instructions for determining, using singular value decomposition, a metagene score for the patient tumor;

(iv) instructions for comparing said metagene score of said patient with a plurality of reference metagene scores in a database, determined for a plurality of reference tumors of known prognosis or outcome, wherein said comparing identifies the reference metagene score in said database closest in value to the metagene score for said patient tumor, whereby the identification of the closest reference metagene score indicates the tumor in said reference database that has the closest reference metagene score is most similar to said patient tumor and therefore most similar in diagnosis and/or prognosis to said patient tumor; and

(v) instructions for transmitting to a user interface the results of said comparison of step (iv), whereby diagnosis and/or prognosis for said patient tumor is predicted.

60. A computer readable physical storage medium having instructions recorded thereon sufficient to implement a process, employing a computer processor, for predicting the prognosis of a cancer patient, the instructions comprising:

(i) instructions for accepting a plurality of input values characterizing gene expression in a tumor from said patient, wherein said values include values characterizing a normalizing gene and a prognostic gene set;

(ii) instructions for normalizing values characterizing gene expression of genes in said prognostic gene set relative to said normalizing gene,

(iii) instructions for determining, using singular value decomposition, a metagene score for the patient tumor;

(iv) instructions for comparing said metagene score of said patient with a plurality of reference metagene scores in a database, determined for a plurality of reference tumors of known prognosis or outcome, wherein said comparing identifies the reference metagene score in said database closest in value to the metagene score for said patient tumor, whereby the identification of the closest reference metagene score indicates the tumor in said reference database that has the closest reference metagene score is most similar to said patient tumor and therefore most similar in diagnosis and/or prognosis to said patient tumor; and

(v) instructions for transmitting to a user interface the results of said comparison of step (iv), whereby diagnosis and/or prognosis for said patient tumor is predicted.

1/22

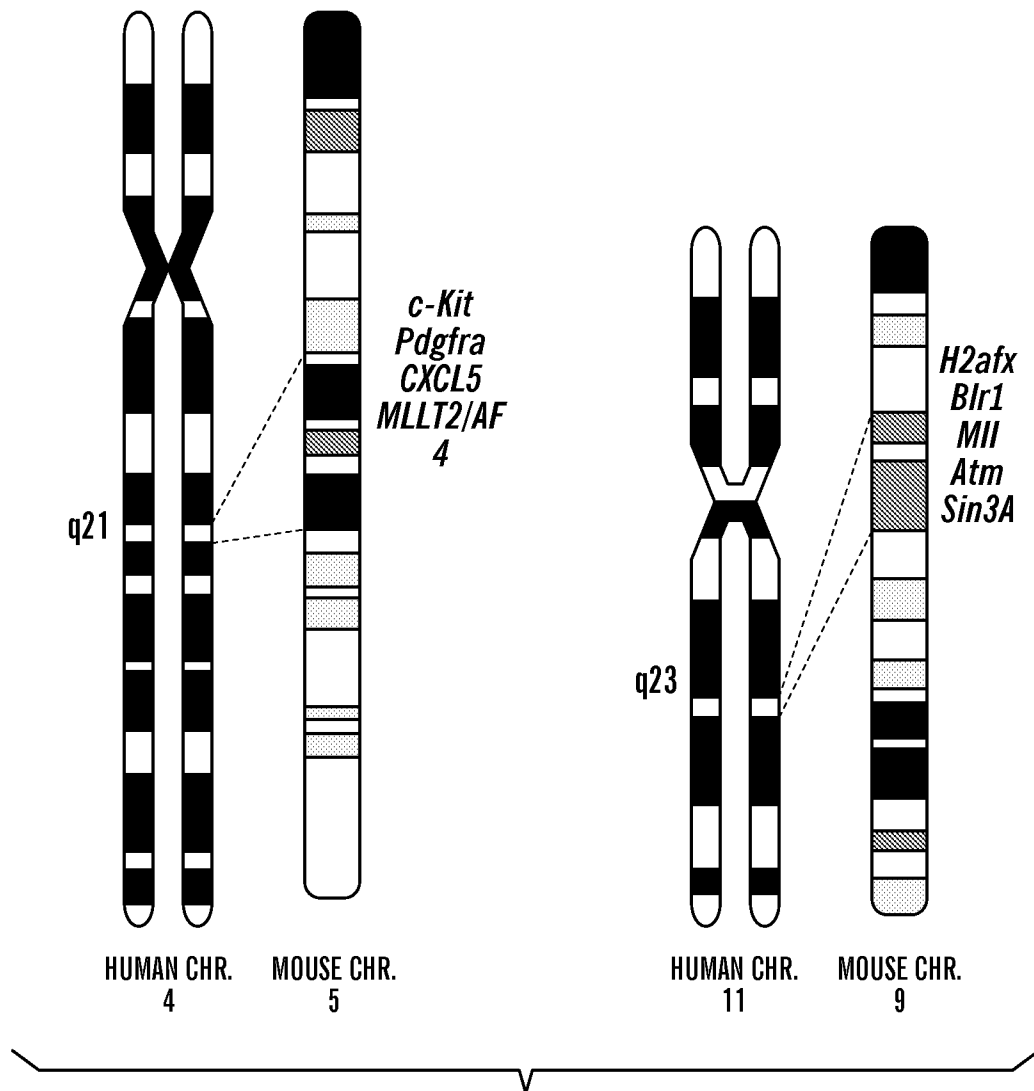
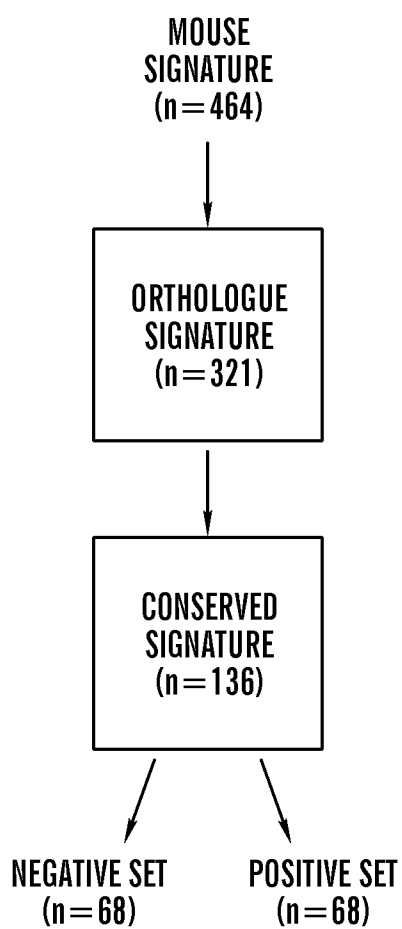
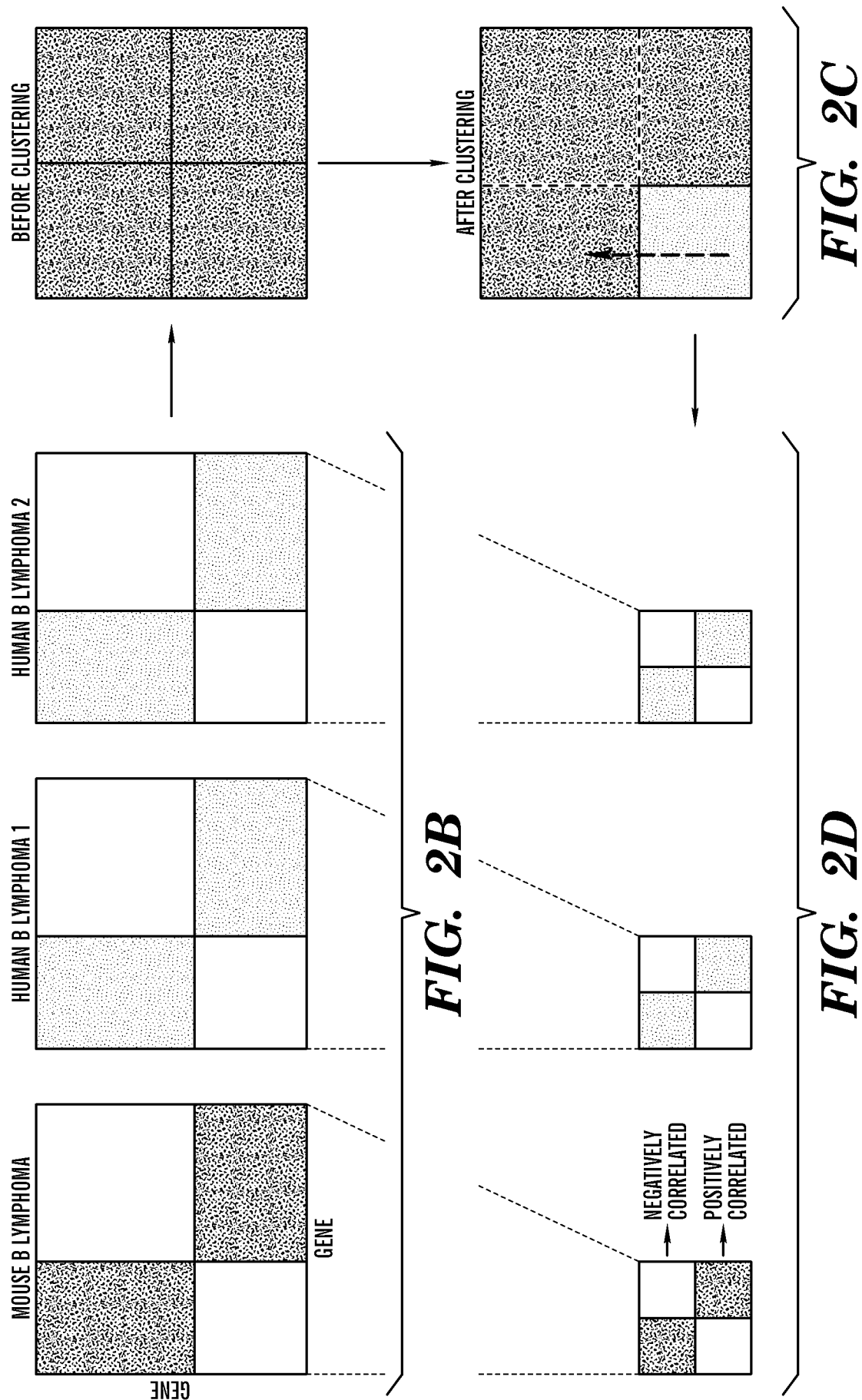


FIG. 1

2/22

***FIG. 2A***



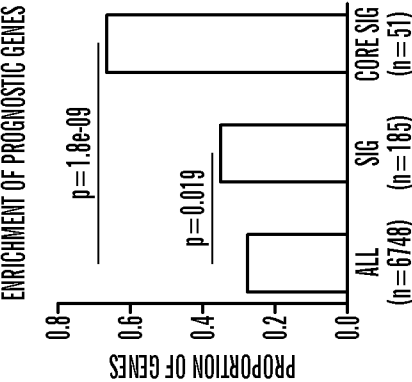


FIG. 3A

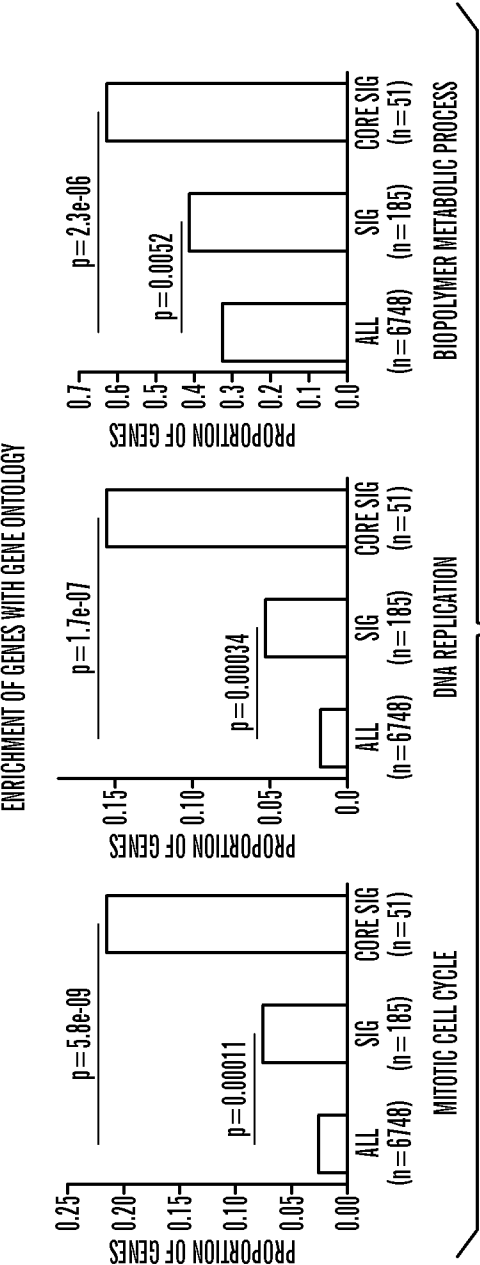


FIG. 3B

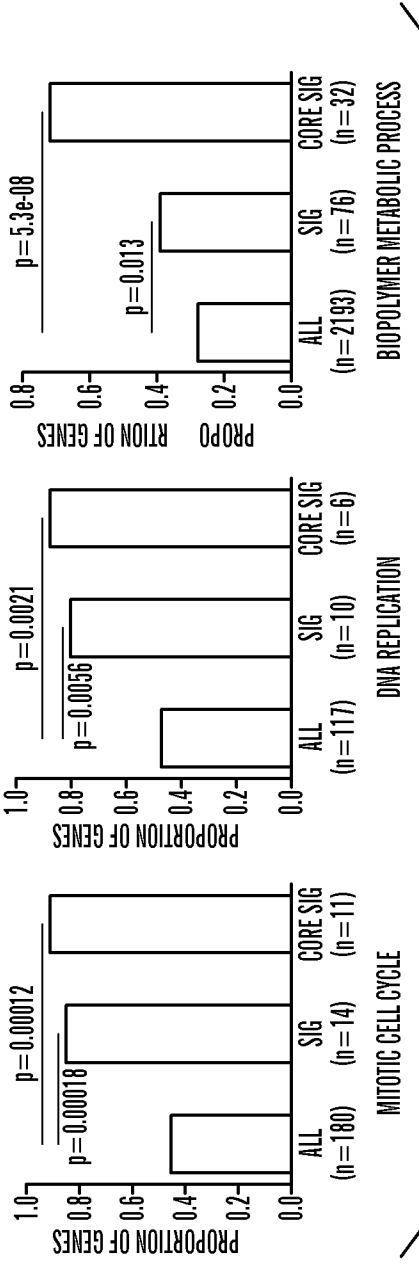


FIG. 3C

5/22

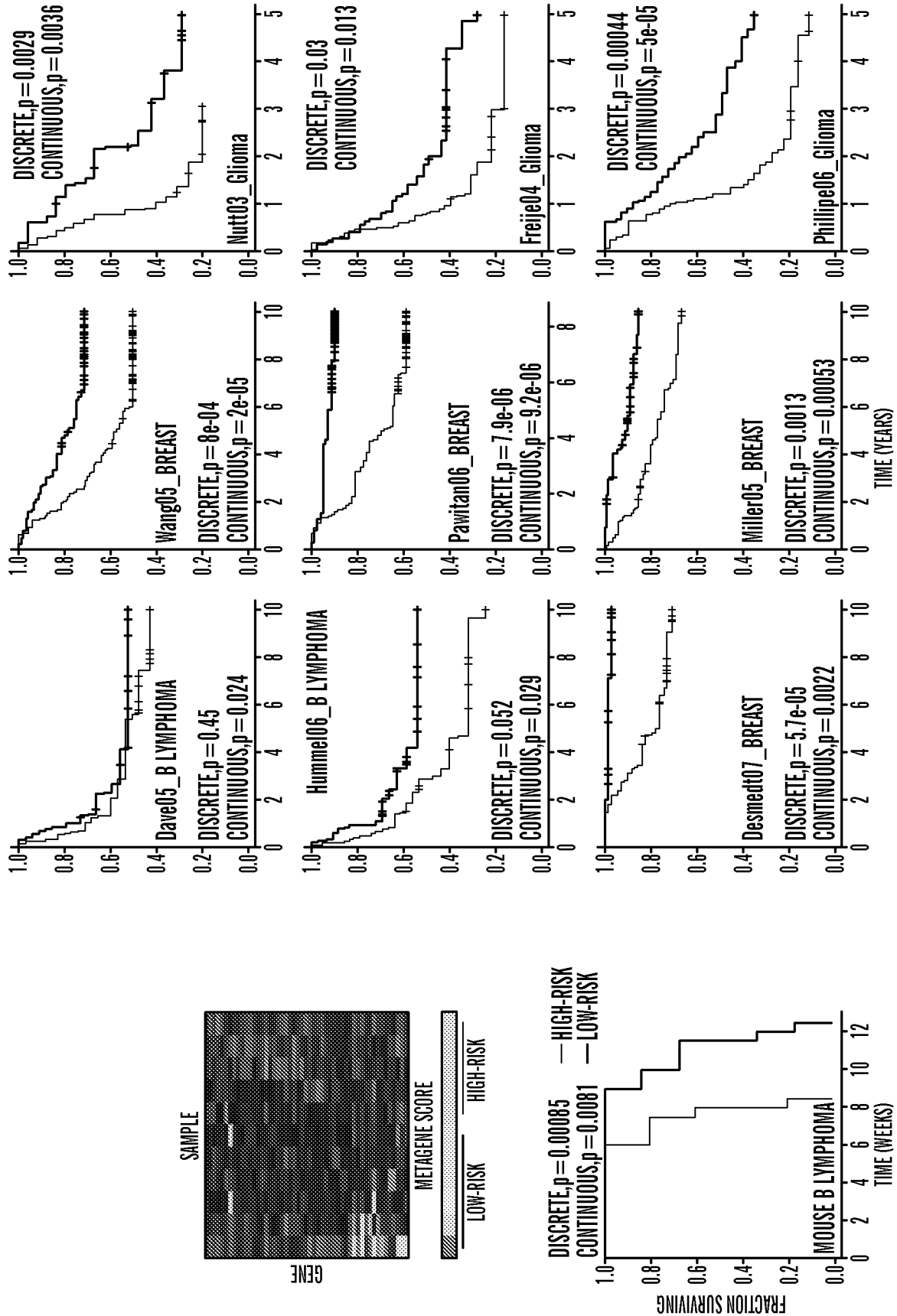


FIG. 4

6/22

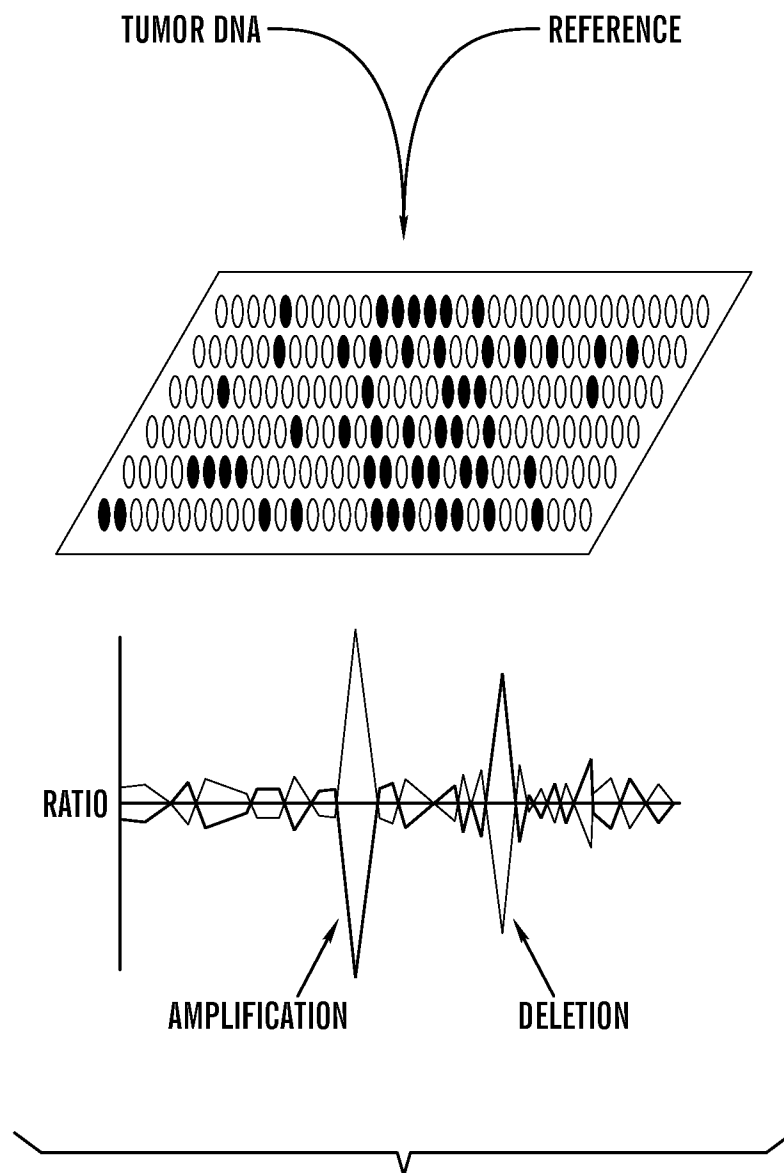
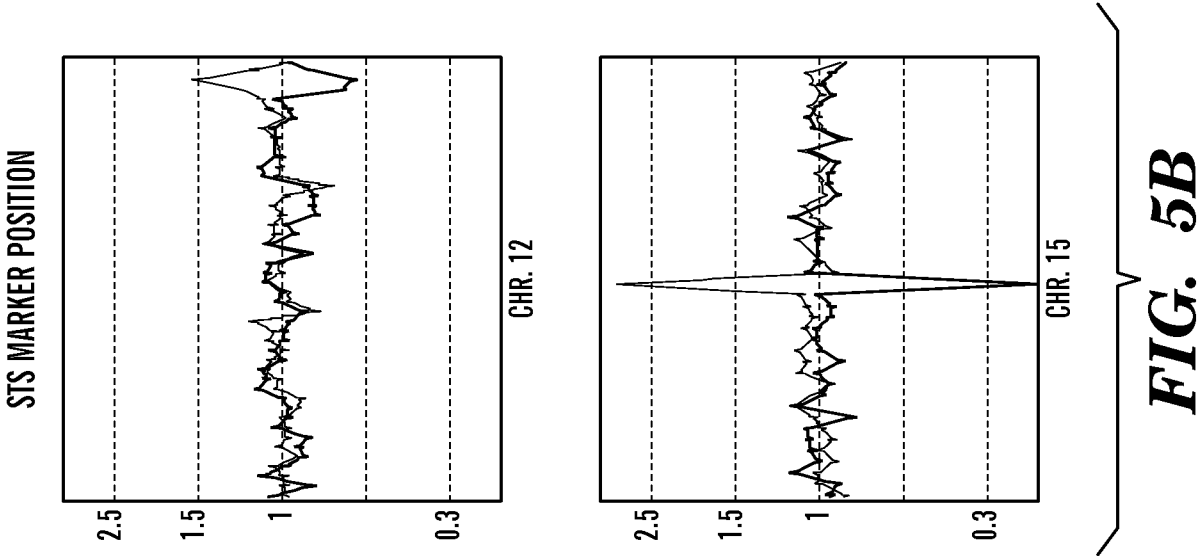
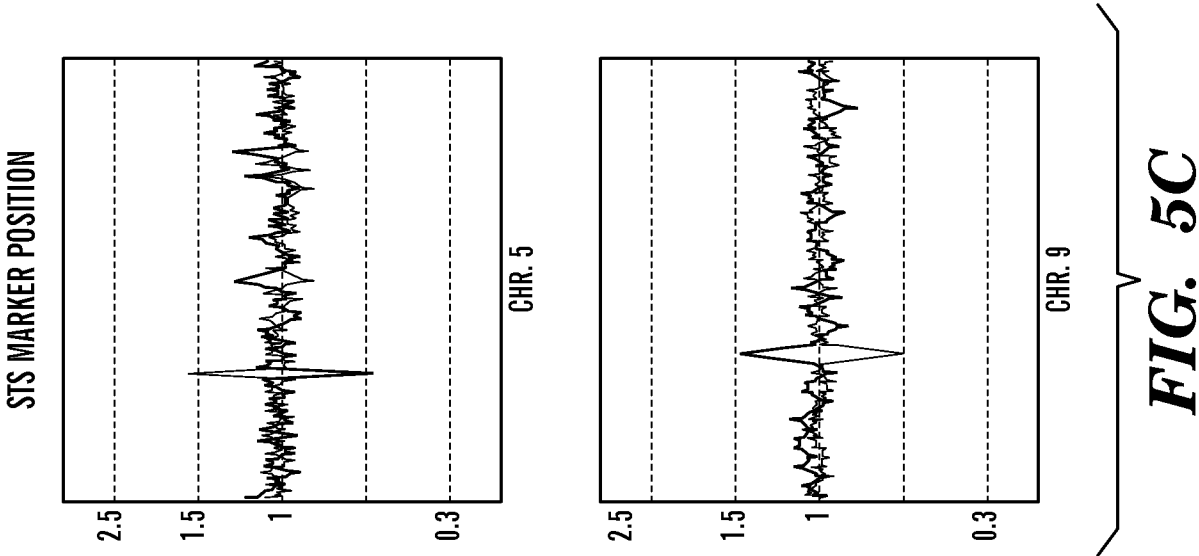
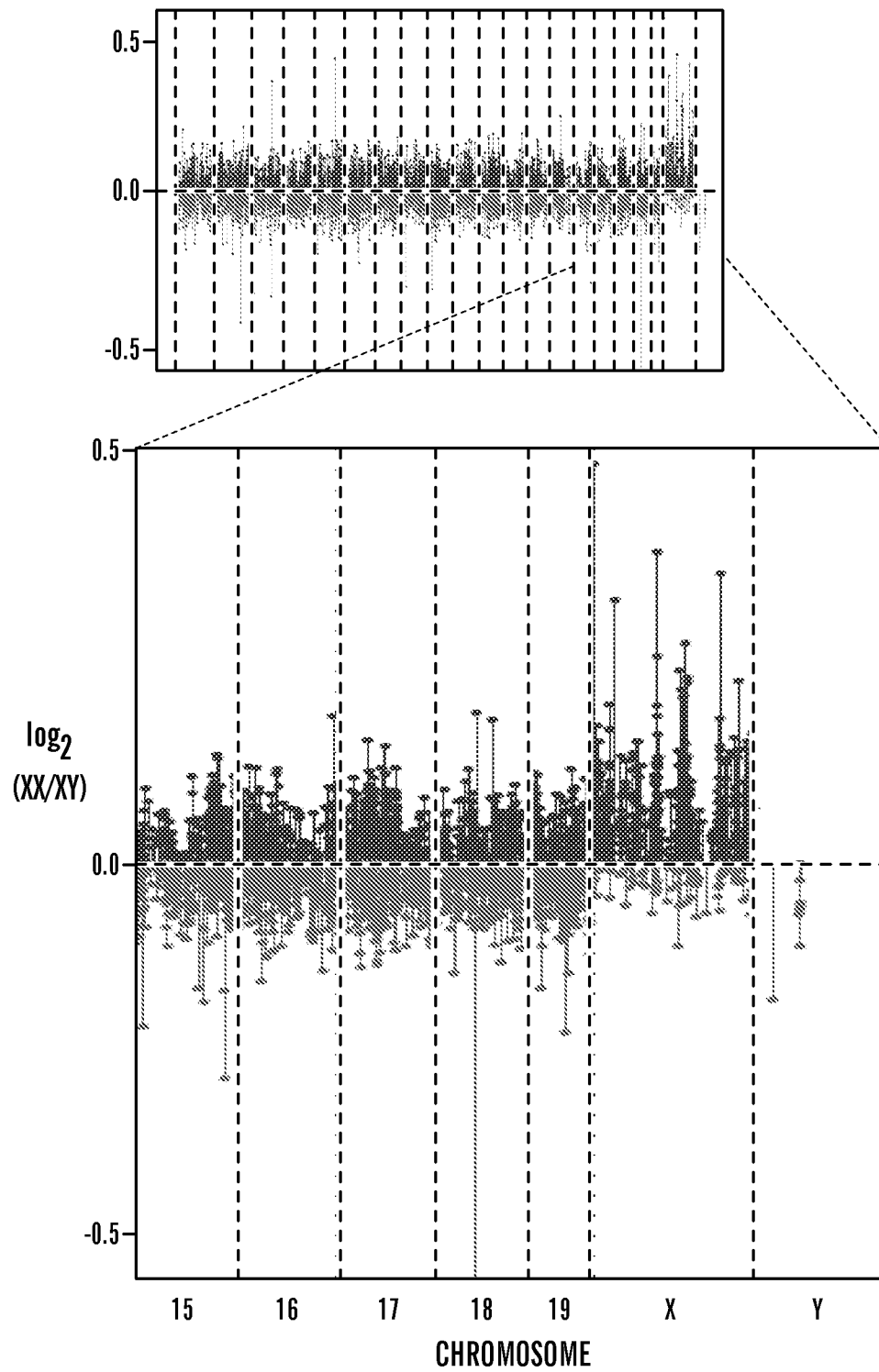


FIG. 5A

7/22



8/22

**FIG. 6**

9/22

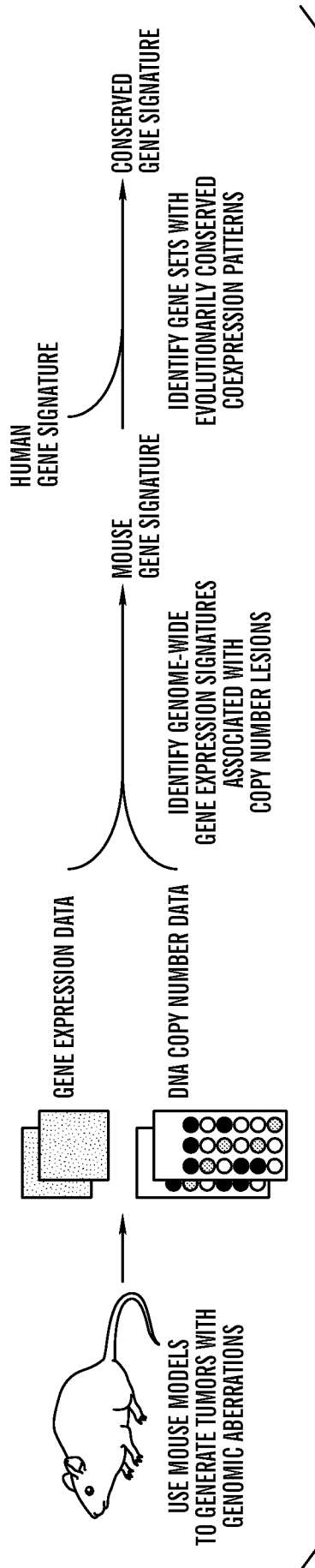


FIG. 7A

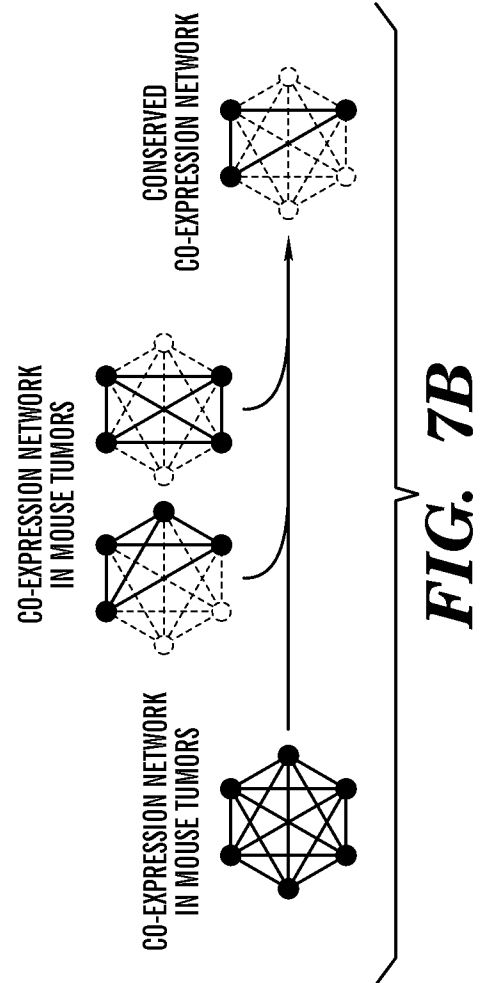


FIG. 7B

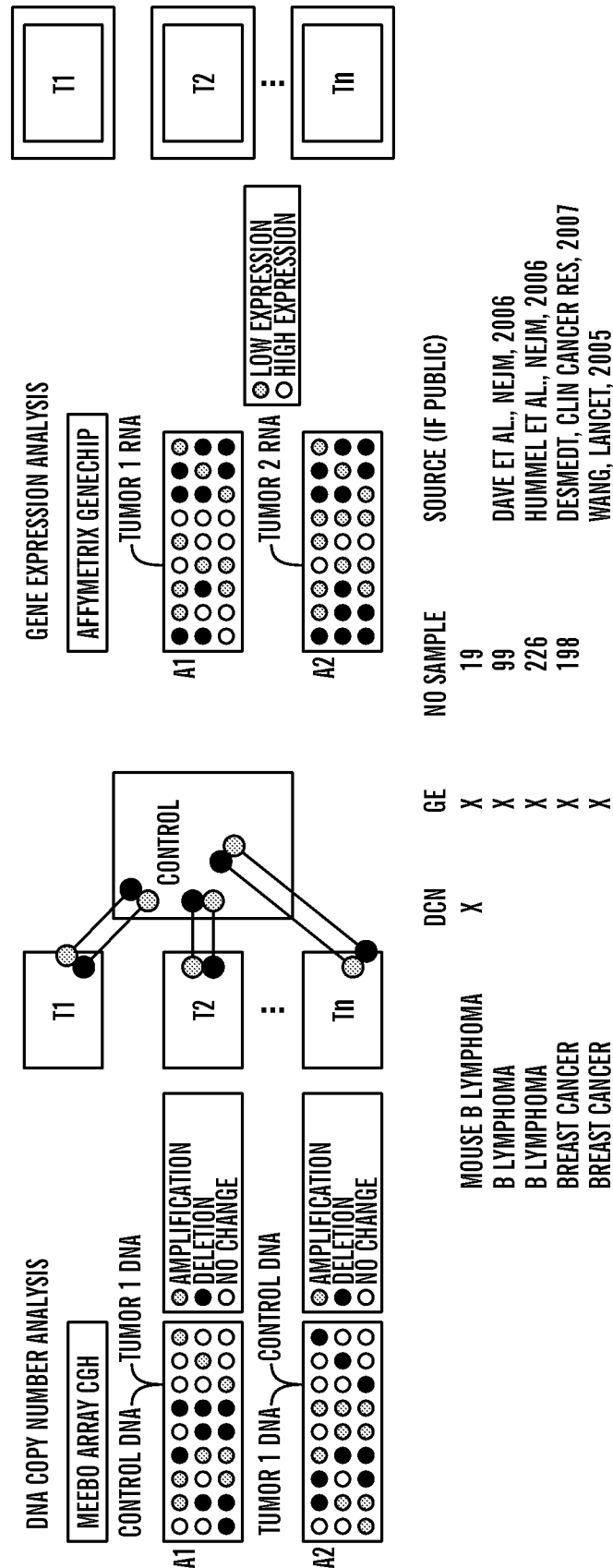
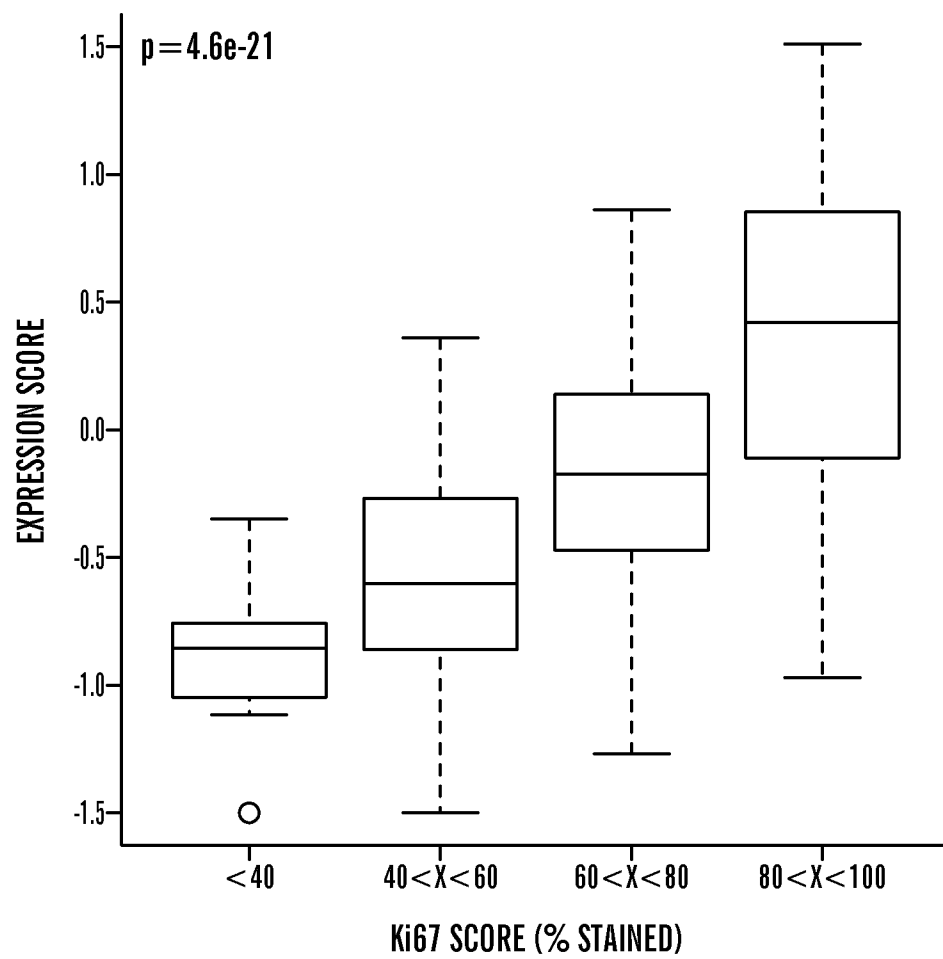
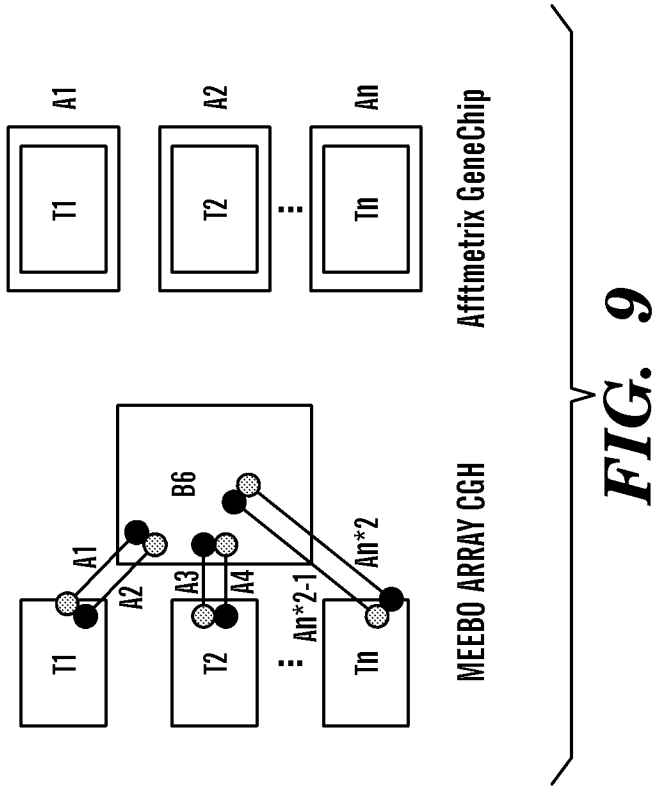
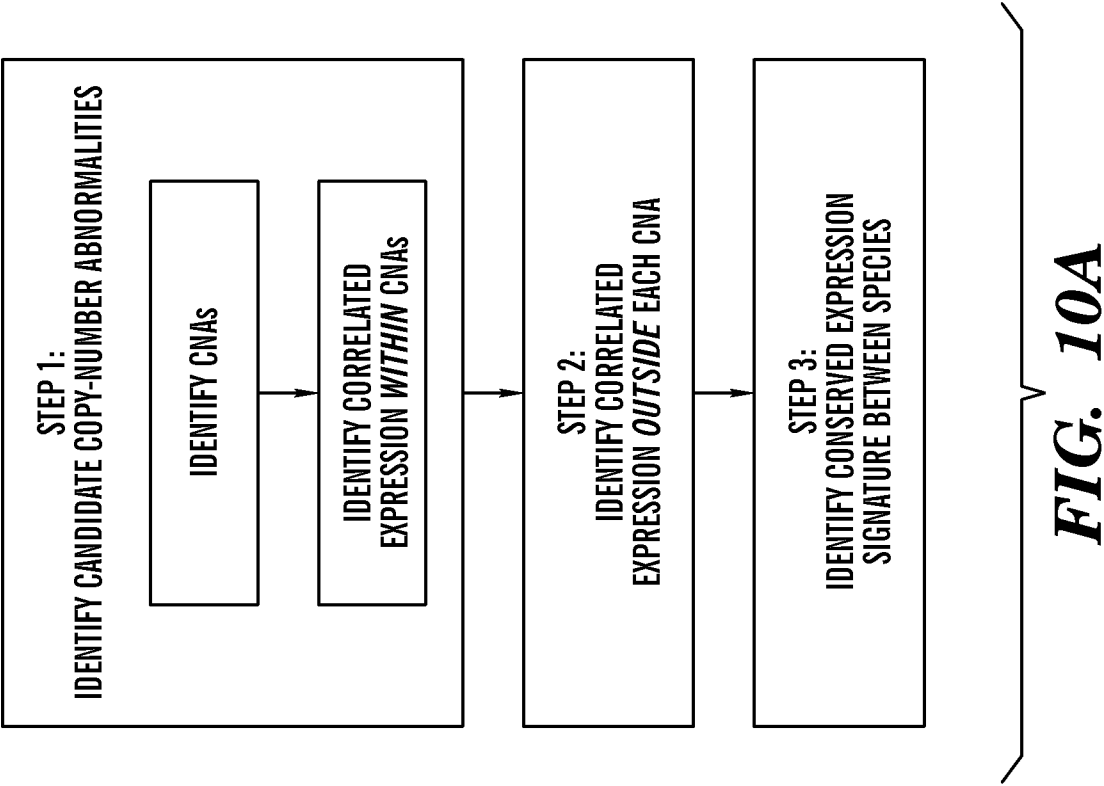
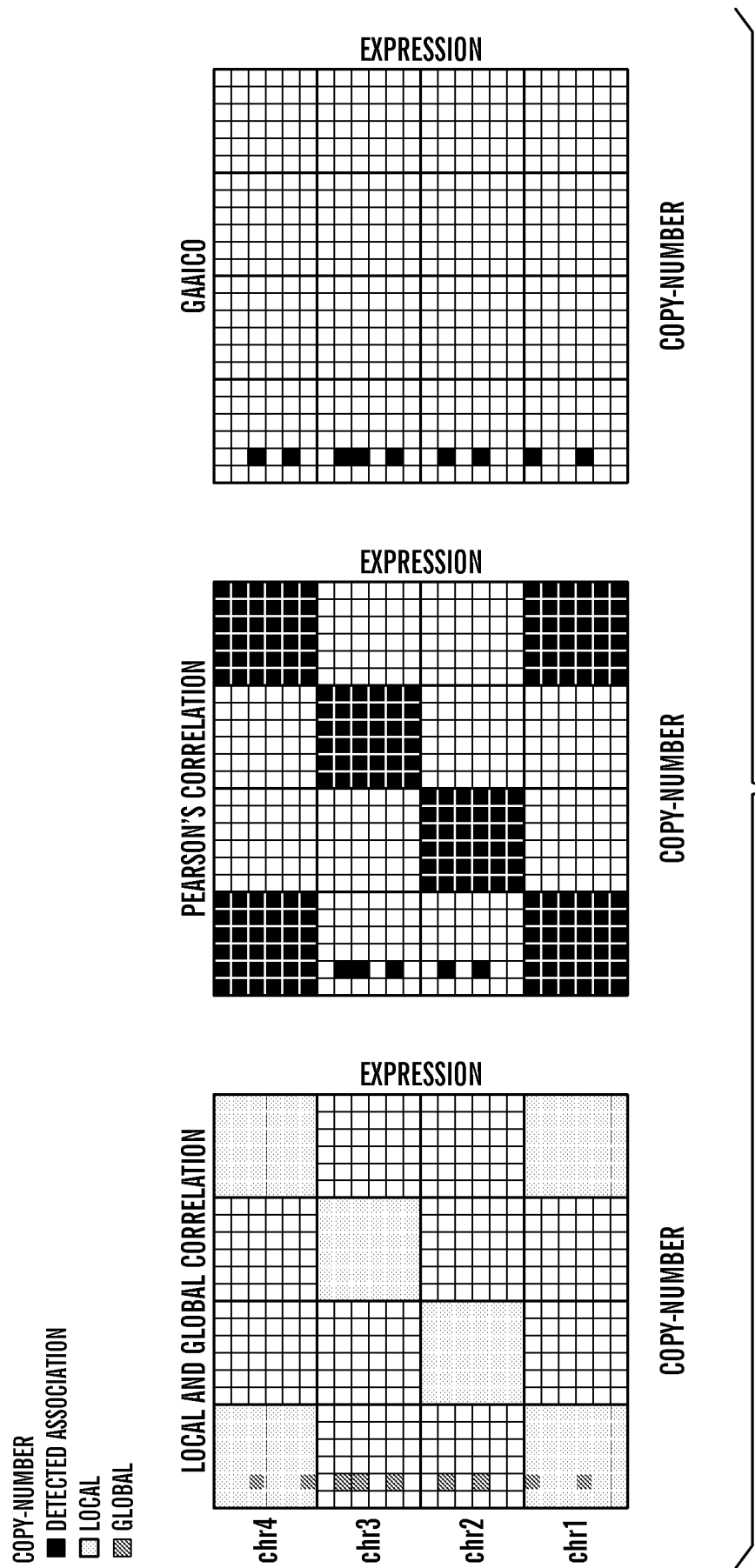


FIG. 7C

11/22

**FIG. 8**





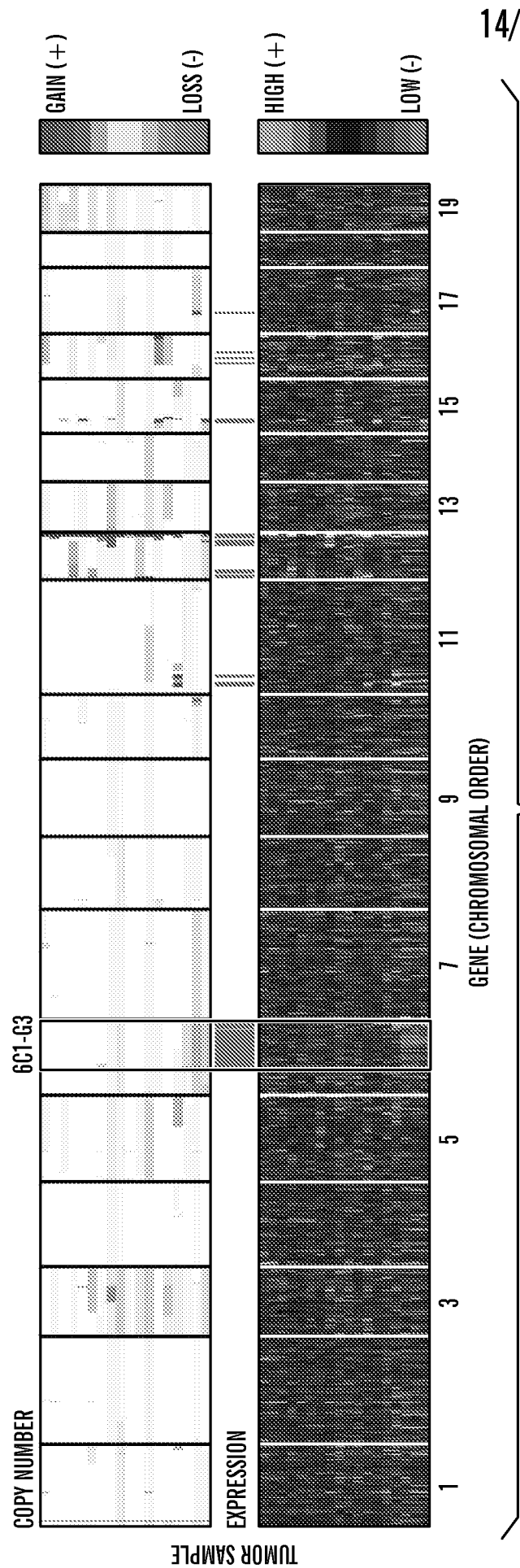
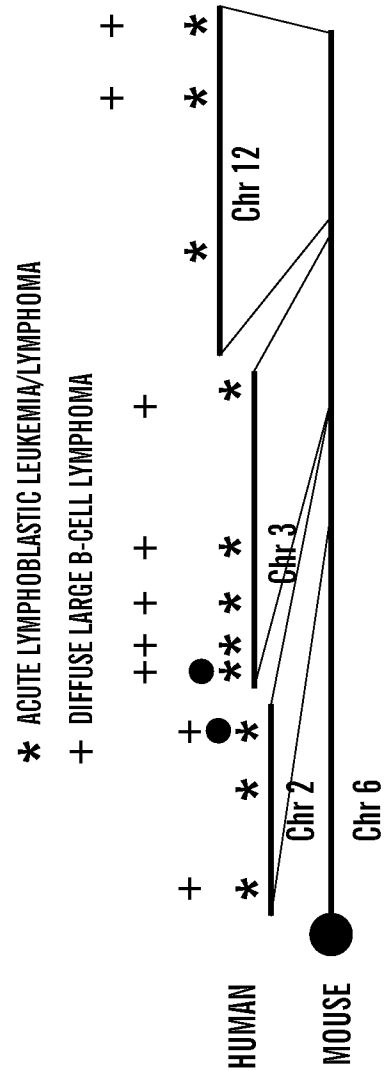
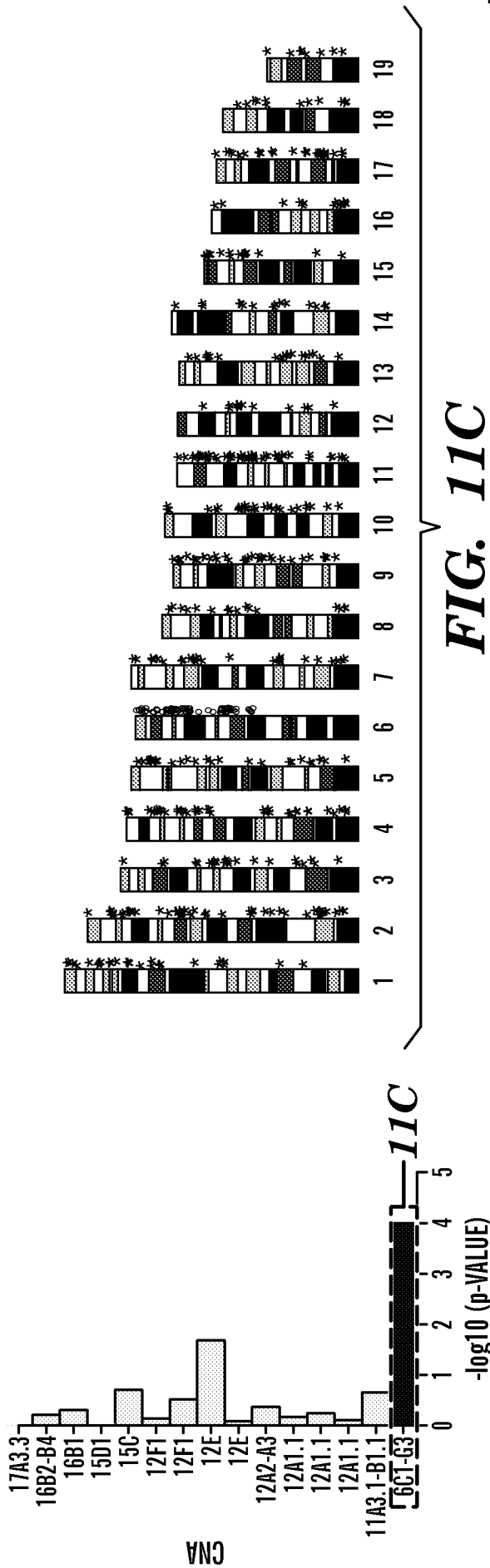
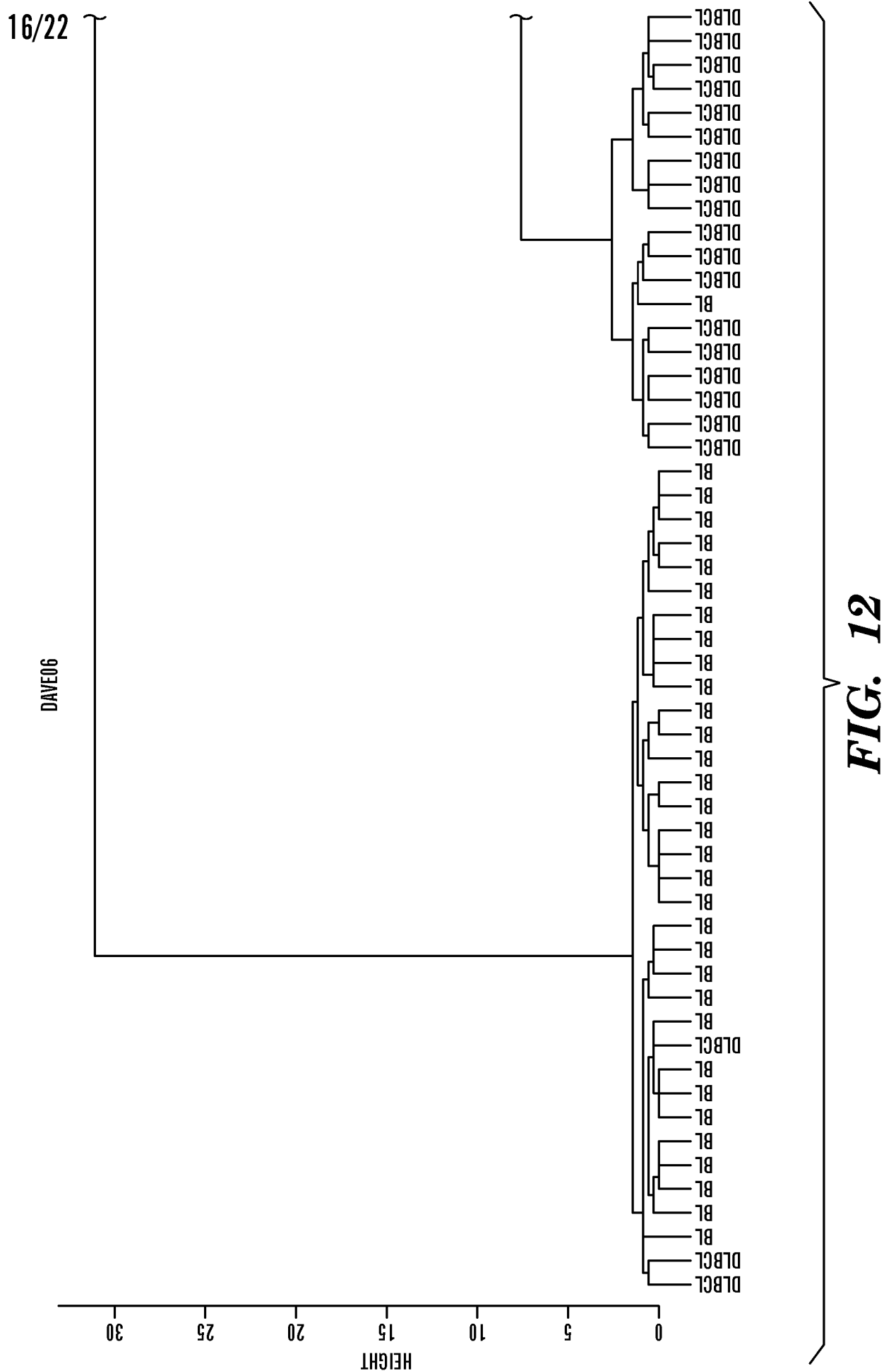
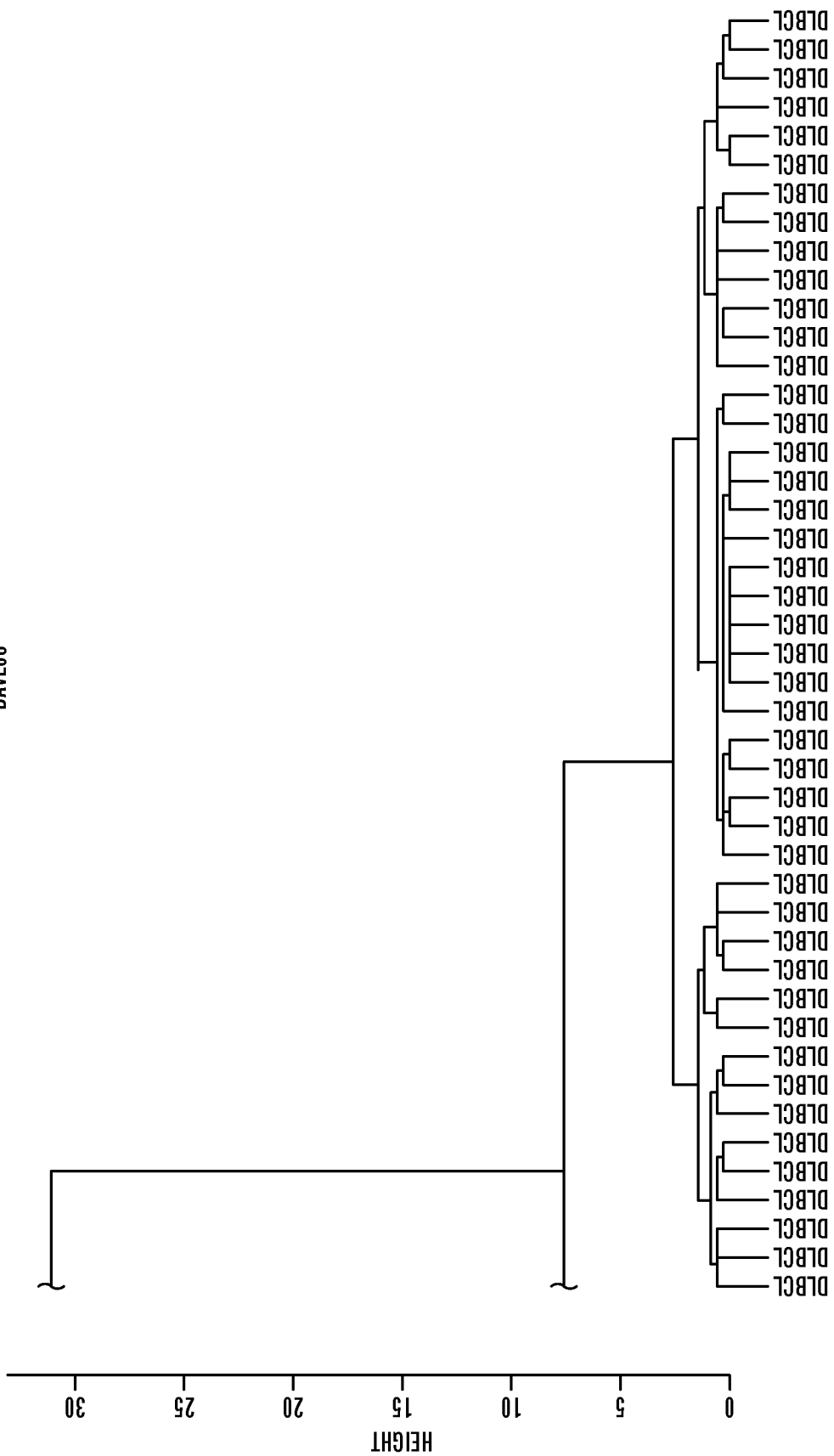


FIG. 11A





DAVE06



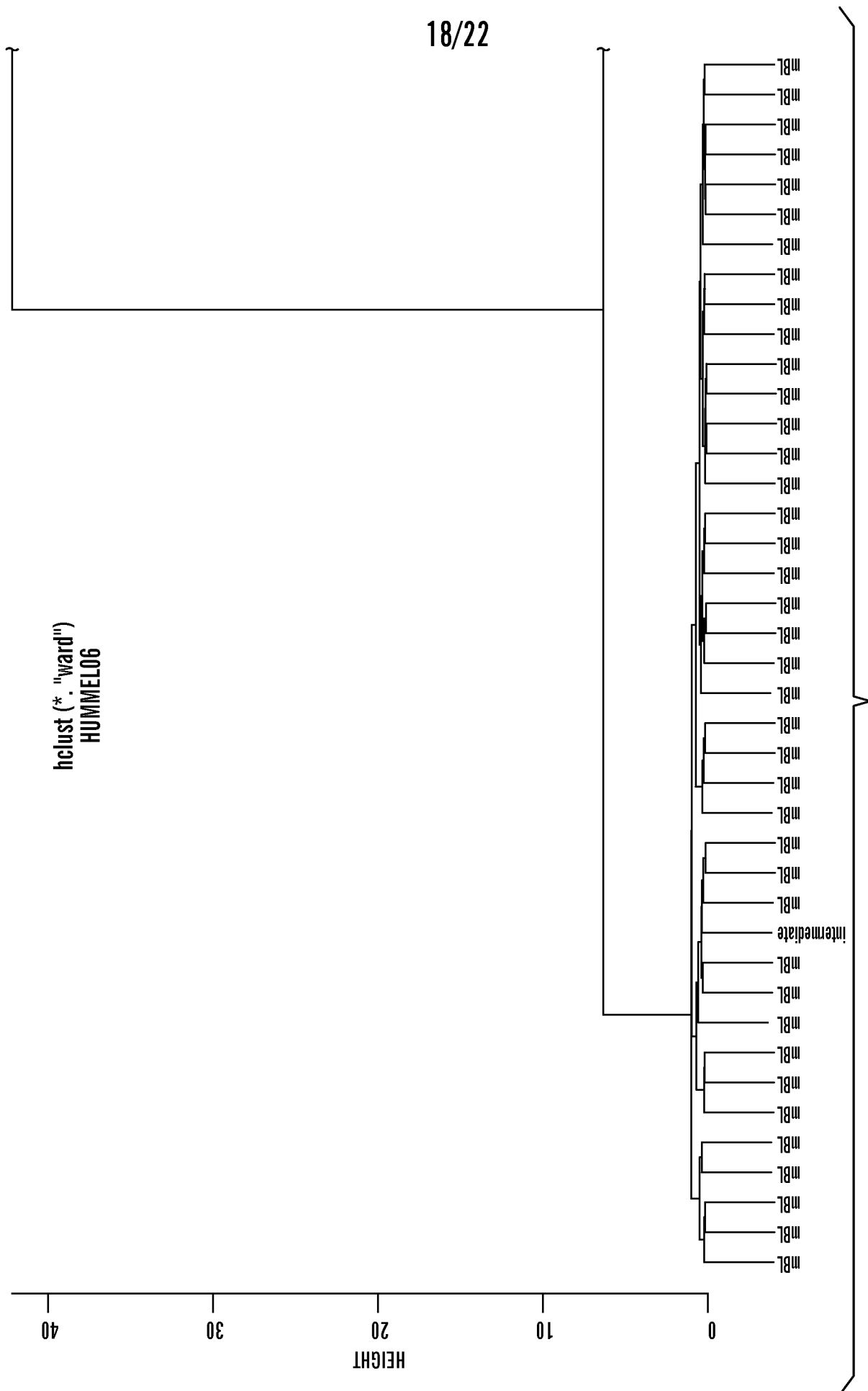
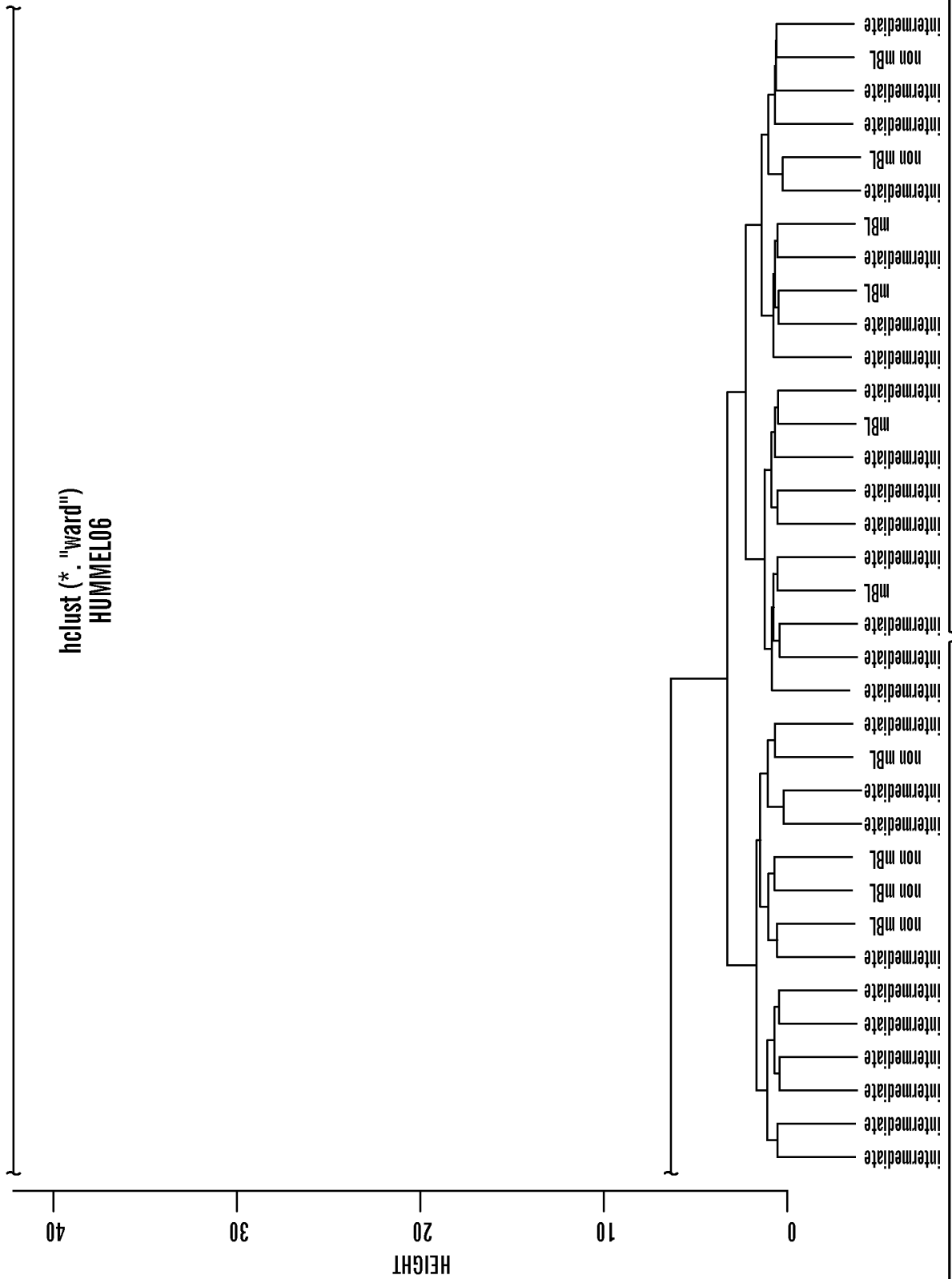


FIG. 12 (cont'd.)



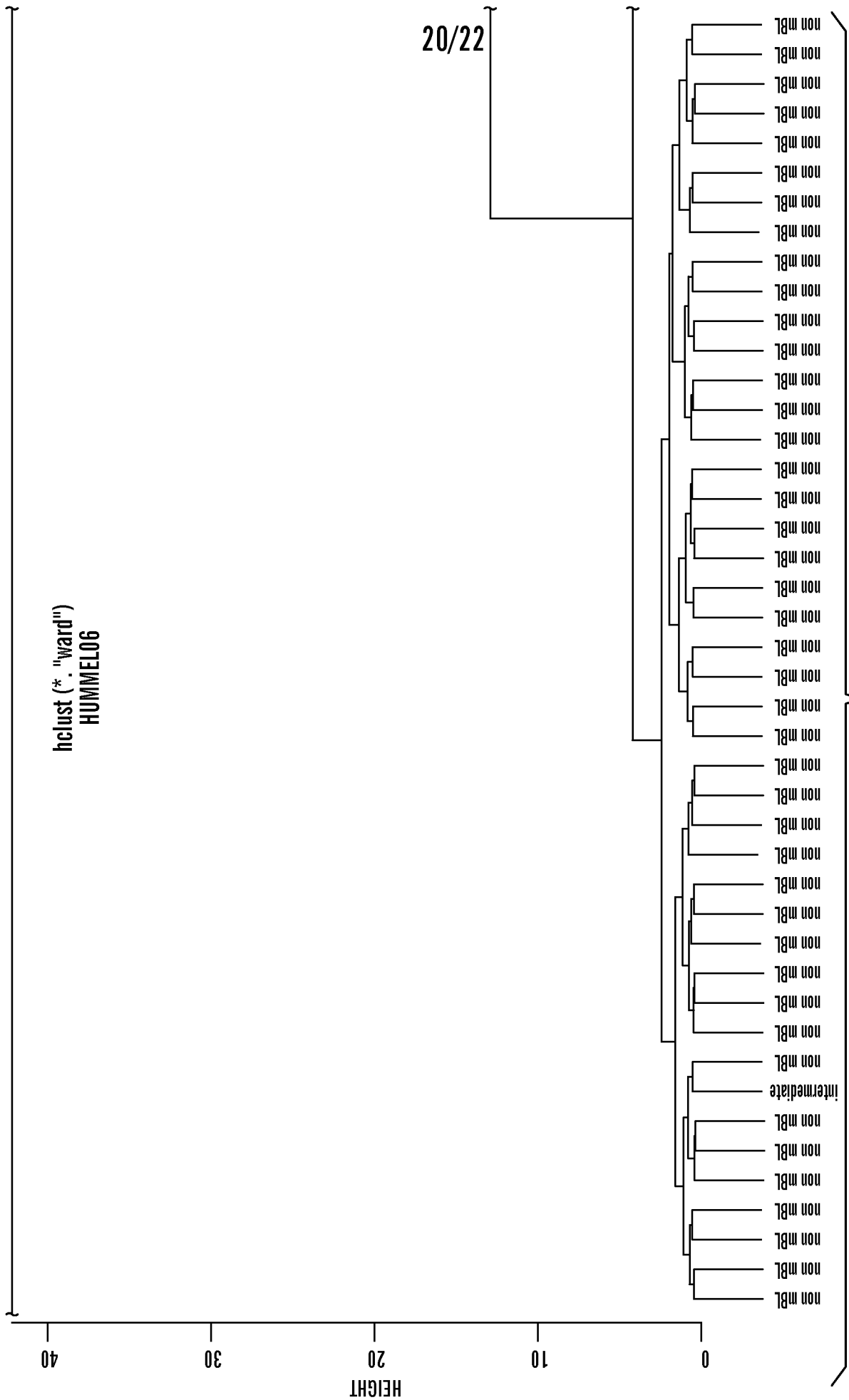


FIG. 12 (cont'd.)

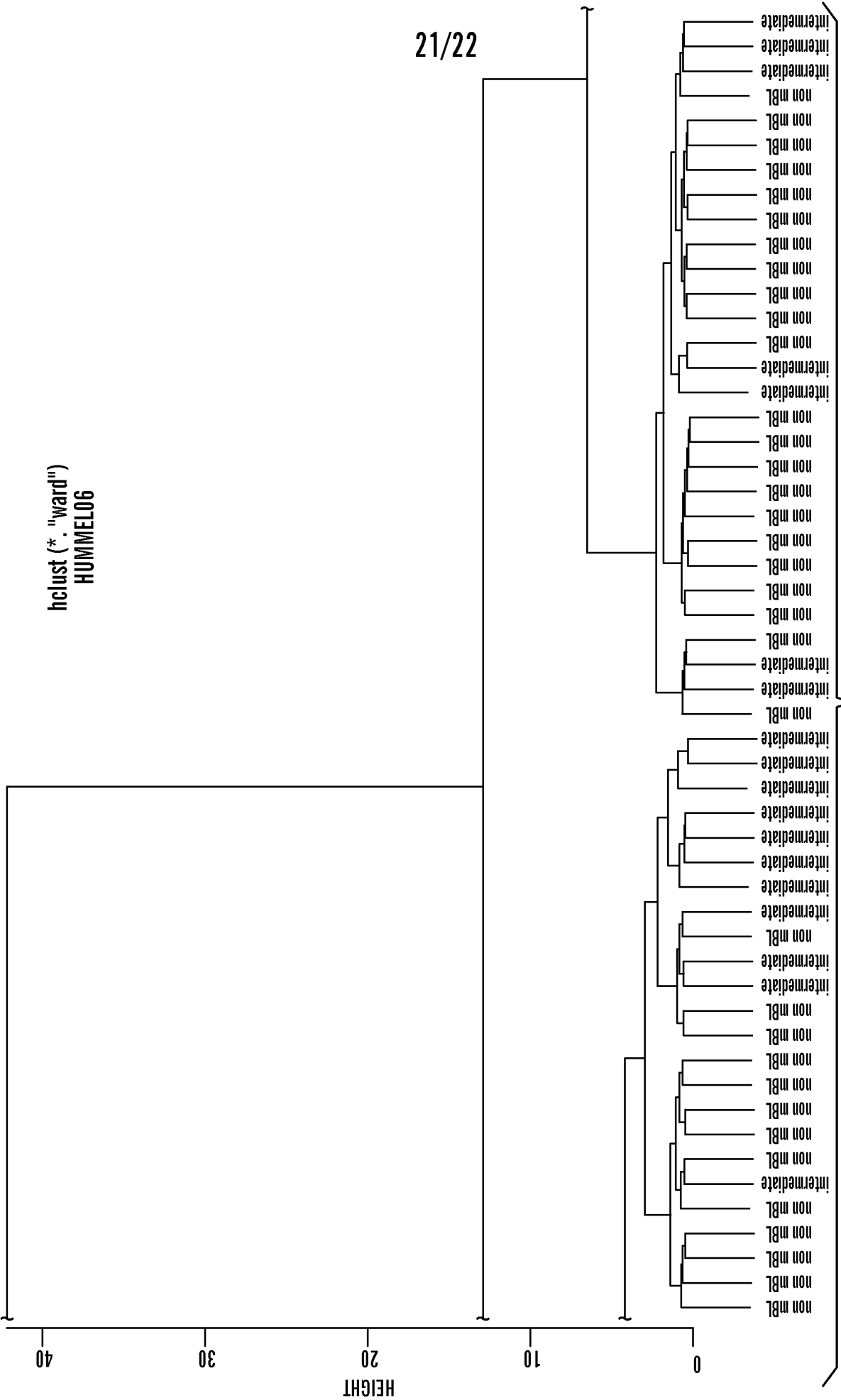


FIG. 12 (cont'd.)



FIG. 12 (cont'd.)