



- (51) International Patent Classification:
C12Q 1/68 (2006.01)
- (21) International Application Number:
PCT/US2015/023197
- (22) International Filing Date:
27 March 2015 (27.03.2015)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
61/971,455 27 March 2014 (27.03.2014) US
61/993,732 15 May 2014 (15.05.2014) US
62/004,727 29 May 2014 (29.05.2014) US
62/092,898 17 December 2014 (17.12.2014) US
- (71) Applicant: **LIFE TECHNOLOGIES CORPORATION**
[US/US]; c/o IP Legal Department Docketing, 5791 Van Allen Way, Carlsbad, California 92008 (US).
- (72) Inventors: **RHODES, Daniel**; Life Technologies Corporation, c/o IP Legal Department Docketing, 5791 Van Allen Way, Carlsbad, California 92008 (US). **SADIS, Seth**; Life Technologies Corporation, c/o IP Legal Department Docketing, 5791 Van Allen Way, Carlsbad, California 92008 (US). **WYNGAARD, Peter**; Life Technologies Corporation, c/o IP Legal Department Docketing, 5791 Van Allen Way, Carlsbad, California 92008 (US). **KHAZANOV, Nikolay**; Life Technologies Corporation, c/o IP Legal Department Docketing, 5791 Van Allen Way, Carlsbad, California 92008 (US). **BANDLA, Santhoshi**; Life Technologies Corporation, c/o IP Legal Department Docketing, 5791 Van Allen Way, Carlsbad, California 92008 (US).

[Continued on next page]

(54) Title: GENE FUSIONS AND GENE VARIANTS ASSOCIATED WITH CANCER

(57) Abstract: The disclosure provides gene fusions, gene variants, and novel associations with disease states, as well as kits, probes, and methods of using the same.

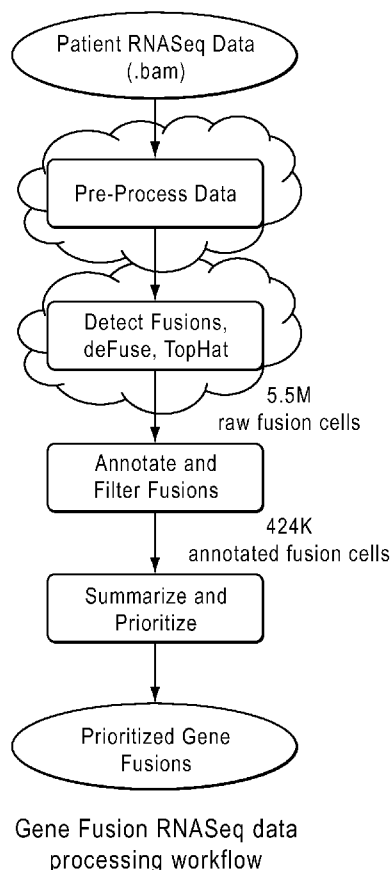


FIG. 1



TOMILO, Mark; Life Technologies Corporation, c/o IP Legal Department Docketing, 5791 Van Allen Way, Carlsbad, California 92008 (US). **EDDY, Sean**; 815 Honey Creek Drive, Ann Arbor, Michigan 48103 (US). **BOWDEN, Emma**; Life Technologies Corporation, c/o IP Legal Department Docketing, 5791 Van Allen Way, Carlsbad, California 92008 (US). **LI, Jia**; 44092 Deep Hollow CT, Northville, Michigan 48168 (US).

(74) **Agent: DUKE, Jennifer**; Life Technologies Corporation, c/o IP Legal Department Docketing, 5791 Van Allen Way, Carlsbad, California 92008 (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, BN, BR, BW, BY, BZ, CA, CH, CL, 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, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY,

TH, 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, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, 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, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

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

— with sequence listing part of description (Rule 5.2(a))

GENE FUSIONS AND GENE VARIANTS ASSOCIATED WITH CANCER

FIELD OF THE INVENTION

[0001] The present invention relates generally to gene fusions and gene variants that are associated with cancer.

BACKGROUND

[0002] Aberrations such as chromosomal translocations and gene variants are frequently found in human cancer cells. Chromosomal translocations may result in a chimeric gene expressing a fusion transcript which is then translated into a fusion protein that affects normal regulatory pathways and stimulates cancer cell growth. Gene variants may also result in aberrant proteins that affect normal regulatory pathways.

[0003] The identification of new fusion genes, new variants of known fusion genes, and gene variants or alleles provides an opportunity for additional diagnostics and cancer treatment targets.

BRIEF SUMMARY OF THE INVENTION

[0004] The disclosure provides novel gene fusion variants and gene fusion-disease state associations. The gene fusions provided herein are associated with certain cancers. The disclosure further provides probes, such as amplification primer sets and detection probes, as well as methods and systems of detection, diagnosis, and treatment and kits that include or detect the gene fusions disclosed herein.

[0005] In certain embodiments, novel mutations and fusions associated with cancer are disclosed in Tables 41-44.

[0006] The disclosure further provides reaction mixtures, sets of probes, kits, methods, isolated nucleic acids comprising or to detect the novel mutations and fusion events of Tables 41-44.

[0007] In one embodiment, the disclosure provides a reaction mixture comprising a probe or a set of probes that specifically recognize a gene fusion selected from Table 1 – Table 3, Table 19, and Table 22. The set of probes can be, for example a set of amplification primers. In another embodiment, provided herein is a reaction mixture that includes a set of primers that flank a gene fusion selected from Table 1 - Table 3, Table

19, and Table 22 in a target nucleic acid. For example, the set of primers can each bind to a target sequence in the human genome within 1000, 750, 500, 250, 100, 90, 80, 75, 70, 65, 50, or 25 nucleotides of opposite sides of the one of the fusion breakpoints identified in Tables 4-6, 20, and 23. The reaction mixture of this embodiment can further include a detector probe that binds to either side of a breakpoint in a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22, or that binds a binding region that spans the breakpoint in a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22, including specific embodiments where the breakpoint is identified in Tables 4-6, 20, and 23. In exemplary embodiments, the detector probe binds to a target sequence in the human genome within 1000, 750, 500, 250, 100, 90, 80, 75, 70, 60, 50, or 25 nucleotides of one of the fusion breakpoints identified in Tables 4-6, 20, and 23. The reaction mixture that includes a detector probe, or does not include a detector probe, can further include a polymerase, a reverse transcriptase, dNTPs, and/or a uracil DNA deglycosylase (UDG). The polymerase, the reverse transcriptase, and the UDG are typically not from human origin. The polymerase in illustrative embodiments is a thermostable polymerase such as a Taq polymerase. In certain embodiments, the dNTPs in the reaction mixture include dUTP, and the reaction mixture can in certain examples, be devoid of dTTP. Furthermore, the reaction mixture can include an amplicon, such as a DNA amplicon that includes one or more deoxyuridine ("dU") residues. In certain embodiments the reaction mixture includes a DNA amplicon that includes one or more dU residues for every deoxythymidine residue in the corresponding human genomic sequence. In certain embodiments, the amplicon includes a segment for which a corresponding sequence is not found in the human genome, such as, for example, a DNA barcode sequence. The non-human segment can be for example, 5-10,000, 5-5000, 5-1000, 5-500, 5-100, 5-50, 5-25, 5-10, 10-10,000, 10-5000, 10-1000, 10-500, 10-100, 10-50, or 10-25 nucleotides in length. In certain embodiments, the amplicon includes segment that corresponds to the region of the human genome that spans an intron, but the amplicon does not include a segment corresponding to the intron. The reaction mixture can further include a target nucleic acid, for example a human target nucleic acid. The human target nucleic acid can be, for example, isolated from a biological sample from a person suspected of having a cancer selected from: BLCA = bladder carcinoma, BRCA = breast carcinoma, CESC = cervical cell carcinoma, COAD = colon adenocarcinoma, GBM = glioblastoma multiforme, HNSC = head and neck squamous cell carcinoma, KIRC = clear cell renal cell carcinoma, KIRP = kidney renal papillary cell carcinoma, LAML = acute myeloid

leukemia, LGG = brain lower grade glioma, LIHC = liver hepatocellular carcinoma, LUAD = lung adenocarcinoma, LUSC = squamous cell lung carcinoma, OV = ovarian serous adenocarcinoma, PRAD = prostate adenocarcinoma, READ = rectal adenocarcinoma, SKCM = cutaneous melanoma, STAD = stomach adenocarcinoma, THCA = thyroid carcinoma, and UCEC = uterine corpus endometrioid carcinoma. In certain embodiments, the target nucleic acid is from a tumor, for example a tumor of one of the cancer types listed in the preceding sentence.

[0008] In another embodiment, a set of probes that specifically recognizes a nucleic acid comprising at least one of SEQ ID NOs: 1-289 (gene fusions) is provided. In another embodiment, provided herein is a set of primers that specifically amplify a target nucleic acid that includes at least 25, 30, 40, 50, 75, 100, 125, 150, 200, or all of SEQ ID NOs: 1-289, or that amplifies up to 25, 30, 40, 50, 75, 100, 125, 150, 200, or all of SEQ ID NOs: 1-289. In another embodiment, provided herein is a qPCR assay, such as a TaqMan™ assay or a Molecular Beacons™ assay, that specifically amplifies and detects a target nucleic acid that includes at least 25, 30, 40, 50, 75, 100, 125, 150, 200, or all of SEQ ID NOs: 1-289.

[0009] The disclosure also provides an isolated nucleic acid comprising at least one sequence selected from a segment that includes at least 25, 30, 40, 50, 75, 100, 125, 150 200, or all of SEQ ID NOs: 1-289 or that includes up to 25, 30, 40, 50, 75, 100, 125, 150, 200, or all of SEQ ID NOs: 1-289. The isolated nucleic acid can include a first primer on a 5' end. Furthermore, the nucleic acid can be single stranded or double stranded. In certain embodiments, the isolated nucleic acid includes a segment for which a corresponding sequence is not found in the human genome, such as, for example, a DNA barcode sequence. The segment can be for example, 5-10,000, 5-5000, 5-1000, 5-500, 5-100, 5-50, 5-25, 5-10, 10-10,000, 10-5000, 10-1000, 10-500, 10-100, 10-50, or 10-25 nucleotides in length.

[0010] The disclosure, in other embodiments, provides a kit that includes a detector probe and/or a set of probes, for example, a set of amplification primers, that specifically recognize a nucleic acid comprising a breakpoint for a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22. For example, in certain embodiments the detector probe or set of amplification primers are designed to amplify and/or detect a nucleic acid that includes up to 25, 30, 40, 50, 75, 100, 125, 150, 200, or all of at least one of SEQ ID NOs: 1-289. The kit can further include, in one or more separate or in the same vessel, at least one component from an amplification reaction mixture, such

as a polymerase, dNTPs, a reverse transcriptase, and/or UDG, typically the reverse transcriptase, polymerase and UDG are not from human origin. In certain embodiments, the dNTPs include dUTP, and in illustrative examples are devoid of dTTP. The polymerase in illustrative embodiments is a thermostable polymerase such as a Taq polymerase. Furthermore, the kit can include a control nucleic acid. For example the control nucleic acid can include a sequence that includes the break point in a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22, such as a nucleic acid that includes at least 25, 30, 40, 50, 75, 100, 125, 150, 200, or all of SEQ ID NOs: 1-289 or a nucleic acid that includes up to 25, 30, 40, 50, 75, 100, 125, 150, 200, or all of SEQ ID NOs: 1-289.

[0011] A method of detecting a cancer is provided comprising amplifying a nucleic acid that spans a breakpoint in a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22, for example the nucleic acid can include a sequence selected from SEQ ID NOs: 1-289, and detecting the presence of the nucleic acid, wherein the presence of the nucleic acid indicates a cancer is present in the sample. In another method, provided herein is a method of detecting a cancer selected from, bladder, colon, breast, endometrial, melanoma, ovarian, glioblastoma, glioma, leukemia, renal cell carcinoma, thyroid, and prostate adenocarcinoma that includes generating an amplicon that includes a sequence selected from SEQ ID NOs: 1-289 and detecting the presence of the amplicon, wherein the presence of the amplicon indicates bladder, colon, melanoma, ovarian, glioblastoma, lung, glioma, leukemia, renal cell carcinoma, thyroid, endometrial endometrioid adenocarcinoma, breast and prostate adenocarcinoma is present in the sample. The amplicon typically includes primers that were extended to form the amplicon. The cancer is selected from bladder urothelial carcinoma, breast carcinoma, endometrial endometrioid adenocarcinoma, colon adenocarcinoma, glioblastoma multiforme, clear cell renal cell carcinoma, papillary renal cell carcinoma, acute myeloid leukemia, brain lower grade glioma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, rectal cutaneous melanoma, and thyroid gland carcinoma. The amplicon that is generated, in certain illustrative embodiments is a DNA amplicon that includes dU residues, and in certain examples includes no dT residues. In the methods provided in this paragraph, the amplicon can be generated using reaction mixtures provided herein. In certain embodiments, the method includes detecting expression of a nucleic acid that spans a breakpoint in a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22. Methods for

detecting expression typically include a step of isolating RNA from a sample, such as a tumor sample, which can be a formalin fixed sample in illustrative embodiments.

[0012] In one embodiment, the reaction mixture includes a dye selected from SYBR Green, SBYR Greener, Fluorescein, Oregon Green, FAM, TET, JOE, VIC, Yakima Yellow, HEX, Cy3, Bodipy TMR, NED, TAMRA, Cy3.5, ROX, Texas Red, LightCycler Red, Bodipy 630/650, Alexa Fluor 647, Cy5, Alexa Fluor 660, or Cy 5.5. In certain embodiments, the dye is attached to a detectably-labeled probe in the reaction mixture. In other embodiments, the dye is bound to the amplicon directly or through a detectably-labeled probe.

[0013] A kit comprising a probe or a set of probes, for example, a detectable probe or a set of amplification primers that specifically recognize a nucleic acid comprising a break point from Tables 4-6, 20, and 23 is provided. The kit can further include, in the same vessel, or in certain preferred embodiments, in a separate vessel, a component from an amplification reaction mixture, such as a polymerase, typically not from human origin, dNTPs, and/or UDG. Furthermore, the kit can include a control nucleic acid. For example the control nucleic acid can include a sequence that includes a break point selected from Tables 4-6, 20, and 23.

[0014] In another embodiment, provided herein a gene fusion that includes the gene fusions identified in Tables 1-3, 19, and 22. In illustrative embodiments, the gene fusions include one of the breakpoints identified in Tables 4-6, 20, and 23. Accordingly, provided herein is an isolated gene fusion nucleic acid of between 100 and 10,000 nucleotides in length and comprising at least 25 nucleotides on either side of one of the break points in Tables 4-6, 20, and 23.

[0015] In a related embodiment, provided herein is an isolated gene fusion nucleic acid comprising at least one of the break points in Tables 4-6, 20, and 23. In certain embodiments, the isolated gene fusion nucleic acid comprises at least 25, 30, 40, 50, 75, 100, 125, 150, 200, or all of SEQ ID NOs: 1-289 or a nucleic acid that includes up to 25, 30, 40, 50, 75, 100, 125, 150, 200, or all of SEQ ID NOs: 1-289. The isolated gene fusion nucleic acid can have a length, for example, of between 50 and 100,000 nucleotides, between 100 and 50,000 nucleotides, between 100 and 25,000 nucleotides, between 100 and 10,000 nucleotides, between 100 and 5,000 nucleotides, between 100 and 2500 nucleotides, between 100 and 1,000 nucleotides, between 100 and 500 nucleotides, between 100 and 250 nucleotides, between 100 and 200 nucleotides, between 250 and 10,000 nucleotides, between 250 and 5,000 nucleotides,

between 250 and 1,000 nucleotides, or between 250 and 500 nucleotides. In certain aspects, the isolated gene fusion nucleic acid is DNA. In certain illustrative embodiments, the isolated nucleic gene fusion is devoid of intron sequences but spans a region that in the genome includes one or more introns. In certain embodiments, the isolated gene fusion nucleic acid is a cDNA.

[0016] In another embodiment, an isolated gene fusion nucleic acid is provided comprising at least one of the break points in Tables 4-6, 20, and 23.

[0017] In another embodiment is a method to detect a cancer selected from bladder urothelial carcinoma, breast carcinoma, endometrial endometrioid adenocarcinoma, colon adenocarcinoma, glioblastoma multiforme, clear cell renal cell carcinoma, papillary renal cell carcinoma, acute myeloid leukemia, brain lower grade glioma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, rectal cutaneous melanoma, and thyroid gland carcinoma in a sample by detecting the presence of a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22.

[0018] The disclosure provides novel gene variants and gene variant-disease state associations. The gene variants can have one or more mutations that result in a variant protein. The gene variants provided herein are associated with certain cancers. The gene variants result in protein variants. The disclosure further provides probes, such as amplification primer sets and detection probes, as well as methods of detection, diagnosis, and treatment and kits that include or detect the gene variants disclosed herein.

[0019] In one embodiment, the disclosure provides a composition and a kit comprising a set of probes that specifically recognize the nucleotide sequence that encodes a gene variant selected from Table 7 and/or Table 11. The set of probes can be, for example a set of amplification primers. In another embodiment, provided herein is a composition that includes a set of primers that flank a gene variant that encodes one or more variants in Table 7 and/or Table 11. The reaction mixture of this embodiment can further include a detector probe that binds to a nucleotide sequence including a gene variant selected from Table 7 and/or Table 11. The reaction mixture that includes a detector probe or does not include a detector probe, can further include a polymerase, dNTPs, and/or a uracil DNA deglycosylase (UDG). The polymerase and UDG are typically not from a human origin. The reaction mixture can further include a target nucleic acid, for example a human target nucleic acid. The human target nucleic acid can be, for example, isolated from a biological sample from a person suspected of

having a cancer. The cancer can be selected from: BLCA = bladder carcinoma, BRCA = breast carcinoma, CESC = cervical cell carcinoma, COAD = colon adenocarcinoma, GBM = glioblastoma multiforme, HNSC = head and neck squamous cell carcinoma, KIRC = clear cell renal cell carcinoma, KIRP = kidney renal papillary cell carcinoma, LAML = acute myeloid leukemia, LGG = brain lower grade glioma, LIHC = liver hepatocellular carcinoma, LUAD = lung adenocarcinoma, LUSC = squamous cell lung carcinoma, OV = ovarian serous adenocarcinoma, PRAD = prostate adenocarcinoma, READ = rectal adenocarcinoma, SKCM = cutaneous melanoma, STAD = stomach adenocarcinoma, THCA = thyroid carcinoma, and UCEC = uterine corpus endometrioid carcinoma.

[0020] The nucleotide sequence that encodes one or more gene variants in Table 7 and/or Table 11 can be any size that encompasses the variation. For example, the nucleotide sequence can be any size that can be easily copied using a primer and/or detected using a probe.

[0021] In another embodiment, a set of probes that specifically recognize a nucleic acid coding for a gene variant selected from Table 7 and/or Table 11 (gene variants) is provided. In another embodiment, provided herein is a set of primers that specifically amplify a target nucleic acid that codes for a gene variant selected from Table 7 and/or Table 11. In another embodiment, provided herein is a qPCR assay, such as, but not limited to, a TaqMan™ assay, a Scorpions assay, or a Molecular Beacons™ assay that specifically amplifies and detects a target nucleic acid that codes for a gene variant selected from Table 7 and/or Table 11.

[0022] The disclosure also provides an isolated nucleic acid comprising at least one sequence that codes for one or more gene variants selected from Table 7 and/or Table 11. The isolated nucleic acid can include a first primer on a 5' end. Furthermore, the nucleic acid can be single stranded or double stranded.

[0023] The disclosure, in other embodiments, provides a kit that includes a detector probe and/or a set of probes, for example, a set of amplification primers that specifically recognize a nucleic acid that codes for a gene variant selected from Table 7 and/or Table 11. For example, in certain embodiments the detector probe or set of amplification primers are designed to amplify and/or detect a nucleic acid that codes for a variant in Table 7 and/or Table 11. The kit can further include, in a separate or in the same vessel, a component from an amplification reaction mixture, such as a polymerase, typically not from human origin, dNTPs, and/or UDG. Furthermore, the kit

can include a control nucleic acid. For example the control nucleic acid can include a sequence that includes the gene variant selected from Table 7 and/or Table 11.

[0024] A method of detecting a cancer is provided comprising amplifying a nucleic acid that encodes a gene variant selected from Table 7 and/or Table 11, for example the nucleic acid can include a sequence from one of the accession numbers in Table 7 and/or Table 11 except that the sequence contains the variant that codes for the gene variants in Table 7 and/or Table 11, and detecting the presence of the nucleic acid, wherein the presence of the nucleic acid indicates a cancer is present in the sample. In another method, provided herein is a method of detecting a cancer that includes generating an amplicon that includes a sequence encoding a variant selected from Table 7 and/or Table 11, and detecting the presence of the nucleic acid, wherein the presence of the nucleic acid indicates a cancer or cancer cell is present in the sample. The amplicon typically includes primers that are extended to form the amplicon. The cancer is selected from bladder carcinoma, breast carcinoma, cervical cell carcinoma, colon adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, clear cell renal cell carcinoma, kidney renal papillary cell carcinoma, acute myeloid leukemia, brain lower grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, squamous cell lung carcinoma, ovarian serous adenocarcinoma, prostate adenocarcinoma, rectal adenocarcinoma, cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, and uterine corpus endometrioid carcinoma.

[0025] A kit comprising a set of probes, for example, a set of amplification primers that specifically recognize a nucleic acid comprising a gene variant from Table 7 and/or Table 11 is provided. The kit can further include, in a separate or in the same vessel, a component from an amplification reaction mixture, such as a polymerase, typically not from human origin, dNTPs, and/or UDG. Furthermore, the kit can include a control nucleic acid. For example the control nucleic acid can include a sequence that includes the gene variant from Table 7 and/or Table 11.

[0026] In certain embodiments, a set of probes that specifically recognize a nucleic acid comprising a gene variant from Table 7 and/or Table 11 is provided.

[0027] In another embodiment, a gene variant is provided comprising at least one of the gene variants in Table 7 and/or Table 11.

[0028] In another embodiment is a method to detect a cancer selected from bladder carcinoma, breast carcinoma, cervical cell carcinoma, colon adenocarcinoma,

glioblastoma multiforme, head and neck squamous cell carcinoma, clear cell renal cell carcinoma, kidney renal papillary cell carcinoma, acute myeloid leukemia, brain lower grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, squamous cell lung carcinoma, ovarian serous adenocarcinoma, prostate adenocarcinoma, rectal adenocarcinoma, cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, and uterine corpus endometrioid carcinoma in a sample by detecting the presence of a gene variant selected from Table 7 and/or Table 11. Gene variants can include, but are not limited to, ZNF479 variants R11Q, R295K, R295T, R295I, R345I, R345T, K438T, and T466K.

[0029] In another embodiment, a method of delivering a drug to a subject is provided, wherein the method includes detecting a genetic event identified in Table 15, and treating the subject with a drug, wherein the drug is believed to positively affect the clinical outcome of patients having the genetic event. In illustrative embodiments, the genetic event is associated with a gene found in Table 8 and the drug is listed in Table 8 as a companion for that gene. In another embodiment, provided herein is a method for determining if a subject receives a drug, the method includes detecting a genetic event identified in Table 15, and then delivering a drug to the subject if the detected genetic event is listed in Table 15 as associated with a poor prognosis, wherein the drug is believed to positively affect the clinical outcome of patients having the genetic event. In illustrative embodiments, the genetic event is associated with a gene found in Table 8 and the drug is listed in Table 8 as a companion for that gene.

[0030] In one embodiment, a kit is provided, wherein the kit comprises a set of probes, wherein each probe specifically hybridizes to a nucleic acid comprising a breakpoint from Tables 4-6, 20, and 23.

[0031] In one embodiment, a method is provided, the method comprising: amplifying a nucleic acid comprising at least one gene fusion from Tables 1-3, 19, and 22 from a sample; and detecting the presence of the at least one gene fusion by at least one of: contacting the composition with at least one probe, wherein each probe specifically hybridizes to the nucleic acid, or observing the presence of a non-natural or non-native chemical structure in the nucleic acid; wherein detecting the presence of the at least one gene fusion indicates that at least one cancer from Tables 1-3, 19, and 22 is present in the sample.

[0032] In one embodiment, a system is provided, the system comprising a nucleic acid amplifier configured to amplify a nucleic acid comprising at least one gene fusion

from Tables 1-3, 19, and 22 from a sample, to yield an amplified nucleic acid; a detector configured to detect the presence of the at least one gene fusion in the amplified nucleic acid by at least one of (i) contacting the composition with at least one probe, wherein each probe specifically hybridizes to the nucleic acid, or (ii) observing the presence of a non-natural or non-native chemical structure in the nucleic acid, and further configured to transmit a detection indication; and a computer system configured to receive the detection indication and determine that at least one cancer from Tables 1-3, 19, and 22 is present in the sample, based on the detection indication.

[0033] In one embodiment, a non-transitory computer readable program storage unit is provided, the non-transitory computer readable program storage unit encoded with instructions that, when executed by a computer, perform a method, comprising receiving an input comprising at least a cancer type and an event type, wherein the cancer type is selected from Table 15 and the event type is selected from Table 15; querying a database for at least one entry comprising a plurality of fields, wherein the plurality of fields comprises at least one of the cancer type and the event type; and transmitting an output comprising at least one field of the plurality from the at least one entry, wherein the at least one field comprises at least one gene, at least one druggable gene, at least one drug targeting the at least one druggable gene, or a prognosis.

[0034] In one embodiment, a method is provided, wherein the method comprises administering to a patient having at least one gene fusion selected from the gene fusions listed in Tables 1-3, 19, and 22 at least one drug selected from the drugs listed in Tables 8, 16-17, 21, and 24.

[0035] In one embodiment, a method is provided, wherein the method comprises contacting a nucleic acid sample from a patient with a reaction mixture comprising a first primer complementary to a first gene and a second primer complementary to a second gene, wherein a fusion of the first gene and the second gene is detectable by the presence of an amplicon generated by the first primer and the second primer, wherein the fusion comprises a breakpoint selected from the breakpoints listed in Tables 4-6, 20, and 23.

[0036] In one embodiment, a non-transitory computer readable program storage unit is provided, the non-transitory computer readable program storage unit encoded with instructions that, when executed by a computer, perform a method, comprising receiving RNA sequence data from at least one cancer cell line; running at least one gene fusion caller on the sequence data, to identify possible breakpoints between fused

genes in the processed data; filtering said possible breakpoints, to retain candidate breakpoints, wherein each candidate breakpoint is in a 5' untranslated region (UTR) or a coding DNA sequence (CDS) of a functional gene region and each candidate breakpoint does not occur in an intron; and annotating the candidate breakpoints with at least one annotation useful in determining a relevance of a gene fusion for at least one of cancer diagnosis, cancer prognosis, or cancer treatment, wherein the gene fusion comprises the candidate breakpoint.

[0037] In one embodiment, a non-transitory computer readable program storage unit is provided, the non-transitory computer readable program storage unit encoded with instructions that, when executed by a computer, perform a method, comprising receiving mutation data from at least one cancer cell line; annotating the mutation data with at least one of variant classification, variant position, or variant change, to yield annotated mutation data; filtering the annotated mutation data, to yield gene region mutation data; classifying the gene region mutation data as hotspot, deleterious, or other; and nominating a gene comprising the gene region mutation as a gain of function, loss of function, or recurrent other gene, based on the relative frequency of mutations in the gene and the classifications of all gene region mutations in the gene.

[0038] In one embodiment, a method is provided, the method comprising detecting one or more gene fusions in a sample from a subject, to yield gene fusion detection data, wherein at least one of the gene fusions is selected from the gene fusions listed in Tables 1-3, 19, and 22, receiving by a computer system the gene fusion detection data, and identifying by the computer system at least one therapeutic option recommended for the subject, based on the gene fusion detection data.

[0039] In one embodiment, a system is provided, the system comprising a detector configured to (i) detect one or more gene fusions in a sample from a subject, to yield gene fusion detection data, wherein at least one of the gene fusions is selected from the gene fusions listed in Tables 1-3, 19, and 22 and (ii) transmit the gene fusion detection data; and a computer system configured to receive the gene fusion detection data and identify at least one therapeutic option recommended for the subject, based on the gene fusion detection data.

[0040] In another embodiment, a novel TP53 WT gene signature is provided as well as methods of detecting expression levels of one or more of the TP53 WT gene signature genes in Table 40.

DESCRIPTION OF THE DRAWINGS

[0041] Figure 1 provides a workflow for gene fusion RNASeq data processing.

[0042] Figure 2 shows the classification scheme for gene variants for Gain of Function and Loss of Function genes.

[0043] Figure 3 summarizes the data flow that integrates the various data types into a Genetic Event Database (GEDB).

[0044] Figure 4 is a flowchart showing the roll up of genetic events

[0045] Figure 5 is a graph showing the TP53 WT expression signature is significantly elevated in TP53 WT breast cancer compared to breast cancer samples harboring a TP53 point mutation.

[0046] Figure 6 is a graph showing the TP53 WT expression signature is significantly elevated in TP53 WT lung cancer compared to lung cancer samples harboring a TP53 mutation.

[0047] Figure 7 is a graph showing the TP53 WT expression signature is significantly elevated in TP53 WT ovarian cancer compared to ovarian cancer samples harboring a TP53 mutation.

[0048] Figure 8 A-D are graphs depicting is Raw RPKM expression values (A-B) vs. z-score normalized values for PLXNB21 and COL7A1 in Ovarian Serous Carcinoma patients (C-D). The population-wide dips in PLXNB1 expression at exons 12, 17 and 23 are smoothed out in the normalized data. A sample predicted to harbor a fusion between these genes, the red diamond indicates the caller-predicted breakpoint exon.

[0049] Figures 9 A-B is a table of frequent TP53 mutations by amino acid position. Mutations displayed that occur with overall frequency in patients of >0.25% in the pan-cancer analysis. A recurrent splice site mutation was identified at the intron-exon junction affecting T-125

[0050] Figures 10 A-B is a table of Tp53 in-frame insertion and deletion mutations. The maximum detected in-frame insertion-deletions identified was 21 bp. Greater than 99% of non-transposon indels across the genome are <100bp.

DETAILED DESCRIPTION

[0051] The disclosure provides novel gene fusions and variants, as well as novel associations of gene fusions and/or gene variants with certain types of cancers. Further provided are probes, reaction mixtures, assays and kits that relate to the gene fusions and/or variants disclosed herein.

Definitions

[0052] The term “marker” or “biomarker” refers to a molecule (typically protein, nucleic acid, carbohydrate, or lipid) that is expressed in the cell, expressed on the surface of a cancer cell or secreted by a cancer cell in comparison to a non-cancer cell, and which is useful for the diagnosis of cancer, for providing a prognosis, and for preferential targeting of a pharmacological agent to the cancer cell. Oftentimes, such markers are molecules that are overexpressed in a cancer cell in comparison to a non-cancer cell, for instance, 1-fold overexpression, 2-fold overexpression, 3-fold overexpression or more in comparison to a normal cell. Further, a marker can be a molecule that is inappropriately synthesized in the cancer cell, for instance, a molecule that contains deletions, additions or mutations in comparison to the molecule expressed on a normal cell. Alternatively, such biomarkers are molecules that are underexpressed in a cancer cell in comparison to a non-cancer cell, for instance, 1-fold underexpression, 2-fold underexpression, 3-fold underexpression, or more. Further, a marker can be a molecule that is inappropriately synthesized in cancer, for instance, a molecule that contains deletions, additions or mutations in comparison to the molecule expressed on a normal cell.

[0053] It will be understood by the skilled artisan that markers may be used in combination with other markers or tests for any of the uses, *e.g.*, prediction, diagnosis, or prognosis of cancer, disclosed herein.

[0054] “Biological sample” includes sections of tissues such as biopsy and autopsy samples, and frozen sections taken for histologic purposes. For example, the biological sample can include a Fresh-Frozen Paraffin-Embedded (FFPE) sample. Alternatively, a biological sample can include blood and blood fractions or products (*e.g.*, serum, plasma, platelets, red blood cells, and the like), sputum, bronchoalveolar lavage, cultured cells, *e.g.*, primary cultures, explants, and transformed cells, stool, urine, etc. A biological sample is typically obtained from a eukaryotic organism, most preferably a

mammal such as a primate e.g., chimpanzee or human; cow; dog; cat; a rodent, e.g., guinea pig, rat, Mouse; rabbit; or a bird; reptile; or fish.

[0055] A “biopsy” refers to the process of removing a tissue sample for diagnostic or prognostic evaluation, and to the tissue specimen itself. Any biopsy technique known in the art can be applied to the diagnostic and prognostic methods of the present invention. The biopsy technique applied will depend on the tissue type to be evaluated (e.g., lung *etc.*), the size and type of the tumor, among other factors. Representative biopsy techniques include, but are not limited to, excisional biopsy, incisional biopsy, needle biopsy, surgical biopsy, and bone marrow biopsy. An “excisional biopsy” refers to the removal of an entire tumor mass with a small margin of normal tissue surrounding it. An “incisional biopsy” refers to the removal of a wedge of tissue from within the tumor. A diagnosis or prognosis made by endoscopy or radiographic guidance can require a “core-needle biopsy”, or a “fine-needle aspiration biopsy” which generally obtains a suspension of cells from within a target tissue. Biopsy techniques are discussed, for example, in *Harrison’s Principles of Internal Medicine*, Kasper, *et al.*, eds., 16th ed., 2005, Chapter 70, and throughout Part V.

[0056] The terms “overexpress,” “overexpression,” or “overexpressed” interchangeably refer to a protein or nucleic acid (RNA) that is translated or transcribed at a detectably greater level, usually in a cancer cell, in comparison to a normal cell. The term includes overexpression due to transcription, post transcriptional processing, translation, post-translational processing, cellular localization (e.g., organelle, cytoplasm, nucleus, cell surface), and RNA and protein stability, as compared to a normal cell. Overexpression can be detected using conventional techniques for detecting mRNA (*i.e.*, RT-PCR, PCR, hybridization) or proteins (*i.e.*, ELISA, immunohistochemical techniques). Overexpression can be 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% or more in comparison to a normal cell. In certain instances, overexpression is 1-fold, 2-fold, 3-fold, 4-fold or more higher levels of transcription or translation in comparison to a normal cell.

[0057] The terms “underexpress,” “underexpression,” or “underexpressed” or “downregulated” interchangeably refer to a protein or nucleic acid that is translated or transcribed at a detectably lower level in a cancer cell, in comparison to a normal cell. The term includes underexpression due to transcription, post transcriptional processing, translation, post-translational processing, cellular localization (e.g., organelle, cytoplasm, nucleus, cell surface), and RNA and protein stability, as compared to a

control. Underexpression can be detected using conventional techniques for detecting mRNA (*i.e.*, RT-PCR, PCR, hybridization) or proteins (*i.e.*, ELISA, immunohistochemical techniques). Underexpression can be 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% or less in comparison to a control. In certain instances, underexpression is 1-fold, 2-fold, 3-fold, 4-fold or more lower levels of transcription or translation in comparison to a control.

[0058] The term “differentially expressed” or “differentially regulated” refers generally to a protein or nucleic acid that is overexpressed (upregulated) or underexpressed (downregulated) in one sample compared to at least one other sample, generally in a cancer patient compared to a sample of non-cancerous tissue in the context of the present invention.

[0059] The term “system” sets forth a set of components, real or abstract, comprising a whole where each component interacts with or is related to at least one other component within the whole.

[0060] The terms “polypeptide,” “peptide” and “protein” are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to amino acid polymers in which one or more amino acid residue is an artificial chemical mimetic of a corresponding naturally occurring amino acid, as well as to naturally occurring amino acid polymers and non-naturally occurring amino acid polymer.

[0061] The term “amino acid” refers to naturally occurring and synthetic amino acids, as well as amino acid analogs and amino acid mimetics that function in a manner similar to the naturally occurring amino acids. Naturally occurring amino acids are those encoded by the genetic code, as well as those amino acids that are later modified, *e.g.*, hydroxyproline, γ -carboxyglutamate, and O-phosphoserine. Amino acid analogs refers to compounds that have the same basic chemical structure as a naturally occurring amino acid, *i.e.*, a carbon that is bound to a hydrogen, a carboxyl group, an amino group, and an R group, *e.g.*, homoserine, norleucine, methionine sulfoxide, methionine methyl sulfonium. Such analogs have modified R groups (*e.g.*, norleucine) or modified peptide backbones, but retain the same basic chemical structure as a naturally occurring amino acid. Amino acid mimetics refer to chemical compounds that have a structure that is different from the general chemical structure of an amino acid, but that functions in a manner similar to a naturally occurring amino acid.

[0062] Amino acids may be referred to herein by either their commonly known three letter symbols or by the one-letter symbols recommended by the IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single-letter codes.

[0063] As to amino acid sequences, one of skill will recognize that individual substitutions, deletions or additions to a nucleic acid, peptide, polypeptide, or protein sequence which alters, adds or deletes a single amino acid or a small percentage of amino acids in the encoded sequence is a "conservatively modified variant" where the alteration results in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. Such conservatively modified variants are in addition to and do not exclude polymorphic variants, interspecies homologs, and alleles of the invention.

[0064] The following eight groups each contain amino acids that are conservative substitutions for one another: 1) Alanine (A), Glycine (G); 2) Aspartic acid (D), Glutamic acid (E); 3) Asparagine (N), Glutamine (Q); 4) Arginine (R), Lysine (K); 5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V); 6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W); 7) Serine (S), Threonine (T); and 8) Cysteine (C), Methionine (M). See, e.g., Creighton, *Proteins* (1984).

[0065] The phrase "specifically (or selectively) binds" when referring to a protein, nucleic acid, antibody, or small molecule compound refers to a binding reaction that is determinative of the presence of the protein or nucleic acid, such as the differentially expressed genes of the present invention, often in a heterogeneous population of proteins or nucleic acids and other biologics. In the case of antibodies, under designated immunoassay conditions, a specified antibody may bind to a particular protein at least two times the background and more typically more than 10 to 100 times background. Specific binding to an antibody under such conditions requires an antibody that is selected for its specificity for a particular protein. For example, polyclonal antibodies can be selected to obtain only those polyclonal antibodies that are specifically immunoreactive with the selected antigen and not with other proteins. This selection may be achieved by subtracting out antibodies that cross-react with other molecules. A variety of immunoassay formats may be used to select antibodies specifically immunoreactive with a particular protein. For example, solid-phase ELISA immunoassays are routinely used to select antibodies specifically immunoreactive with a protein (see, e.g., Harlow & Lane, *Antibodies, A Laboratory Manual* (1988) for a

description of immunoassay formats and conditions that can be used to determine specific immunoreactivity).

[0066] The phrase “functional effects” in the context of assays for testing compounds that modulate a marker protein includes the determination of a parameter that is indirectly or directly under the influence of a biomarker of the invention, e.g., a chemical or phenotypic. A functional effect therefore includes ligand binding activity, transcriptional activation or repression, the ability of cells to proliferate, the ability to migrate, among others. “Functional effects” include *in vitro*, *in vivo*, and *ex vivo* activities.

[0067] By “determining the functional effect” is meant assaying for a compound that increases or decreases a parameter that is indirectly or directly under the influence of a biomarker of the invention, e.g., measuring physical and chemical or phenotypic effects. Such functional effects can be measured by any means known to those skilled in the art, e.g., changes in spectroscopic characteristics (e.g., fluorescence, absorbance, refractive index); hydrodynamic (e.g., shape), chromatographic; or solubility properties for the protein; ligand binding assays, e.g., binding to antibodies; measuring inducible markers or transcriptional activation of the marker; measuring changes in enzymatic activity; the ability to increase or decrease cellular proliferation, apoptosis, cell cycle arrest, measuring changes in cell surface markers. The functional effects can be evaluated by many means known to those skilled in the art, e.g., microscopy for quantitative or qualitative measures of alterations in morphological features, measurement of changes in RNA or protein levels for other genes expressed in placental tissue, measurement of RNA stability, identification of downstream or reporter gene expression (CAT, luciferase, β -gal, GFP and the like), e.g., via chemiluminescence, fluorescence, colorimetric reactions, antibody binding, inducible markers, etc.

[0068] “Inhibitors,” “activators,” and “modulators” of the markers are used to refer to activating, inhibitory, or modulating molecules identified using *in vitro* and *in vivo* assays of cancer biomarkers. Inhibitors are compounds that, e.g., bind to, partially or totally block activity, decrease, prevent, delay activation, inactivate, desensitize, or down regulate the activity or expression of cancer biomarkers. “Activators” are compounds that increase, open, activate, facilitate, enhance activation, sensitize, agonize, or up regulate activity of cancer biomarkers, e.g., agonists. Inhibitors, activators, or modulators also include genetically modified versions of cancer biomarkers, e.g.,

versions with altered activity, as well as naturally occurring and synthetic ligands, antagonists, agonists, antibodies, peptides, cyclic peptides, nucleic acids, antisense molecules, ribozymes, RNAi and siRNA molecules, small organic molecules and the like. Such assays for inhibitors and activators include, e.g., expressing cancer biomarkers *in vitro*, in cells, or cell extracts, applying putative modulator compounds, and then determining the functional effects on activity, as described above.

[0069] A "probe" or "probes" refers to a polynucleotide that is at least eight (8) nucleotides in length and which forms a hybrid structure with a target sequence, due to complementarity of at least one sequence in the probe with a sequence in the target region. The polynucleotide can be composed of DNA and/or RNA. Probes in certain embodiments, are detectably labeled, as discussed in more detail herein. Probes can vary significantly in size. Generally, probes are, for example, at least 8 to 15 nucleotides in length. Other probes are, for example, at least 20, 30 or 40 nucleotides long. Still other probes are somewhat longer, being at least, for example, 50, 60, 70, 80, 90 nucleotides long. Yet other probes are longer still, and are at least, for example, 100, 150, 200 or more nucleotides long. Probes can be of any specific length that falls within the foregoing ranges as well. Preferably, the probe does not contain a sequence complementary to the sequence(s) used to prime for a target sequence during the polymerase chain reaction.

[0070] The terms "complementary" or "complementarity" are used in reference to polynucleotides (that is, a sequence of nucleotides) related by the base-pairing rules. For example, the sequence "A-G-T," is complementary to the sequence "T-C-A." Complementarity may be "partial," in which only some of the nucleic acids' bases are matched according to the base pairing rules. Alternatively, there may be "complete" or "total" complementarity between the nucleic acids. The degree of complementarity between nucleic acid strands has significant effects on the efficiency and strength of hybridization between nucleic acid strands.

[0071] "Oligonucleotide" or "polynucleotide" refers to a polymeric form of nucleotides of any length, either deoxyribonucleotide or ribonucleotide. These terms include, but are not limited to, a single-, double- or triple-stranded DNA, genomic DNA, cDNA, RNA, DNA-RNA hybrid, or a polymer comprising purine and pyrimidine bases or other natural chemically, biochemically modified non-natural or derivatized nucleotide bases.

[0072] “Amplification detection assay” refers to a primer pair and matched probe wherein the primer pair flanks a region of a target nucleic acid, typically a target gene, that defines an amplicon, and wherein the probe binds to the amplicon.

[0073] The terms “genetic variant” and “nucleotide variant” are used herein interchangeably to refer to changes or alterations to the reference human gene or cDNA sequence at a particular locus, including, but not limited to, nucleotide base deletions, insertions, inversions, and substitutions in the coding and noncoding regions. Deletions may be of a single nucleotide base, a portion or a region of the nucleotide sequence of the gene, or of the entire gene sequence. Insertions may be of one or more nucleotide bases. The “genetic variant” or “nucleotide variant” may occur in transcriptional regulatory regions, untranslated regions of mRNA, exons, introns, or exon/intron junctions. The “genetic variant” or “nucleotide variant” may or may not result in stop codons, frame shifts, deletions of amino acids, altered gene transcript splice forms or altered amino acid sequence.

[0074] The term “gene” refers to a polynucleotide (e.g., a DNA segment), that encodes a polypeptide and includes regions preceding and following the coding regions as well as intervening sequences (introns) between individual coding segments (exons). Parent genes or protein sequences are presented as Entrez Gene IDs or accession numbers. For example, the ZNF479 Entrez Gene ID is 90827. If any changes have been made to the sequence in the Gene ID in Entrez, the change is indicated after the Gene ID with a decimal and the number of the change (e.g., 90827.1). Further, for example, TPM1 has the accession number NM_004304.

[0075] The term “allele” or “gene allele” is used herein to refer generally to a naturally occurring gene having a reference sequence or a gene containing a specific nucleotide variant.

[0076] As used herein, “haplotype” is a combination of genetic (nucleotide) variants in a region of an mRNA or a genomic DNA on a chromosome found in an individual. Thus, a haplotype includes a number of genetically linked polymorphic variants which are typically inherited together as a unit.

[0077] As used herein, the term “amino acid variant” is used to refer to an amino acid change to a reference human protein sequence resulting from “genetic variant” or “nucleotide variant” to the reference human gene encoding the reference protein. The term “amino acid variant” is intended to encompass not only single amino acid

substitutions, but also amino acid deletions, insertions, and other significant changes or amino acid sequence in the reference protein. Variants of the invention are described by the following nomenclature: [original amino acid residue/position/substituted amino acid residue]. For example, the substitution of leucine for arginine at position 76 is represented as R76L.

[0078] The term "genotype" as used herein means the nucleotide characters at a particular nucleotide variant marker (or locus) in either one allele or both alleles of a gene (or a particular chromosome region). With respect to a particular nucleotide position of a gene of interest, the nucleotide(s) at that locus or equivalent thereof in one or both alleles form the genotype of the gene at that locus. A genotype can be homozygous or heterozygous. Accordingly, "genotyping" means determining the genotype, that is, the nucleotide(s) at a particular gene locus. Genotyping can also be done by determining the amino acid variant at a particular position of a protein which can be used to deduce the corresponding nucleotide variant (s).

[0079] A set of probes typically refers to a set of primers, usually primer pairs, and/or detectably-labeled probes that are used to detect the target genetic variations. The primer pairs are used in an amplification reaction to define an amplicon that spans a region for a target genetic variation for each of the aforementioned genes. The set of amplicons are detected by a set of matched probes. In an exemplary embodiment, the invention is a set of TaqMan™ (Roche Molecular Systems, Pleasanton, CA) assays that are used to detect a set of target genetic variations used in the methods of the invention.

[0080] In one embodiment, the set of probes are a set of primers used to generate amplicons that are detected by a nucleic acid sequencing reaction, such as a next generation sequencing reaction. In these embodiments, for example, AmpliSEQ™ (Life Technologies/Ion Torrent, Carlsbad, CA) or TruSEQ™ (Illumina, San Diego, CA) technology can be employed. In other embodiments, the two or more probes are primer pairs.

[0081] A modified ribonucleotide or deoxyribonucleotide refers to a molecule that can be used in place of naturally occurring bases in nucleic acid and includes, but is not limited to, modified purines and pyrimidines, minor bases, convertible nucleosides, structural analogs of purines and pyrimidines, labeled, derivatized and modified nucleosides and nucleotides, conjugated nucleosides and nucleotides, sequence modifiers, terminus modifiers, spacer modifiers, and nucleotides with backbone

modifications, including, but not limited to, ribose- modified nucleotides, phosphoramidates, phosphorothioates, phosphonamidites, methyl phosphonates, methyl phosphoramidites, methyl phosphonamidites, 5'- β - cyanoethyl phosphoramidites, methylenephosphonates, phosphorodithioates, peptide nucleic acids, achiral and neutral internucleotidic linkages.

[0082] "Hybridize" or "hybridization" refers to the binding between nucleic acids. The conditions for hybridization can be varied according to the sequence homology of the nucleic acids to be bound. Thus, if the sequence homology between the subject nucleic acids is high, stringent conditions are used. If the sequence homology is low, mild conditions are used. When the hybridization conditions are stringent, the hybridization specificity increases, and this increase of the hybridization specificity leads to a decrease in the yield of non-specific hybridization products. However, under mild hybridization conditions, the hybridization specificity decreases, and this decrease in the hybridization specificity leads to an increase in the yield of non-specific hybridization products.

[0083] "Stringent conditions" refers to conditions under which a probe will hybridize to its target subsequence, typically in a complex mixture of nucleic acids, but to no other sequences. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, *Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Probes*, "Overview of principles of hybridization and the strategy of nucleic acid assays" (1993). Generally, stringent conditions are selected to be about 5-10°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength pH. The T_m is the temperature (under defined ionic strength, pH, and nucleic concentration) at which 50% of the probes complementary to the target hybridize to the target sequence at equilibrium (as the target sequences are present in excess, at T_m , 50% of the probes are occupied at equilibrium). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide. For selective or specific hybridization, a positive signal is at least two times background, preferably 10 times background hybridization. Exemplary stringent hybridization conditions can be as following: 50% formamide, 5x SSC, and 1% SDS, incubating at 42°C, or, 5x SSC, 1% SDS, incubating at 65°C, with wash in 0.2x SSC, and 0.1% SDS at 65°C.

[0084] Nucleic acids that do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, for example, when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code. In such cases, the nucleic acids typically hybridize under moderately stringent hybridization conditions. Exemplary "moderately stringent hybridization conditions" include a hybridization in a buffer of 40% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 1X SSC at 45°C. A positive hybridization is at least twice background. Those of ordinary skill will readily recognize that alternative hybridization and wash conditions can be utilized to provide conditions of similar stringency. Additional guidelines for determining hybridization parameters are provided in numerous reference, e.g., and *Current Protocols in Molecular Biology*, ed.

[0085] Hybridization between nucleic acids can occur between a DNA molecule and a DNA molecule, hybridization between a DNA molecule and a RNA molecule, and hybridization between a RNA molecule and a RNA molecule.

[0086] A "mutein" or "variant" refers to a polynucleotide or polypeptide that differs relative to a wild-type or the most prevalent form in a population of individuals by the exchange, deletion, or insertion of one or more nucleotides or amino acids, respectively. The number of nucleotides or amino acids exchanged, deleted, or inserted can be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more such as 25, 30, 35, 40, 45 or 50. The term mutein can also encompass a translocation, for example the fusion of the polypeptides encoded by the ALK and TPM1 genes (TPM1/ALK).

[0087] "Gene fusion" refers to a chimeric genomic DNA resulting from the fusion of at least a portion of a first gene to a portion of a second gene. The point of transition between the sequence from the first gene in the fusion to the sequence from the second gene in the fusion is referred to as the "breakpoint" or "fusion point."

[0088] Transcription of the gene fusion results in a chimeric mRNA.

[0089] "Single nucleotide polymorphism" or "SNP" refers to a DNA sequence variation that occurs when a single nucleotide (A, T, G, or C) in the genome differs between members of a biological species or paired chromosomes in a human.

[0090] "Mutation" is defined herein as a specific change at a genomic location, i.e.: Chromosome, start, stop, reference base, alternate base, variant type (SNP, INS, DEL) etc.

[0091] "Annotation" is defined herein as a transcript-specific set of properties that describe the effect of the mutation, i.e.: Gene, transcript, variant classification, variant change, variant codon position, etc.

[0092] A "primer" or "primer sequence" refers to an oligonucleotide that hybridizes to a target nucleic acid sequence (for example, a DNA template to be amplified) to prime a nucleic acid synthesis reaction. The primer may be a DNA oligonucleotide, a RNA oligonucleotide, or a chimeric sequence. The primer may contain natural, synthetic, or modified nucleotides. Both the upper and lower limits of the length of the primer are empirically determined. The lower limit on primer length is the minimum length that is required to form a stable duplex upon hybridization with the target nucleic acid under nucleic acid amplification reaction conditions. Very short primers (usually less than 3-4 nucleotides long) do not form thermodynamically stable duplexes with target nucleic acids under such hybridization conditions. The upper limit is often determined by the possibility of having a duplex formation in a region other than the pre-determined nucleic acid sequence in the target nucleic acid. Generally, suitable primer lengths are in the range of about 10 to about 40 nucleotides long. In certain embodiments, for example, a primer can be 10-40, 15-30, or 10-20 nucleotides long. A primer is capable of acting as a point of initiation of synthesis on a polynucleotide sequence when placed under appropriate conditions.

[0093] The primer will be completely or substantially complementary to a region of the target polynucleotide sequence to be copied. Therefore, under conditions conducive to hybridization, the primer will anneal to the complementary region of the target sequence. Upon addition of suitable reactants, including, but not limited to, a polymerase, nucleotide triphosphates, etc., the primer is extended by the polymerizing agent to form a copy of the target sequence. The primer may be single-stranded or alternatively may be partially double-stranded.

[0094] "Detection," "detectable" and grammatical equivalents thereof refers to ways of determining the presence and/or quantity and/or identity of a target nucleic acid sequence. In some embodiments, detection occurs amplifying the target nucleic acid sequence. In other embodiments, sequencing of the target nucleic acid can be characterized as "detecting" the target nucleic acid. A label attached to the probe can include any of a variety of different labels known in the art that can be detected by, for example, chemical or physical means. Labels that can be attached to probes may include, for example, fluorescent and luminescence materials.

[0095] "Amplifying," "amplification," and grammatical equivalents thereof refers to any method by which at least a part of a target nucleic acid sequence is reproduced in a template-dependent manner, including without limitation, a broad range of techniques for amplifying nucleic acid sequences, either linearly or exponentially. Exemplary means for performing an amplifying step include ligase chain reaction (LCR), ligase detection reaction (LDR), ligation followed by Q-replicase amplification, PCR, primer extension, strand displacement amplification (SDA), hyperbranched strand displacement amplification, multiple displacement amplification (MDA), nucleic acid strand-based amplification (NASBA), two-step multiplexed amplifications, rolling circle amplification (RCA), recombinase-polymerase amplification (RPA)(TwistDx, Cambridge, UK), and self-sustained sequence replication (3SR), including multiplex versions or combinations thereof, for example but not limited to, OLA/PCR, PCR/OLA, LDR/PCR, PCR/PCR/LDR, PCR/LDR, LCR/PCR, PCR/LCR (also known as combined chain reaction-CCR), and the like. Descriptions of such techniques can be found in, among other places, Sambrook et al. *Molecular Cloning*, 3rd Edition; Ausbel et al.; *PCR Primer: A Laboratory Manual*, Diffeenbach, Ed., Cold Spring Harbor Press (1995); *The Electronic Protocol Book*, Chang Bioscience (2002), Msuih et al., *J. Clin. Micro.* 34:501-07 (1996); *The Nucleic Acid Protocols Handbook*, R. Rapley, ed., Humana Press, Totowa, N.J. (2002).

[0096] Analysis of nucleic acid markers can be performed using techniques known in the art including, without limitation, sequence analysis, and electrophoretic analysis. Non-limiting examples of sequence analysis include Maxam-Gilbert sequencing, Sanger sequencing, capillary array DNA sequencing, thermal cycle sequencing (Sears et al., *Biotechniques*, 13:626-633 (1992)), solid-phase sequencing (Zimmerman et al., *Methods Mol. Cell Biol.*, 3:39-42 (1992)), sequencing with mass spectrometry such as matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF/MS; Fu et al., *Nat. Biotechnol.*, 16:381-384 (1998)), and sequencing by hybridization. Chee et al., *Science*, 274:610-614 (1996); Drmanac et al., *Science*, 260:1649-1652 (1993); Drmanac et al., *Nat. Biotechnol.*, 16:54-58 (1998). Non-limiting examples of electrophoretic analysis include slab gel electrophoresis such as agarose or polyacrylamide gel electrophoresis, capillary electrophoresis, and denaturing gradient gel electrophoresis. Additionally, next generation sequencing methods can be performed using commercially available kits and instruments from companies such as the Life Technologies/Ion Torrent PGM or Proton, the Illumina HiSEQ or MiSEQ, and the Roche/454 next generation sequencing system.

[0097] In some embodiments, the amount of probe that gives a fluorescent signal in response to an excited light typically relates to the amount of nucleic acid produced in the amplification reaction. Thus, in some embodiments, the amount of fluorescent signal is related to the amount of product created in the amplification reaction. In such embodiments, one can therefore measure the amount of amplification product by measuring the intensity of the fluorescent signal from the fluorescent indicator.

[0098] "Detectably labeled probe" or "detector probe" refers to a molecule used in an amplification reaction, typically for quantitative or real-time PCR analysis, as well as end-point analysis. Such detector probes can be used to monitor the amplification of the target nucleic acid sequence. In some embodiments, detector probes present in an amplification reaction are suitable for monitoring the amount of amplicon(s) produced as a function of time. Such detector probes include, but are not limited to, the 5'-exonuclease assay (TAQMAN® probes described herein (see also U.S. Pat. No. 5,538,848) various stem-loop molecular beacons (see for example, U.S. Pat. Nos. 6,103,476 and 5,925,517 and Tyagi and Kramer, 1996, *Nature Biotechnology* 14:303-308), stemless or linear beacons (see, e.g., WO 99/21881), PNA Molecular Beacons™ (see, e.g., U.S. Pat. Nos. 6,355,421 and 6,593,091), linear PNA beacons (see, for example, Kubista et al., 2001, *SPIE* 4264:53-58), non-FRET probes (see, for example, U.S. Pat. No. 6,150,097), Sunrise®/Amplifluor™ probes (U.S. Pat. No. 6,548,250), stem-loop and duplex Scorpion probes (Solinas et al., 2001, *Nucleic Acids Research* 29:E96 and U.S. Pat. No. 6,589,743), bulge loop probes (U.S. Pat. No. 6,590,091), pseudo knot probes (U.S. Pat. No. 6,589,250), cyclicons (U.S. Pat. No. 6,383,752), MGB Eclipse™ probe (Epoch Biosciences), hairpin probes (U.S. Pat. No. 6,596,490), peptide nucleic acid (PNA) light-up probes, self-assembled nanoparticle probes, and ferrocene-modified probes described, for example, in U.S. Pat. No. 6,485,901; Mhlanga et al., 2001, *Methods* 25:463-471; Whitcombe et al., 1999, *Nature Biotechnology* 17:804-807; Isacsson et al., 2000, *Molecular Cell Probes* 14:321-328; Svanvik et al., 2000, *Anal Biochem* 281:26-35; Wolffs et al., 2001, *Biotechniques* 766:769-771; Tsourkas et al., 2002, *Nucleic Acids Research* 30:4208-4215; Riccelli et al., 2002, *Nucleic Acids Research* 30:4088-4093; Zhang et al., 2002 *Shanghai* 34:329-332; Maxwell et al., 2002, *J. Am. Chem. Soc.* 124:9606-9612; Broude et al., 2002, *Trends Biotechnol.* 20:249-56; Huang et al., 2002, *Chem. Res. Toxicol.* 15:118-126; and Yu et al., 2001, *J. Am. Chem. Soc.* 123:11155-11161.

[0099] Detector probes can also include quenchers, including without limitation black hole quenchers (Biosearch), Iowa Black (IDT), QSY quencher (Molecular Probes), and Dabsyl and Dabcel sulfonate/carboxylate Quenchers (Epoch).

[00100] Detector probes can also include two probes, wherein for example a fluor is on one probe, and a quencher is on the other probe, wherein hybridization of the two probes together on a target quenches the signal, or wherein hybridization on the target alters the signal signature via a change in fluorescence. Detector probes can also comprise sulfonate derivatives of fluorescein dyes with SO_3 instead of the carboxylate group, phosphoramidite forms of fluorescein, phosphoramidite forms of CY 5 (commercially available for example from Amersham). In some embodiments, interchelating labels are used such as ethidium bromide, SYBR® Green I (Molecular Probes), and PicoGreen® (Molecular Probes), thereby allowing visualization in real-time, or end point, of an amplification product in the absence of a detector probe. In some embodiments, real-time visualization can comprise both an intercalating detector probe and a sequence-based detector probe can be employed. In some embodiments, the detector probe is at least partially quenched when not hybridized to a complementary sequence in the amplification reaction, and is at least partially unquenched when hybridized to a complementary sequence in the amplification reaction. In some embodiments, the detector probes of the present teachings have a T_m of 63-69° C., though it will be appreciated that guided by the present teachings routine experimentation can result in detector probes with other T_m s. In some embodiments, probes can further comprise various modifications such as a minor groove binder (see for example U.S. Pat. No. 6,486,308) to further provide desirable thermodynamic characteristics.

[00101] In some embodiments, detection can occur through any of a variety of mobility dependent analytical techniques based on differential rates of migration between different analyte species. Exemplary mobility-dependent analysis techniques include electrophoresis, chromatography, mass spectroscopy, sedimentation, for example, gradient centrifugation, field-flow fractionation, multi-stage extraction techniques, and the like. In some embodiments, mobility probes can be hybridized to amplification products, and the identity of the target nucleic acid sequence determined via a mobility dependent analysis technique of the eluted mobility probes, as described for example in Published P.C.T. Application WO04/46344 to Rosenblum et al., and WO01/92579 to Wenz et al. In some embodiments, detection can be achieved by

various microarrays and related software such as the Applied Biosystems Array System with the Applied Biosystems 1700 Chemiluminescent Microarray Analyzer and other commercially available array systems available from Affymetrix, Agilent, Illumina, and Amersham Biosciences, among others (see also Gerry et al., J. Mol. Biol. 292:251-62, 1999; De Bellis et al., Minerva Biotech 14:247-52, 2002; and Stears et al., Nat. Med. 9:14045, including supplements, 2003). It will also be appreciated that detection can comprise reporter groups that are incorporated into the reaction products, either as part of labeled primers or due to the incorporation of labeled dNTPs during an amplification, or attached to reaction products, for example but not limited to, via hybridization tag complements comprising reporter groups or via linker arms that are integral or attached to reaction products. Detection of unlabeled reaction products, for example using mass spectrometry, is also within the scope of the current teachings.

[00102] "Aberration" Means a genomic structural variation or alteration of DNA. Examples include: over- /under-expression; copy number amplification/deletion; mutation; gene fusion; etc.

[00103] "Driver Event" means a genomic aberration, representing a Gain of Function (GoF) mutation, a fusion, or copy number peak.

[00104] "Recurrent" means occurrence of an event in 3 or more tumor samples.

[00105] "Mitelman" means a database of Chromosome Aberrations and Gene Fusions in Cancer manually curated from literature. <http://goo.gl/PnXMT>

[00106] GENE FUSIONS

Table 1: Gene Fusions

Cancer Type	Gene A Symbol	Gene B Symbol	orientation	Druggable gene
Bladder Urothelial Carcinoma	ALK	TPM1	TPM1/ALK	ALK
Colon Adenocarcinoma	ALK	PRKAR1A	PRKAR1A/ALK	ALK
Cutaneous Melanoma	ALK	NCOA1	NCOA1/ALK	ALK
Ovarian Serous Cystadenocarcinoma	CASR	LPP	LPP/CASR	CASR
Glioblastoma	EGFR	MDM2	MDM2/EGFR	EGFR
Lower Grade Glioma	ELAVL3	FGFR3	FGFR3/ELAVL3	FGFR3
Acute Myeloid Leukemia	B2M	GNAS	B2M/GNAS	GNAS
Clear Cell Renal Cell Carcinoma	DOCK8	JAK2	DOCK8/JAK2	JAK2
Papillary Renal Cell Carcinoma	HNF1B	NOTCH1	HNF1B/NOTCH1	NOTCH1
Glioblastoma	NFASC	NTRK1	NFASC/NTRK1	NTRK1
Thyroid Gland Carcinoma	NTRK1	SSBP2	SSBP2/NTRK1	NTRK1
Thyroid Gland Carcinoma	NTRK1	SQSTM1	SQSTM1/NTRK1	NTRK1
Prostate Adenocarcinoma	PIK3CA	TBL1XR1	TBL1XR1/PIK3CA	PIK3CA
Thyroid Gland Carcinoma	AKAP13	RET	AKAP13/RET	RET
Thyroid Gland Carcinoma	FKBP15	RET	FKBP15/RET	RET
Thyroid Gland Carcinoma	RET	TBL1XR1	TBL1XR1/RET	RET

Cancer Type	Gene A Symbol	Gene B Symbol	orientation	Druggable gene
Glioblastoma	CEP85L	ROS1	CEP85L/ROS1	ROS1
Thyroid Gland Carcinoma	ALK	GTF2IRD1	GTF2IRD1/ALK	ALK
Ovarian Serous Cystadenocarcinoma	BRS3	HTATSF1	HTATSF1/BRS3	BRS3
Invasive Breast Carcinoma	CCDC132	CDH1	CDH1/CCDC132; CCDC132/CDH1	CDH1
Invasive Breast Carcinoma	ERBB2	SLC29A3	ERBB2/SLC29A3	ERBB2
Thyroid Gland Carcinoma	MET	TFG	MET/TFG; TFG/MET	MET
Ovarian Serous Cystadenocarcinoma	MNDA	NOTCH2	NOTCH2/MNDA	NOTCH2
Thyroid Gland Carcinoma	IRF2BP2	NTRK1	IRF2BP2/NTRK1	NTRK1
Ovarian Serous Cystadenocarcinoma	EIF2C2	PTK2	EIF2C2/PTK2	PTK2
Invasive Breast Carcinoma	HOXB3	RARA	RARA/HOXB3	RARA
Prostate Adenocarcinoma	ETV4	STAT3	STAT3/ETV4	STAT3
Invasive Breast Carcinoma	C17orf64	TOP1	TOP1/C17orf64	TOP1
Prostate Adenocarcinoma	KIAA0753	TP53	TP53/KIAA0753	TP53
Glioblastoma	GFAP	VIM	GFAP/VIM; VIM/GFAP	VIM
Thyroid Gland Carcinoma	LTK	UACA	UACA/LTK	LTK
Papillary Renal Cell Carcinoma	ALK	STRN	STRN/ALK	ALK
Thyroid Gland Carcinoma	ALK	STRN	STRN/ALK	ALK
Cutaneous Melanoma	BRAF	CDC27	CDC27/BRAF	BRAF
Thyroid Gland Carcinoma	BRAF	MACF1	MACF1/BRAF	BRAF
Thyroid Gland Carcinoma	BRAF	MKRN1	MKRN1/BRAF	BRAF
Cutaneous Melanoma	BRAF	TAX1BP1	TAX1BP1/BRAF	BRAF
Prostate Adenocarcinoma	BRAF	JHDM1D	JHDM1D/BRAF	BRAF

Table 2: Gene Fusions

Cancer Type	Gene A Symbol	Gene B Symbol	Orientation	Druggable gene
Cutaneous Melanoma	CLCN6	RAF1	CLCN6/RAF1	RAF1
Cutaneous Melanoma	TRAK1	RAF1	TRAK/RAF1	RAF1
Colon Adenocarcinoma	AKT1	PRKACA	PRKACA/AKT1	AKT1
Endometrial Endometrioid Adenocarcinoma	AKT1	PRKACA	PRKACA/AKT1	AKT1
Colon Adenocarcinoma	AKT2	PRKACA	PRKACA/AKT2	AKT2
Lung Adenocarcinoma	FYN	MLL	MLL/FYN	FYN
Lung Adenocarcinoma	ECHD1	FYN	ECHD1/FYN	FYN
Invasive Breast Carcinoma	JAK2	TTC13	TTC13/JAK2	JAK2
Gastric Adenocarcinoma	CAB39	ERBB2	CAB39/ERBB2	ERBB2
Endometrial Endometrioid Adenocarcinoma	BRAF	EXOC4	EXOC4/BRAF	BRAF
Invasive Breast Carcinoma	HOOK3	IKBKB	HOOK3/IKBKB	IKBKB
Invasive Breast Carcinoma	CDK6	KRIT1	KRIT1/CDK6	CDK6
Gastric Adenocarcinoma	CAPZA2	MET	CAPZA2/MET	MET
Invasive Breast Carcinoma	ACE	MLLT6	MLLT6/ACE	ACE
Endometrial Endometrioid Adenocarcinoma	HLA-C	MUC16	HLA-C/MUC16	MUC16
Head and Neck Squamous Cell Carcinoma	LYN	NTRK3	LYN/NTRK3	LYN, NTRK3
Ovarian Serous Cystadenocarcinoma	MUC16	OR7G2	MUC16/OR7G2	MUC16

Cancer Type	Gene A Symbol	Gene B Symbol	Orientation	Druggable gene
Ovarian Serous Cystadenocarcinoma	MDK	RAB11B	RAB11B/MDK	MDK
Squamous Cell Lung Carcinoma	GADD45GIP1	RB1	RB1/GADD45GIP1	RB1
Gastric Adenocarcinoma	PRKAR2A	RHOA	PRKAR2A/RHOA	RHOA
Cutaneous Melanoma	MAPK1	SHANK3	SHANK3/MAPK1	MAPK1
Thyroid Gland Carcinoma	RET	SPECC1L	SPECC1L/RET	RET
Ovarian Serous Cystadenocarcinoma	IGFBP2	SPP1	IGFBP2/SPP1	IGFBP2, SPP1
Invasive Breast Carcinoma	PAPD7	SRD5A1	PAPD7/SRD5A1; SRD5A1/ PAPD7	SRD5A1
Glioblastoma	RARA	TAOK1	TAOK1/RARA	RARA
Gastric Adenocarcinoma	CDK12	THRA	THRA/CDK12	THRA
Invasive Breast Carcinoma	NARS2	TOP1	NARS2/TOP1	TOP1
Gastric Adenocarcinoma	PTK2	TRAPPC9	PTK2/TRAPPC9; TRAPPC9/PTK2	PTK2
Invasive Breast Carcinoma	CBL	UBE4A	CBL/UBE4A	CBL
Lower Grade Glioma	GFAP	VIM	GFAP/VIM; VIM/GFAP	VIM
Invasive Breast Carcinoma	ADAM9	WRN	WRN/ADAM9	ADAM9
Colon and Rectal Adenocarcinoma	MAP2K2	YWHAE	YWHAE/MAP2K2	MAP2K2
Head and Neck Squamous Cell Carcinoma	ALK	CLIP4	CLIP4/ALK	ALK
Squamous Cell Lung Carcinoma	ALK	CLIP4	CLIP4/ALK	ALK
Thyroid Gland Carcinoma	ALK	MEMO1	MEMO1/ALK	ALK
Thyroid Gland Carcinoma	BRAF	SND1	BRAF/SND1; SND1/BRAF	BRAF
Thyroid Gland Carcinoma	BRAF	ZC3HAV1	ZC3HAV1/BRAF	BRAF

Table 3: Gene Fusions

Cancer Type	Gene A Symbol	Gene B Symbol	orientation	Druggable gene	Cancer type precedent
Thyroid Gland Carcinoma	NOTCH1	SEC16A	SEC16A-NOTCH1	NOTCH1	breast cancer
Invasive Breast Carcinoma	ERC1	RET	ERC1-RET	RET	thyroid cancer
Ovarian Serous Cystadenocarcinoma	CCDC170	ESR1	ESR1/CCDC170	ESR1	Invasive Breast Carcinoma
Head and Neck Squamous Cell Carcinoma	RPS6KB1	VMP1	RPS6KB1/VMP1; VMP1/RPS6KB1	RPS6KB1	Invasive Breast Carcinoma
Lung Adenocarcinoma	RPS6KB1	VMP1	RPS6KB1/VMP1	RPS6KB1	Invasive Breast Carcinoma
Squamous Cell Lung Carcinoma	RPS6KB1	VMP1	RPS6KB1/VMP1	RPS6KB1	Invasive Breast Carcinoma
Ovarian Serous Cystadenocarcinoma	RPS6KB1	VMP1	RPS6KB1/VMP1	RPS6KB1	Invasive Breast Carcinoma
Cutaneous Melanoma	RPS6KB1	VMP1	RPS6KB1/VMP1	RPS6KB1	Invasive Breast Carcinoma
Gastric Adenocarcinoma	RPS6KB1	VMP1	RPS6KB1/VMP1	RPS6KB1	Invasive Breast Carcinoma

Table 4: Breakpoint Sequence for Table 1

Table 4 Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chrom	3' Gene Symbol	3' Accession	3' Gene Breakpoint	Breakpoint Sequence	SEQ ID NO
TPM1/ALK	chr15	TPM1	NM_000366	63,354,844	chr2	ALK	NM_004304	29446394	TGCGGAGAGGTAGTAACATAATTCGAGAAAAAGCATTCATGACTTAGAAAGTGTACCGCCGGAAG CACCAGGAGCTGCAAGCCATGCGAGATGGAGCTGCAG	1
PRKARI/ALK	chr17	PRKARIA	NM_002734	66,511,717	chr2	ALK	NM_004304	29446263	CTGAGAGACCCATGGCAATTCCTCAGGGAAATACCTTGAGAGGTGGAGAAAGACCTCTCTCCATCAGT GACCTGAAGGAGGTGCGCGGAAAAACATACCCCT	2
NCOA1/ALK	chr2	NCOA1	NM_003743	24,991,142	chr2	ALK	NM_004304	30143047	GTGCAACAGGTTACAGGTGTTTCTGACGTCAGTGTAACAGTGAATCTGGTAGCGGCTGTGGGGC TGCTCCAGTTCAATCTCAGCGAGCTGTTCAGTTGG	3
LPP/CASR	chr3	LPP	NM_005578	188,202,492	chr3	CASR	NM_000388	121972795	GAAACTTTCCTCTCCACCACTCTTGATGAGAGGCTTTCAAAGTACAGIAGGCAATCACAGGAG GCCCTGCAATGATGTGGCTTCCAAAGACTCAAGCA	4
MDM2/EGFR	chr12	MDM2	NM_002392	69,203,072	chr7	EGFR	NM_005228	55231426	GATGGTCTGTAAACCACTCACAGATTCAGCTTCGGAACAAGAGACCCTGTGAGCCAAGGAG TTTGTGGAGAACTCTGAGTGCATACAGTGCACCCCA	5
FGFR3/ELAVL3	chr4	FGFR3	NM_000142	1,808,638	chr19	ELAVL3	NM_001420	11577572	GCCCTCCAGAGGCCCACTTCACAGCTGTGGAGGACCTGGACCGTGTCTTGGTACAAATG GAGCCACTGAGCAGCAGGACAGCAACCTCATCTGTC	6
B2M/GNAS	chr15	B2M	NM_004048	45,003,811	chr20	GNAS	NM_000516	57470667	TAGTGTGCTCGGCTACTCTCTCTTCTGGCTGGAGGCTATCCAGCGTGTCTGGAGAACTCTGGTA AAAGCACCATTGTGAAGCAGATGAGGATCTCTGCA	7
DOCK8/JAK2	chr9	DOCK8	NM_203447	340,321	chr9	JAK2	NM_004972	5050686	GAGATTTTGGAAATTTCCACACGAGAAATATATGTCCCTCACACTGTGTATACGTGGCGGCATGAT TTTGTGACGGATGGATAAAAGTACTGTGACTCA	8
HNFB/NOTCH1	chr17	HNFB	NM_000458	36,099,431	chr9	NOTCH1	NM_017617	139396940	TGCGCTCTGTACACTGTTAGTCAAGAAACGAGAGATCTCTCCGACGTGAGACCGTGGAG CGCCCGCCCGCGGCGGAGCTGCACTTCATGTACGTG	9
NFASC/NTRK1	chr1	NFASC	NM_015090	204,951,148	chr1	NTRK1	NM_002529	156844363	GGGAAAGGCGCTGAGCCAGAGTCCGTCTATCGGTACTCCGGGAGAGATTACACTAACAGCACAT CTGGAGACCCCGTGGAGAGAGGACGAAACACCTT	10
SSBP2/NTRK1	chr5	SSBP2	NM_012446	80,742,687	chr1	NTRK1	NM_002529	156845312	TCCAGGAGGTGGAGGGCCACACGGAACACCAATCATGTGCTAGTCCAGCAGCGCCCGCTGTCTGTG GCTCCAGAGGATGGGCTGGCCATGTCTCTGCAATTC	11
SQSTM1/NTK1	chr5	SQSTM1	NM_003900	179,252,226	chr1	NTRK1	NM_002529	156844363	TTTCTGAGAGAACGTTGGGAGAGTGTGGACGCTGCGCTTAGCCCTCTGGIACACTAACAGCACAT CTGGAGACCCCGTGGAGAGAGGACGAAACACCTT	12
TBL1XR1/PIK3CA	chr3	TBL1XR1	NM_024665	176,914,909	chr3	PIK3CA	NM_006218	178916338	CATATAAACTACTTTAAGGAATTAGATGTATGGTTCTCCAAAGCAGAAACCTGGAAACCGGTG GCCTCCAAACCGCGCTCCCGCTCCCGGGAATGGAGG	13
AKAP13/RET	chr15	AKAP13	NM_006738	86,286,839	chr10	RET	NM_020630	43612067	CGCCATCTGCACCTTCCATAGCCAAATTCAGGGTCAATGGACTCAGAACTTGGTTCTTGGAAAAAC TCTAGGAGAGGGGAAATTTGGAAAAAGTGGTCAAGG	14
FKBP15/RET	chr9	FKBP15	NM_015258	115,932,802	chr10	RET	NM_020630	43612032	AATCTTACAAATGGCAGGACCATTTCTGGGAACCATATGAATACGATCAAGGAGGATCCAAAGTG GGAAATCCCTCGGAGAACTTGGTTCTTGGAAAAAC	15
TBL1XR1/RET	chr3	TBL1XR1	NM_024665	176,765,103	chr10	RET	NM_020630	43610136	GGCCTATATTTGCAATTAATAGGAATAAGAAAGGAAATTTTCATCTAAAGTGTCTGGACTCCATGGA GAACCAAGGTCTCCGTGCTGCTTCAAGATCTCTGG	16
CEP85L	chr6	CEP85L	387119	118,802,942	chr6	ROSI	NM_002911	117641193	TTAATATGCCAGAAAAAGAAAGAAAGAGGTAGTAACCTTACAGTACTCTTCCAAACCC TTAATATGCCAGAAAAAGAAAGAAAGAGGTAGTAACCTTACAGTACTCTTCCAAACCC	17

Table 4 Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chrom	3' Gene Symbol	3' Accession	3' Gene Breakpoint	Breakpoint Sequence	SEC ID NO
/ROS1							44		AAGAGGAGATTGAAATCTCTCCCTCCCTCGGG	
CCDC132/CDH1	CCDC132	NM_017667	chr7	92,940,584	CDH1	NM_004360	chr16	68,857,494	GAATGCACCTATTAAACAAATCAACATTTGAACGTATGTTGTCTTGGGGAATCTTGG GAGGAATCTCTGCTTGTCTAATCTGATCTGCTG	30
CDH1/CCDC132	CDH1	NM_004360	chr16	68,857,529	CCDC132	NM_017667	chr7	92,952,923	AACATCAAGAGCAATGGCTTAAGAAATGTCATCATCTGCATATATTTTCTTAGCAAGCAAGAA TTCTCCCAAGAAATCCCAAGATGGAGGAAATTTGC	31
CDH1/CCDC132	CDH1	NM_004360	chr16	68,857,529	CCDC132	NM_017667	chr7	92,952,923	GCAATCTCTGCCATCTCTGGGATCTCTGGAGAAATCTTCTGCTTGTCTAAGAAATATATGCAGA TGATGAACATCTCTAAGCAATTCGCTTTGATGTT	32
EIF2C2/PTK2	EIF2C2	NM_012154	chr8	141,645,584	PTK2	NM_005607	chr8	141,685,598	GCTGCAGGATCTGGTTTACCCACAGGCTGATATATATGTTGTTTCCAAATCCGGGCGGCTCCCG AGTACATGTTGGCGCGCGCGGCTCCGGGGCC	33
EIF2C2/PTK2	EIF2C2	NM_012154	chr8	141,645,584	PTK2	NM_005607	chr8	141,685,598	GGCCCGGAGCCCTCTGGCGGGCCCAACATGTACTCGGAGCCGGCCCGCAATTGGAAACCAACA TATATATACAGCTGTGGGTAAACGAGATGCTGAGC	34
EIF2C2/PTK2	EIF2C2	NM_012154	chr8	141,645,584	PTK2	NM_005607	chr8	141,712,806	CCCCGGAGCCCTCTGGCGGGCCCAACATGTACTCGGAGCCGGCCCGGTTTCTGGCTACCTGG TTCAATGGAATCACAGCAATGGCTGGGAGCATCT	35
EIF2C2/PTK2	EIF2C2	NM_012154	chr8	141,645,584	PTK2	NM_005607	chr8	141,762,415	CGAAGTACAGTTTATACATGTTTAAATTGCAACCGCCAAAGCTGGATTCTCCGGGCGGCTCCC GAGTACATGTTGGCGCGCGAGGGCTCCGGGGCC	36
EIF2C2/PTK2	EIF2C2	NM_012154	chr8	141,645,584	PTK2	NM_005607	chr8	141,675,096	GGCCCGGAGCCCTCTGGCGGGCCCAACATGTACTCGGAGCCGGCCCGCAAGTCCGCTTGGC CTTGAGGACATTAATTGGGCACTGTGGATGAGACCAAT	37
ERBB2/SLC29A3	ERBB2	NM_004448	chr17	37,883,211	SLC29A3	NM_018344	chr10	73,115,986	ACACATGGGCGGCAAGAACAGGCTCATGTACTCTGGCATCTCCAGCGCCGGGCGGAGGGT CTGGACAGAAAGAGCCCTCTGGGTTACAGATACT	38
ERBB2/SLC29A3	ERBB2	NM_004448	chr17	37,883,548	SLC29A3	NM_018344	chr10	73,121,774	GGCGCTGGGGGCAATGTTCCACACAGGACCCGAGCTCATCTACCAGGA/CTCCCTCACTGCCCC TTCCGTGGCTCCAGATTTGATTTCCACACAC	39
ERBB2/SLC29A3	ERBB2	NM_004448	chr17	37,883,598	SLC29A3	NM_018344	chr10	73,121,726	GTGGCGTGGGAGCTGACACTAGGCTGGAGCCCTCTGAAAGAGAGGCGCTGTCTTTCGCGCCC ATGTGTTTCTGGTGAAGAGGAGCTTCCCAAGGACT	40
ERBB2/SLC29A3	ERBB2	NM_004448	chr17	37,883,205	SLC29A3	NM_018344	chr10	73,115,911	CTGAGGAGTATCTGGTATCCCAAGAGGCTTCTTCTGTCCAGACCTGCCTGCCTGCGCTT CTTCTGACGGCCACTGTCTTCTCTGCTGCTGCA	41
ERBB2/SLC29A3	ERBB2	NM_004448	chr17	37,882,078	SLC29A3	NM_018344	chr10	73,115,911	TGCAGACACGAGGAAGACAGTGGCGCTCAGGAAGAGGCCAGGGCGCTGGGTGCAGATGGGG GGCTGGGCGAGCGCTCCCTCTTTCAGAGGAGGTGAG	42
GFAP/VIM	GFAP	NM_002055	chr17	42,987,987	VIM	NM_003380	chr10	17,277,377	AGGAGAACCGGATCACCATTCCCGTGCAGACCTTCTCCAACTGCAGATTTCGAGGAGAGCAGGA TTCTGCTCTTCTCCAACTTCTTCTGCTGAACT	43
GFAP/VIM	GFAP	NM_002055	chr17	42,988,732	VIM	NM_003380	chr10	17,277,285	ACGTGCGGAGGCGGCGGATTCAGAGGCGCTGGCGGGCTGGAGGAATAGGCTGCTCACTT TCGTGAATACCAAGAGCTGCTCAATGTTAAGTGGC	44
GFAP/VIM	GFAP	NM_002055	chr17	42,987,987	VIM	NM_003380	chr10	17,277,377	AGGAGAACCGGATCACCATTCCCGTGCAGACCTTCTCCAACTGCAGATTTCGAGGAGAGCAGGA TTTCTGCTCTTCTCCAACTTCTTCTGCTGAACT	45
GFAP/VIM	GFAP	NM_002055	chr17	42,988,622	VIM	NM_003380	chr10	17,277,371	AATGTCAAGCTGGCGCTGGACATCGAGATCGCCACCTACAGGAAGCTGCTTGGAGGCGGAGGA GCAGGATTTCTGCTCTTCTCCAACTTCTGCTGCTGCT	46
GFAP/VIM	GFAP	NM_002055	chr17	42,983,511	VIM	NM_003380	chr10	17,277,237	ATCACCATCCCGTGGAGACCTTCTCCAACTGCAGATTTCGAGGAACAGGACACTATTGGCGGC CTGCAGGATGAGATTCAGGAATATGAAGAGGAGAAAT	47

Table 4		Breakpoint Sequence										SEF ID NO
Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chrom	3' Gene Symbol	3' Accession	3' Gene Breakpoint				
GFAP/VIM	GFAP	NM_001131	chr17	42,987,602	VIM	NM_003380	chr10	17,277,286	CTTCTCAAACCTGCAGATTCCGAGGGGCAAAAGCAACCAAGACGGGAAATGGCTCTGTCACCTTCGTGAATACCAAGACCTGCTCAATGTTAAGATGCC		48	
GFAP/VIM	GFAP	NM_002055	chr17	42,987,983	VIM	NM_003380	chr10	17,278,298	GAACCGGATCACCATTCCCGGTGCAGACCTTCTCAAACCTGCAGATTTCGAGJAATCTGGATTCACTCCCTCTGTTGATACCCACTCAAAAAGGACACTTCT		49	
GFAP/VIM	GFAP	NM_002055	chr17	42,992,594	VIM	NM_003380	chr10	17,271,785	CAGAGATGATGGAGCTCAATACCCGCTTTGGCAGCTACATCGAGAAAGGTTTCGCTTCCTCTGGAGCAGCAGATAAGATCTCTGCTGGCAGCTCGACAGCT		50	
GFAP/VIM	GFAP	NM_002055	chr17	42,985,469	VIM	NM_003380	chr10	17,277,285	GAACACGCTGCAGACCAAGTCTGTGTCAGAAGCCACCTCAAGAGGAAMATGCTCGTCACTTCGTGAATACCAAGACCTGCTCAATGTTAAGATGCC		51	
GFAP/VIM	GFAP	NM_002055	chr17	42,988,779	VIM	NM_003380	chr10	17,277,168	CACGAACGAGTCCCTGGAGAGCAGATGCGGACGAGGAGGAGCGGCACGAATGAGTCCCTGGAAACCCAGATGCCGTGAAATGGAAAGAGAACTTTGCCGT		52	
GFAP/VIM	GFAP	NM_002055	chr17	42,988,637	VIM	NM_003380	chr10	17,277,351	GGCAGAGAAATCTCTGCTCTCTCGCCTTCCAGCAGCTTCTCTGATGGTGCGTGTCGATGTCGATGTCAGAGGCCAGCTTGACATITGAGCAGGTCGTGGTA		53	
GFAP/VIM	GFAP	NM_002055	chr17	42,992,627	VIM	NM_003380	chr10	17,271,752	CTGGCTTCAAGGAGACCCGGGCGCATGTAGCGGGGCAAGATGATGGAGCTCAATGACCCGCTTCGCCTCATCATCGACAAGTGGCTCTCTCTGGAGCAGCA		54	
GFAP/VIM	GFAP	NM_002055	chr17	42,988,742	VIM	NM_003380	chr10	17,277,351	GAGGACGGCACGTGCGGGAGGGCGGCAGTTATCAGGAGGGCCTGGCCGCGCCACCTACAGGAAGCTCTGGAAAGCGGAGGACGACGAGGATTTCTTGGC		55	
GFAP/VIM	GFAP	NM_002055	chr17	42,988,025	VIM	NM_003380	chr10	17,276,771	CATCGAGATCGCCACCTACAGGAAGCTGCTAGAGGGCGAGGAACCGAAGACAGGTGTCAGTCCCTACCTGTGAAGTGTGATGCCCTTAAGAGAAACCAAT		56	
GFAP/VIM	GFAP	NM_002055	chr17	42,988,742	VIM	NM_003380	chr10	17,277,367	GAGGAGCGCACGTGCGGGAGGGCGGCAGTTATCAGGAGGGCCTGGCCGCGTCTGGAAAGGCGGAGGAGCAGGATTTCTCTGCTCTTCCAAACTTTCC		57	
GFAP/VIM	GFAP	NM_002055	chr17	42,988,642	VIM	NM_003380	chr10	17,277,351	GGCAGAGAAATCTCTCTCTCTCGCCTTCCAGCAGCTTCTGTAGTGGTGGTGTCTCTGATGTCCAGGGCCAGCTTGACATTGACGAGGTCCTGTCTCTCT		58	
GFAP/VIM	GFAP	NM_002055	chr17	42,988,642	VIM	NM_003380	chr10	17,277,351	AGGAGTACCAAGCACTCTCAATGTCAAGCTGGCCCTGGACATCGAGATCGCCACCTACAGGAAGCTCTGGAAGGGCGAGGACAGGATTTCTCTGCC		59	
GFAP/VIM	GFAP	NM_002055	chr17	42,992,612	VIM	NM_003380	chr10	17,271,824	CCCGGGCCAGTACGCGGCAGAGATGATGAGCTCAATGACCGCTTTGCGCTCGAGCAGCTCAAAGGCCAAGCTCGCGCTTGGGGACCTCTACGA		60	
GFAP/VIM	GFAP	NM_002055	chr17	42,992,610	VIM	NM_003380	chr10	17,271,769	CGGGCAGTGAAGCGGCGAGAGATGATGAGCTCAATGACCGCTTTGGCAGTCTACATCGACAAGGTGGCTCTCTGGAGCAGCAGATAAAGACTCTGCTGG		61	
GTF2IRD1/A	GTF2IRD1	NM_005685	chr7	73,935,627	ALK	NM_004304	chr2	29,446,394	ACGTCCATGCTCCAAAGCGAGATTTCTTCTCCATCTGCTCCATGCAAGTCAAGTGTACCGCCGGGAAGCAGCAGGAGCTGCAAGCCATGCAAGATGGAGCTGCA		62	
HTATSF1/BR	HTATSF1	NM_014500	chrX	135,586,622	BR33	NM_001727	chrX	135,572,292	CCATGACGAGCTTGTCATCACTACAAAGATATGTTTTCATCTCTATGGATTTTGTGAGATACAAGGCAGTTGTGAAGGCCACTTGACGACAGCCCTCCCAATGCCAT		63	
IRF2BP2/NT	IRF2BP2	NM_182972	chr1	234,744,241	NTRK1	NM_002529	chr1	156,844,363	CTCGGGCCCTTCGAGAGCAAGTTTAAAGAGGAGCGCGCCCTGACTGCAJACACTAACAACGACA		64	
IRF2BP2/NT	IRF2BP2	NM_182972	chr1	234,744,241	NTRK1	NM_002529	chr1	156,844,363	TCTGGAGACCCGGTGGAGAGAAAGAGGAGAAACACTT		65	
MET/TFG	MET	NM_000245	chr7	116,412,043	TFG	NM_006070	chr3	100,455,420	CTGCTCTTAAACTTGCTCTCGAAGGGGCGCGGAG		66	

Table 4	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chrom	3' Gene Symbol	3' Accession	3' Gene Breakpoint	Breakpoint Sequence	Seq ID NO
	me								CTGCAGAAAGATCGTTCAGGAACACCCGACAGCAATT	
MET/TFG	MET	NM_000245	chr7	116,412,013	TFG	NM_006070	chr3	100,455,435	TGTAAGTGCCCGAAGTGTAAGGCCAACTACAGAAATGGTTTCAATGAAATCTGCAGAGAAGATCGTT	67
MET/TFG	MET	NM_000245	chr7	116,414,937	TFG	NM_006070	chr3	100,455,447	CAGGAACCCGGACAGCAATGCTTCTCTCTGCTCA	68
MET/TFG	MET	NM_000245	chr7	116,415,078	TFG	NM_006070	chr3	100,455,435	AATGGTTTCAATGAAATCTGTAGACTACCGAGCTACTTTTCCAGAAAGATGCTGTTCCAGGAACACCCG	69
NOTCH2/MNDA	NOTCH2	NM_024408	chr1	120,478,095	MNDA	NM_002432	chr1	158,815,377	ACAGCATGCTTCTCTCTCTCAGCAGCTCACCA	70
NOTCH2/MNDA	NOTCH2	NM_024408	chr1	120,478,095	MNDA	NM_002432	chr1	158,815,377	TATATCCAGTCCAATTACTGCAAAATACTGTGCCAATGTAGCCTCAGTGTCTGTCGACGAAGATCGTTTC	71
RARA/HOXB3	RARA	NM_000964	chr17	38,508,759	HOXB3	NM_002146	chr17	46,632,980	GCCCAACGGCAGGTGGATGCAAGAGAAATGTTC	72
STAT3/ETV4	STAT3	NM_003150	chr17	40,468,807	ETV4	NM_001986	chr17	41,611,353	GTATTGACCTTTGTGAACCAATTTCAAGTGTCTTGCCCAACGACGACCTGGGGAATGTTC	73
STAT3/ETV4	STAT3	NM_003150	chr17	40,468,860	ETV4	NM_001986	chr17	41,613,825	CCATCCGACACAGATCACCTCTCAAGGCTGCCTGCCCTGGACATCCCTGAGGGGAGATTGTTC	74
STAT3/ETV4	STAT3	NM_003150	chr17	40,468,846	ETV4	NM_001986	chr17	41,610,042	GCCTGCCCTCGCTCTGGGGCTCGAATGTAATA	75
TFG/MET	TFG	NM_006070	chr3	100,451,516	MET	NM_000245	chr7	116,414,935	GTTTGGAAATAATGGTGAAGGTGCTGAACCCCTCAGCAGGAGGGCAGTTTGTAGCTTTCCACAGCC	76
TOP1/C17orf64	TOP1	NM_003286	chr20	39,729,993	C17orf64	NM_181707	chr17	58,503,144	CCACCAACGGATCAAGAAAGGAGCCCAAGTCCC	77
TOP1/C17orf64	TOP1	NM_003286	chr20	39,729,993	C17orf64	NM_181707	chr17	58,503,144	AGCAATACCAATTGACCTGCGGATGTCGCCCGCACITTTAGATTCAATTGATTCAGATTGTCCTGTAT	78
TOP1/C17orf64	TOP1	NM_003286	chr20	39,728,797	C17orf64	NM_181707	chr17	58,503,167	TTCCATTCAAAAAACCTAGCTTTCCACAGGCCA	79
TP53/KIAA0753	TP53	NM_000546	chr17	7,590,695	KIAA0753	NM_014804	chr17	6,498,373	CGCTGGAGACAAGCTCCAGCGCACGCGCTGTAGA	80
TP53/KIAA0753	TP53	NM_000546	chr17	7,579,529	KIAA0753	NM_014804	chr17	6,493,323	TGGCATGGCGCATGAGGAGTCTCTACAGAGGCTGGCGCTGGAGCTTGTGTCCAGGAGGCTCTAT	81
TP53/KIAA0753	TP53	NM_000546	chr17	7,590,695	KIAA0753	NM_014804	chr17	6,498,373	CTTGAAGTTAGCAATCTCTCTTTGTGGTTATCCA	82
VIM/GFAP	VIM	NM_003380	chr10	17,277,325	GFAP	NM_002055	chr17	42,987,987	TCAGCATATCGGATTTTATATATCTTTGACGAACAGACTCTCTGGTATTTTCCATCCACGGGAAGCG	83
VIM/GFAP	VIM	NM_003380	chr10	17,277,325	GFAP	NM_002055	chr17	42,988,666	TGTCCAGCTCTGGAAAGCAGCGCTCCAGCCGA	84

Table 4 Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chrom	3' Gene Symbol	3' Accession	3' Gene Breakpoint	Breakpoint Sequence	SEQ ID NO
VIM/GFAP	VIM	NM_003380	chr10	17,277,255	GFAP	NM_002055	chr17	42,987,987	TTGAAGCTCTACTCAACAGACACTATTGGCCGCTGCTCGAGAGAAACACGCTGGACACCAAGTCTGTGTGAGAAGGCCACCTCAAGA	85
VIM/GFAP	VIM	NM_003380	chr10	17,277,370	GFAP	NM_002055	chr17	42,988,621	AATGTTAAGATGGCCCTTGACATTGAGATTGCCACCTACAGGAAGCTGCTTAGAGGGCGAGGAGAACCGGATACCAATTCCGTGCGAGACCTTCTCCAAACC	86
VIM/GFAP	VIM	NM_003380	chr10	17,271,830	GFAP	NM_002055	chr17	42,992,688	GGTGGCTTCTCGGAGCAGCAGAAATAAGATCTGCTGGCCGAGCTCGAGCTGGGCACTCAATGCTGGCTTCAAGGAGACCCGGCGAGTGAAGCGGGGAGAG	87
VIM/GFAP	VIM	NM_003380	chr10	17,277,350	GFAP	NM_002055	chr17	42,988,641	GTGAATACCAAGACCTGCTCAATGTTAAGATGGCCCTTGACATTGAGATTGGCCACCTACAGGAAGCTGTAGAGGGGAGGAGAACCGGATACCAATTCC	88
VIM/GFAP	VIM	NM_003380	chr10	17,277,877	GFAP	NM_002055	chr17	42,988,655	GAAGCGAGGAGGAGCAGGATTTCTCTGCTCTTCCAACTTTTCTCTCCCTTGGACATCGAGATCGCCACCTACAGGAAGCTGTAGAGGGGAGGAGAAC	89
VIM/GFAP	VIM	NM_003380	chr10	17,277,877	GFAP	NM_002055	chr17	42,988,655	GGAAGCGAGGAGGAGCAGGATTTCTCTGCTCTTCCAACTTTTCTCTCCCTTGGACATCGAGATCGCCACCTACAGGAAGCTGTAGAGGGGAGGAGAAC	90
UACA/LTK	chr15	UACA	NM_018003	70,957,001	chr15	LTK	NM_002344	41,799,372	TGATTGACACTCTGCAGCACCAAGTGAAAATCTCTGGAGCAACAGCTGGCCGTGGGCTTGGCCCGGCCAGTCTGGCTCTGCCACCAAGGTGTACCCGA	184
STRN/ALK	chr2	STRN	NM_003162	37,143,221	chr2	ALK	NM_004304	29446394	TACGGGACAGAAITGAATCAGGGAGATATGAAGCTTCCAAAGCTATGATTCTGTACCCCGGAAACACAGGAGCTGAAAGCCATGCAATGCAGATGGAGCTGC	185
STRN/ALK	chr2	STRN	NM_003162	37,143,221	chr2	ALK	NM_004304	29446394	TACGGGACAGAAITGAATCAGGGAGATATGAAGCTTCCAAAGCTATGATTCTGTACCCCGGAAACACAGGAGCTGAAAGCCATGCAATGCAGATGGAGCTGC	186
JHDMID/BRAF	chr7	JHDMID	NM_030647	139,810,895	chr7	BRAF	NM_004333	140481493	TAGACCTGGACACCTTATTAAGAACTTTCTAAAGTAAATTCGAGCAATAGAGTAGACCTTGGTAGACGGGACTCGAGTGATGATTGGGAGATTCTCTGAT	187
JHDMID/BRAF	chr7	JHDMID	NM_030647	139,810,895	chr7	BRAF	NM_004333	140481493	AAACACTTGGTAGACGGGACTCGAGTGATGATTGGGAGATTCTCTGATGACCTGACACCTTATTAAGAACTTTCTAAAGTAAATTCGAGCAATAGAGAA	188
TAXIBP1/BRAF	chr7	TAXIBP1	NM_006024	27,827,222	chr7	BRAF	NM_004333	140481493	CTGAAAAGGAAATCTGAAAAGAACTTTCTCTTACAACTTACAACTTCAAGTAAATAA	189
MRN1/BRAF	chr7	MRN1	NM_013446	140,158,807	chr7	BRAF	NM_004333	140487384	AACACTTGGTAGACGGGACTCGAGTGATGATTGGGAGATTCTCTGATGGCAGTCTGCAATGGATGCTGCCAGAGATCGCAGCATATCAAAGA	190
MACF1/BRAF	chr1	MACF1	NM_012090	39,896,580	chr7	BRAF	NM_004333	140487384	TTGGACAAAGGGTGGATGAAATTTGATGCTGCTGTTTTCAGAGATCAACACAGGCTTGATAGAGACCAAGGATTCGTTGGTGGAGGATCAACCCACAGG	191
CDC27/BRAF	chr17	CDC27	NM_001256	45,206,816	chr7	BRAF	NM_004333	140487365	CTTGATTAGAGACCAAGGATTCGTTGGTGGAGGATCGCAGCATATCAAAGA	192

Table 5: Breakpoint sequences for Table 2

Table 5 Cancer	Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chromosome	3' Gene Symbol	3' Accession	3' Gene Breakpoint	Breakpoint Sequence
Melanoma	CLCN6/RAF1	chr1	CLCN6	NM_001286	11867247	chr3	RAF1	NM_002880	12641914	GAGAAACACAGGAGGAGGAGGATGAGATTCTTCCAAAGGAAAGACTATGAGTGATGCAATTCGAAGTCACAGGGAATCAGCCTCACCTTTCAGCCCTGCCAG SEQ ID NO: 18

Table 5 Cancer	Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Break point	3' Gene Chromosome	3' Gene	3' Accession	3' Gene Breakpoint	Breakpoint Sequence
Melanoma	TRAK1/RAF1	chr3	TRAK1	NM_014965	42235390	chr3	RAF1	NM_002880	12641914	TCAGCATCTGGGGGCTGCTAAGGATGCCAGCGGCGAGCTCACAGC CGAGCATGCAATTCGAATTCAGACGGAATCAGCCTCACCTTCAGCC CTGTCCAG SEQ ID NO: 19
Colon adenocarcinoma	PRKAC A/AKT1	chr19	PRKAC A	NM_002730	14208406	chr14	AKT1	NM_005163	1.05E+08	AGGCGCGACCTTGACCTTGTGGCGCACCCCTGAGTACCTGGCCCC TGAGTGTCTGGAGGACAAATGACTACGGCCGTGCACTGGACTGGTG GGGCTGGG SEQ ID NO: 20
Colon adenocarcinoma	PRKAC A/AKT1	chr19	PRKAC A	NM_002730	14208406	chr14	AKT1	NM_005163	1.05E+08	AGGCGCGACCTTGACCTTGTGGCGCACCCCTGAGTACCTGGCCCC TGAGTGTCTGGAGGACAAATGACTACGGCCGTGCACTGGACTGGTG GGGCTGGG SEQ ID NO: 21
Colon adenocarcinoma	PRKAC A/AKT1	chr19	PRKAC A	NM_002730	14208406	chr14	AKT1	NM_005163	1.05E+08	AGGCGCGACCTTGACCTTGTGGCGCACCCCTGAGTACCTGGCCCC TGAGTGTCTGGAGGACAAATGACTACGGCCGTGCACTGGACTGGTG GGGCTGGG SEQ ID NO: 22
Endometrial adenocarcinoma	PRKAC A/AKT1	chr19	PRKAC A	NM_002730	14208406	chr14	AKT1	NM_005163	1.05E+08	AGGCGCGACCTTGACCTTGTGGCGCACCCCTGAGTACCTGGCCCC TGAGTGTCTGGAGGACAAATGACTACGGCCGTGCACTGGACTGGTG GGGCTGGG SEQ ID NO: 23
Endometrial adenocarcinoma	PRKAC A/AKT1	chr19	PRKAC A	NM_002730	14208406	chr14	AKT1	NM_005163	1.05E+08	AGGCGCGACCTTGACCTTGTGGCGCACCCCTGAGTACCTGGCCCC TGAGTGTCTGGAGGACAAATGACTACGGCCGTGCACTGGACTGGTG GGGCTGGG SEQ ID NO: 24
Colon adenocarcinoma	PRKAC A/AKT2	chr19	PRKAC A	NM_002730	14208406	chr19	AKT2	NM_001626	4074201	AGGCGCGACCTTGACCTTGTGGCGCACCCCTGAGTACCTGGCCCC TGAGTGTCTGGAGGACAAATGACTACGGCCGTGCACTGGACTGGTG GGGCTGGG SEQ ID NO: 25
Lung adenocarcinoma	MLL/FY N	chr11	MLL	NM_005933	1.18E+08	chr6	FYN	NM_002037	1.12E+08	CCAGGAAGCTCGATCAATGCGCCCTAAAGCAGCTCTCATTTGCA GTGTGTACTTTGGAAACTTGGCCGCAAGATGCTGAGCGACAGC TATTGTCT SEQ ID NO: 26
Lung adenocarcinoma	ECHDC 1/FYN	chr6	ECHDC 1	NM_001002030	1.28E+08	chr6	FYN	NM_002037	1.12E+08	CAAGTTGGGATTTGGTGGAGGACAGAAATTTACTACAGCATGTG ATTTCAGGGAAGGAGATTGGTGGAAAGCCCGCTCTCTTGACAACTGG AGAGACAG SEQ ID NO: 27
Breast carcinoma	TTC13/JAK2	chr1	TTC13	NM_024525	2.31E+08	chr9	JAK2	NM_004972	5055786	CTTCATATCAGAGGACTATGCAACAGCCCATGAAGACTTTTCAGCAG TCCTCTGGAATTTGAATTAAGTCACTTAAGGGAAGCTTTGTCTTCG TGTCAAT SEQ ID NO: 28
Gastric Adenocarcinoma	CAB39/ERBB2	chr2	CAB39	NM_016289	231,577,945	chr17	ERBB2	NM_004448	37,863,243	GGGACAGCGACGACGCGGAGGAGGAGAAAGGGAAGCCCGGCCCA GCCCTGTGACCGGACAGACATGAAGTGGCGCTCCCTGCCAGT CCCGAGACC SEQ ID NO: 29
Gastric Adenocarcinoma	CAPZA 2/MET	chr7	CAPZA 2	NM_006136	116,502,704	chr7	MET	NM_000245	116,435,709	CCAGAAAGGAGATGGCGGATCTGGAGGAGCAGTTGTGTGTAAG AGAAAGTGTCTTTGGCGTCTCTCTCTGGGAGCTGATGACAAGAGG AGCCCAACC SEQ ID NO: 30
Invasive Breast Carcinoma	CBL/UB E4A	chr11	CBL	NM_005188	119,158,656	chr11	UBEA4	NM_004788	118,261,372	CAAAATCAACCTCTCTCACTGCGCAATGCCATTTATTTCTCTGGCTG CCAAGGATGAGGAGAAATTTCTGTGCCACTGTGCCCAAGGATGGAC GTTCCAT SEQ ID NO: 31
Endometrial Adenocarcinoma	EXOC4/BRAF	chr7	EXOC4	NM_021807	133,164,892	chr7	BRAF	NM_004333	140,434,570	TCTCGAGAACACAGAGAGGAGTCTATAGTCGGAGTGGAGAACT GCAAGTATCTCCCTATTGAGCTGTGGCCGCTCATTTGCCAAA AATTCACCG SEQ ID NO: 32
Low Grade	VIM/GF AP	chr10	VIM	NM_003380	17,271,860	chr17	GFAP	NM_002055	42,992,778	CCTGCTGGCGAGCTCGAGCAGCTCAAGGGCCAAAGCAAGTCGCGC CTGGCTCTGGCGCGCTGCTGGTCTGCTGCGACCCGCTCTCCCTGGC

Table 5 Cancer	Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chromosome	3' Gene	3' Accession	3' Gene Breakpoint	Breakpoint Sequence
Glioma										TCGAATG SEQ ID NO: 95
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,276,745	chr17	GFAP	NM_002055	42,988,692	CTGACCTCTCTGAGGCTGCCAACCGGAACAAATGACGCCCTGCGGCA GGCAICAGGAGTACAGGACCTGCTCAATGTCAAGCTGGCCCTTGA CATCGAGT SEQ ID NO: 96
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,276,789	chr17	GFAP	NM_002055	42,990,649	CAGGCAAGCAGGAGTCCACTGAGTACCGGAGACAGTGCAGTCC CTCAGTACCGCTCAAGTTTGCAGACCTGACAGACGCTGCTGCC GCAACGGG SEQ ID NO: 97
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,276,817	chr17	GFAP	NM_002055	42,988,824	TTTCCGTTGAAGCTGCTAACTACCAAGACACTATTGGCCGCTGCA GGAITACAGGACCTGCTCAATGTCAAGCTGGCCCTGGACATCGAG ATCGCCA SEQ ID NO: 98
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,276,817	chr17	GFAP	NM_002055	42,988,824	CATTGAGATTGCCACTACAGGAAGCTGCTGGAAGGCGAGGAGAG CAGAGTACAGGACCTGCTCAATGTCAAGCTGGCCCTGGACATCG AGATCGCCA SEQ ID NO: 99
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,255	chr17	GFAP	NM_002055	42,987,988	GAACCTTGGCTTGAAGCTGCTAACTACCAAGACACTATTGGCCGCTGCA CTGCTTCGAGAAACCAAGCTTGCACACCAAGTCTGTGTCAGAAAGGCC ACCTCAAG SEQ ID NO: 100
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,259	chr17	GFAP	NM_002055	42,988,687	TTTCCGTTGAAGCTGCTAACTACCAAGACACTATTGGCCGCTGCA GGAITACAGGACCTGCTCAATGTCAAGCTGGCCCTGGACATCGAG ATCGCCA SEQ ID NO: 101
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,259	chr17	GFAP	NM_002055	42,988,687	TTTCCGTTGAAGCTGCTAACTACCAAGACACTATTGGCCGCTGCA GGAITACAGGACCTGCTCAATGTCAAGCTGGCCCTGGACATCGAG ATCGCCA SEQ ID NO: 102
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,323	chr17	GFAP	NM_002055	42,988,623	TGAAGAGGAAATGGCTGCTCACTTCTGTGAATACCAAGACCTGCTCA CAATCTAGAGGCGGAGGAAACCGGATCACCATTCCTGTCAGAC CTTCTCCAA SEQ ID NO: 103
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,325	chr17	GFAP	NM_002055	42,988,666	AAGGAGAAATGGCTGCTCACTTCTGTGAATACCAAGACCTGCTCA ATGTCAGCTGGCCCTGGACATCGAGATCGCCACCTACAGGAAGCT GCTAGAGG SEQ ID NO: 104
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,370	chr17	GFAP	NM_002055	42,988,621	AATGTTAAGATGGCCCTTGACATTGACATTGCCACCTACAGGAAGC TGCTAGAGGCGGAGGAAACCGGATCACCATTCCTGTCAGACCTT CTCCAACC SEQ ID NO: 105
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,370	chr17	GFAP	NM_002055	42,988,621	AATGTTAAGATGGCCCTTGACATTGACATTGCCACCTACAGGAAGC TGCTAGAGGCGGAGGAAACCGGATCACCATTCCTGTCAGACCTT CTCCAACC SEQ ID NO: 106
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,370	chr17	GFAP	NM_002055	42,988,621	AATGTTAAGATGGCCCTTGACATTGACATTGCCACCTACAGGAAGC TGCTAGAGGCGGAGGAAACCGGATCACCATTCCTGTCAGACCTT CTCCAACC SEQ ID NO: 107
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,370	chr17	GFAP	NM_002055	42,988,621	AATGTTAAGATGGCCCTTGACATTGACATTGCCACCTACAGGAAGC TGCTAGAGGCGGAGGAAACCGGATCACCATTCCTGTCAGACCTT CTCCAACC SEQ ID NO: 108
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,375	chr17	GFAP	NM_002055	42,988,677	TAAGATGGCCCTTGACATTGACATTGCCACCTACAGGAAGCTGCTG GAAGCGGAGCGGCGGAGTATCAGGAGGCGCTGGCGCGGCTGG AGGAAGAGG SEQ ID NO: 109
Low Grade Glioma	VIM/GF AP	chr10	VIM	NM_003380	17,277,877	chr17	GFAP	NM_002055	42,988,655	GAAGCGGAGGAGGAGGATTTCTGCTTCTCCAAACTTTTCTC CTTGTGACATCGAGATCGCCACCTACAGGAAGCTGCTAGAGGCGG AGGAGAAC SEQ ID NO: 110

Table 5	Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chromosome	3' Gene	3' Accession	3' Gene Breakpoint	Breakpoint Sequence
Low Grade Glioma	VIM/GFAP	chr10	VIM	NM_003380	17,277,877	chr17	GFAP	NM_002055	42,988,655	GAAGCGGAGGAGCAGGATTCTCTGCCTCTTCCAAACATTTTCTCCTCCTTTGGACATCGAGATCGCCACCTACAGGAAGCTGCTAGAGGGCGAGGAGAAC SEQ ID NO: 111
Low Grade Glioma	GFAP/VIM	chr17	GFAP	NM_002055	42,984,756	chr10	VIM	NM_003380	17,278,322	CCTCAAGAGGAAACATCGTGGTGAAGACCGTGGAGATCGCGGGATGGAGAGGATACCCACTCAAAAGGACATCTCTGATTAAGACCGGTTGAACTAGAGA SEQ ID NO: 112
Low Grade Glioma	GFAP/VIM	chr17	GFAP	NM_002055	42,985,436	chr10	VIM	NM_003380	17,277,187	GGCCACTCAAGAGGAAACATCGTGGTGAAGACCGTGGAGATCGGGGATGATGAGATCGGTGAATGGAAGAGAACTTTGCCCGTTGAAGCTGCTAACTACCAA SEQ ID NO: 113
Low Grade Glioma	GFAP/VIM	chr17	GFAP	NM_002055	42,985,438	chr10	VIM	NM_003380	17,277,380	AAGGCCCTCAAGAGGAAACATCGTGGTGAAGACCGTGGAGATGCGGGATGAGAGGATTTCTCTGCCTCTTCCAAACTTTTCTCCCTGGAACCTGA SEQ ID NO: 114
Low Grade Glioma	GFAP/VIM	chr17	GFAP	NM_002055	42,985,452	chr10	VIM	NM_003380	17,277,278	CAAGTCTGTGTGACAGGCGCCACTCAAGAGGAAACATCGTGGTGAAGACCGGGAGAAATGGCTCGTCACTTCTGCGTGAATACCAAGACCTGCTCAATGTGA SEQ ID NO: 115
Low Grade Glioma	GFAP/VIM	chr17	GFAP	NM_001131019	42,987,510	chr10	VIM	NM_003380	17,277,303	TTATACCAATACAGGCTCACAGATTGTAAATGGAACGCCGCCGGCTCGCGAATACCAAGACCTGCTCAATGTTAAGATGGCCCTTGACATTTGAGATTGC SEQ ID NO: 116
Low Grade Glioma	GFAP/VIM	chr17	GFAP	NM_002055	42,987,987	chr10	VIM	NM_003380	17,277,377	AGGAGAACCGGATCACCAATCCCGTGCAGACCTTCTCCAACTGCAATTCCGAGGAGAGCAGGATTTCTCTGCCTCTTCCAAACTTTCTCTCCCTGAACC SEQ ID NO: 117
Low Grade Glioma	GFAP/VIM	chr17	GFAP	NM_002055	42,987,987	chr10	VIM	NM_003380	17,277,377	AGGAGAACCGGATCACCAATCCCGTGCAGACCTTCTCCAACTGCAATTCCGAGGAGAGCAGGATTTCTCTGCCTCTTCCAAACTTTCTCTCCCTGAACC SEQ ID NO: 118
Low Grade Glioma	GFAP/VIM	chr17	GFAP	NM_002055	42,988,642	chr10	VIM	NM_003380	17,277,351	AGGAGTACCAGGACCTGCTCAATGTCAAGCTGGCCCTGGACATCGAGTCCACCTACAGGAAGTGTGGAAGGGCAGGAGAGCAGGATTTCTCTGCC SEQ ID NO: 119
Low Grade Glioma	GFAP/VIM	chr17	GFAP	NM_002055	42,988,655	chr10	VIM	NM_003380	17,277,336	GCCCGCACTTGCAGGATACAGGACCTGCTCAATGTCAAGCTGGCCCTTGTGACATTTGAGATTGCCACCTACAGGAAGCTGCTGGAAAGCGAGGAG SEQ ID NO: 120
Endometrial Adenocarcinoma	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTCTTCCCAACAGGTGGAAGGAGGAGCTGCTCTCAGGCTCGGTCCAGCAACAGTGGCCAGGCTACTACCAGTCACACCTAGACCTGGAGGATC SEQ ID NO: 121
Endometrial Adenocarcinoma	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTCTTCCCAACAGGTGGAAGGAGGAGCTGCTCTCAGGCTCGGTCCAGCAACAGTGGCCAGGCTACTACCAGTCACACCTAGACCTGGAGGATC SEQ ID NO: 122
Endometrial Adenocarcinoma	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTCTTCCCAACAGGTGGAAGGAGGAGCTGCTCTCAGGCTCGGTCCAGCAACAGTGGCCAGGCTACTACCAGTCACACCTAGACCTGGAGGATC SEQ ID NO: 123

Table 5 Cancer	Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chromosome	3' Gene	3' Accession	3' Gene Breakpoint	Breakpoint Sequence
Adenocarcinoma										
Endometrial	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTTCTTCCACAGGTGGAAAAGGAGGGAGGCTGCTCTCAGGCT GCGTCCAGCAACAGTGCCAGGCTACTACCAAGTACACCTAGACCT GGAGGATC SEQ ID NO: 124
Endometrial										
Adenocarcinoma										
Endometrial	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTTCTTCCACAGGTGGAAAAGGAGGGAGGCTGCTCTCAGGCT GCGTCCAGCAACAGTGCCAGGCTACTACCAAGTACACCTAGACCT GGAGGATC SEQ ID NO: 125
Endometrial										
Adenocarcinoma										
Endometrial	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTTCTTCCACAGGTGGAAAAGGAGGGAGGCTGCTCTCAGGCT GCGTCCAGCAACAGTGCCAGGCTACTACCAAGTACACCTAGACCT GGAGGATC SEQ ID NO: 126
Endometrial										
Adenocarcinoma										
Endometrial	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTTCTTCCACAGGTGGAAAAGGAGGGAGGCTGCTCTCAGGCT GCGTCCAGCAACAGTGCCAGGCTACTACCAAGTACACCTAGACCT GGAGGATC SEQ ID NO: 127
Endometrial										
Adenocarcinoma										
Endometrial	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTTCTTCCACAGGTGGAAAAGGAGGGAGGCTGCTCTCAGGCT GCGTCCAGCAACAGTGCCAGGCTACTACCAAGTACACCTAGACCT GGAGGATC SEQ ID NO: 128
Endometrial										
Adenocarcinoma										
Endometrial	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTTCTTCCACAGGTGGAAAAGGAGGGAGGCTGCTCTCAGGCT GCGTCCAGCAACAGTGCCAGGCTACTACCAAGTACACCTAGACCT GGAGGATC SEQ ID NO: 129
Endometrial										
Adenocarcinoma										
Endometrial	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTTCTTCCACAGGTGGAAAAGGAGGGAGGCTGCTCTCAGGCT GCGTCCAGCAACAGTGCCAGGCTACTACCAAGTACACCTAGACCT GGAGGATC SEQ ID NO: 130
Endometrial										
Adenocarcinoma										
Endometrial	HLA-C/MUC16	chr6	HLA-C	NM_002117	31,237,270	chr19	MUC16	NM_024690	8,959,665	GCATTTTCTTCCACAGGTGGAAAAGGAGGGAGGCTGCTCTCAGGCT GCGTCCAGCAACAGTGCCAGGCTACTACCAAGTACACCTAGACCT GGAGGATC SEQ ID NO: 131
Endometrial										
Adenocarcinoma										
Invasive Breast Carcinoma	HOOK3/IKBK	chr8	HOOK3	NM_032410	42,798,568	chr8	IKBK	NM_001556	42,147,725	GATGAGCAGAGCTTGGAAAGGATGCTTACGCTCATCTTAGGCTGTG CTGTGAACCTTGGCGCCCAATGACCTGCCCTGCTGGCCATGGAGTA CTGCCAAG SEQ ID NO: 131

Table 5 Cancer	Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chromosome	3' Gene	3' Accession	3' Gene Breakpoint	Breakpoint Sequence
ma										
Invasive Breast Carcinoma	HOOK3/ IKBK	chr8	HOOK3	NM_032410	42,798,588	chr8	IKBK	NM_001556	42,162,705	GGATGCTTCAGCTCATCTTAGGCTGTGCTGTGAACCTGTGTGAACAGAA GCAAGCCTCTGGCTTAGATACCTTCATGAAAAACAGAAATCATCCAT CGGATCT SEQ ID NO: 132
Ovarian Serous Cystadenocarcinoma	IGFBP2/ SPP1	chr2	IGFBP2	NM_000597	217,528,783	chr4	SPP1	NM_000582	88,896,866	GGAGCCCCCACCCTCCGGGGGACCCCGAGTGTATCTCTTCTAC AATGACAGCAGAGGAGGAGGAGCAGACAGCATCGTCGGGACAGAGA CTCGTCTCA SEQ ID NO: 133
Ovarian Serous Cystadenocarcinoma	IGFBP2/ SPP1	chr2	IGFBP2	NM_000597	217,528,783	chr4	SPP1	NM_000582	88,896,866	TGAGACGAGTCTGTCCCGACGATGTGTGTCTTGCCTCTCTCTGCT GCTCATTTAGAGAGATGACACTCGGGGTCCCCCGGATGTGTGGG GGCTCCC SEQ ID NO: 134
Invasive Breast Carcinoma	KRIT1/ CDK6	chr7	KRIT1	NM_004912	91,842,555	chr7	CDK6	NM_001259	92,462,486	ATATTTACAAAGGCAAGCCGCCAGCAATCATATAAGTATCCCTGTGT ATGTAGGAGGATGCCGCTCTCCACCATCCGGGAGGTGGCGGTGC TGAGGCAC SEQ ID NO: 135
Head and Neck Squamous Cell Carcinoma	LYN/NT RK3	chr8	LYN	NM_002350	56,866,524	chr15	NTRK3	NM_002530	88,670,398	AGATCCCGGGGAGTCCATCAAGTTGGTGAAGAGCTTGGCGCTGG GCAGTTTGGGTATCCATAGCAGTTGGACTTGTGCTTTTGCCTGTG TCCTGTT SEQ ID NO: 136
Invasive Breast Carcinoma	MLLT6/ ACE	chr17	MLLT6	NM_005937	36,868,267	chr17	ACE	NM_000789	61,573,755	CCAGCAGCAGGAGAGAGCACCACCCACCCACGAGAGGGGCCAGA AGAAAGTACTTTGTACGCTTCATCATCCAGTTCCAGTTCCACGAGG CACTGTGCC SEQ ID NO: 137
Invasive Breast Carcinoma	MLLT6/ ACE	chr17	MLLT6	NM_005937	36,868,267	chr17	ACE	NM_000789	61,573,755	CCTGGCACAGTGCCTCGTGGAACTGGAACTGGATGATGAAGCTGAC AAAGTACTTCTTCTGGCCCCCTCTCTGTGGGTGGGTGCTTCTTCC TGCTGCG SEQ ID NO: 138
Ovarian Serous Cystadenocarcinoma	MUC16/ OR7G2	chr19	MUC16	NM_024690	9,024,134	chr19	OR7G2	NM_001005193	9,213,935	AGTGATCTCAGAACTCAGGACTCCATCTCTCTCTCCAGCCCCA CAATTCATCATCAACAGCATGGAAGCGAGAAACCAACACAGCTAT TTCAAAAT SEQ ID NO: 139
Ovarian Serous Cystadenocarcinoma	MUC16/ OR7G2	chr19	MUC16	NM_024690	9,045,564	chr19	OR7G2	NM_001005193	9,213,935	ATTTGAAATAGCTGTTTGTCTTCTCGCTTCCATGCTGTGTGATGATG AAITTTCTTGTGAGTCACTCTCTCAGGGCCAAAGGTGGACATCCC AGGTGTG SEQ ID NO: 140
Invasive Breast	NARS2/ TOP1	chr11	NARS2	NM_024678	78,189,672	chr20	TOP1	NM_003286	39,721,138	GGAACCTTCAAGGCTCAACAATGATGGTTCTCTCAAAATGTCTT GAAGTGCATCAAGTGGAAATTCCTAGAACATAAAGGTCCAGTATTTG

Table 5 Cancer	Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Accession	5' Gene Breakpoint	3' Gene Chromosome	3' Gene	3' Accession	3' Gene Breakpoint	Breakpoint Sequence
Carcinoma										CCCCACCA SEQ ID NO: 141
Invasive Breast Carcinoma	SRD5A1 / PAPD7	chr5	SRD5A1	NM_001047	6,633,982	chr5	PAPD7	NM_006999	6,738,796	GGGCCCCAATGCGATCTCTGCGCCATGTTCTCTGCTCCACTAGGGGCA TCGJTACAGATATTTGGCAGCTTTAGTACAGGTCTTTATCTTCCAACTAGCGA SEQ ID NO: 142
Invasive Breast Carcinoma	PAPD7 / SRD5A1	chr5	PAPD7	NM_006999	6,746,451	chr5	SRD5A1	NM_001047	6,662,933	GGGAGAAATTTTAAATTTACTTTGAAAACCGGTATTTAGAAATCAAAAGAG GAGJCTTATTTGAATACGTAACTGCAGCAACTATTTTGGAGAAATCATGGAGT SEQ ID NO: 143
Gastric Adenocarcinoma	PRKAR2A / RHOA	chr3	PRKAR2A	NM_004157	48,845,082	chr3	RHOA	NM_001664	49,405,981	GACGAGGACTTGAAGTTCCAGTTCTTAGCAGATTTTAAATAGACGAG TATCIAGGTAGAGTTGGCTTTGTGGGACACAGCTGGGCGAGGAAGATTATGATCGC SEQ ID NO: 144
Gastric Adenocarcinoma	TRAPP C9 / PTK2	chr8	TRAPP C9	NM_031466	141,460,889	chr8	PTK2	NM_005607	141,900,868	CTCTGTCCCGTTTGAGAAAAAGGACTTTGTAGGACTGGACACAG ACAGCAGAATATGACAGATACCTAGCATCTAGCAAAAATAATGGCA GCTGCTTAC SEQ ID NO: 145
Gastric Adenocarcinoma	PTK2 / TRAPP C9	chr8	PTK2	NM_005607	142,011,224	chr8	TRAPP C9	NM_031466	141,034,176	CCGCCCGTCTGCTGCTTCCGTTTCAACGGGCGGAGCCGGGTC CGAICCTGGAGCTGCTGCTGAATTTCAAAATCTCTGGAGGCCCGG GCCACAC SEQ ID NO: 146
Ovarian Serous Cystadenocarcinoma	RAB11B / MDK	chr19	RAB11B	NM_004218	8,468,319	chr11	MDK	NM_002391	46,404,173	AGGAAGCATTCAGAAACATCTCTCACAGAGATCTACCGCATCTGTC ACAGTGTATGGGCGACAGCACCAAGTCCGCCAAGGACCCCTGAAGAAGCG SEQ ID NO: 147
Ovarian Serous Cystadenocarcinoma	RAB11B / MDK	chr19	RAB11B	NM_004218	8,468,374	chr11	MDK	NM_002391	46,404,248	GATCGCAGCGCGGCTGCCACGACGAGTCCCGGGGGAACACGCTG GTGJCATCCGCGTACCAAGCCCTGCACCCCAAGACCAAGCA AAGGCCAAA SEQ ID NO: 148
Squamous Lung Carcinoma	RBI / GAD45GIP1	chr13	RBI	NM_000321	48,955,574	chr19	GAD45GIP1	NM_052850	13,065,313	AAAAATTTAGAACGATGTGAACATCGAATCATGGAATCCCTTGCA TGGCACAAGATGCCACAGATGATTTGTGAACCTGGCAGCAGCAGCAGC GGGAGAACT SEQ ID NO: 149
Cutaneous Melanoma	SHANK3 / MAPK1	chr22	SHANK3	NM_033517	51,115,121	chr22	MAPK1	NM_002745	22,153,417	TTTATGCCAGAACCTCATCGATGATAAGCAGTTTGGCAAGCTTCAC ACAAGATCTGTGACTTTGGCCTGGCCCTGTTGCAGATCCAGACC ATGATCA SEQ ID NO: 150
Thyroid Gland Carcinoma	SPECC1L / RET	chr22	SPECC1L	NM_015330	24,734,416	chr10	RET	NM_020630	43,610,055	TGCAGCTGCAATTCCTCGAAGCCCTGAGCCCAAGTCTATGAAA ACCCTCTCAGCTGAGTACCTTCGCGAGGCCCGCCAGGCCCTT CCCGGTCA SEQ ID NO: 151
Glioblastoma	TAOK1 / RARA	chr17	TAOK1	NM_020791	27,718,042	chr17	RARA	NM_000964	38,504,568	GGGAGGCTGGGCACTATCTCTTCAGAACTGTGCTCTGGGTCTCA ATGJCTTCCCGCAGAGTCTGGGGGAGCAGGAGGCGCAGCAGC CCCCTGCA SEQ ID NO: 152
Gastric	THRA / C	chr17	THRA	NM_003	38,245,	chr17	CDK	NM_01508	37,686,8	CAACCCAGCGAAACACAACTTCGCGCACTTCTGGCCCAAGCTGCTG

Table 5 Cancer	Fusion Name	5' Gene Chromosome	5' Gene Symbol	5' Gene Accession	5' Gene Break point	3' Gene Chromosome	3' Gene	3' Accession	3' Gene Breakpoint	Breakpoint Sequence
Adenocarcinoma	DK12			250	586		12	3	84	ATGATGAGAGAGAGGCCCTGAGCCCGGACCTCCACCCGCGC CACCTCCAC SEQ ID NO: 153
Invasive Breast Carcinoma	WRN/A DAM9	chr8	WRN	NM_000553	30,982,516	chr8	ADA M9	NM_003816	38,871,484	TCCTTGGGAATTATGGGAAGTAAAAATGCTGTGATAATTCGAGGT CCAGIAGACCTTTTCCTGAAGATTGTTGTTTATACCTTACAACAA GGAAGGGA SEQ ID NO: 154
Colon and Rectal Adenocarcinoma	YWHA E/MAP2 K2	chr17	YWHA E	NM_006761	1,303,359	chr19	MAP 2K2	NM_030662	4,123,868	CGCTATGGATGATCGAGAGGATCTGGTGTACCAAGGCGAAGCTGGCC GAGCTGGCCCGGAGGAGGCGGTCTGCCGCGCTCACCATCAAC CCTACCATC SEQ ID NO: 155
Thyroid Gland Carcinoma	ZC3HA V1/BR AF	chr7	ZC3HA V1	NM_020119	138,758,639	chr7	BRA F	NM_004333	140482825	ACCAAGCCAGCCAAATCTGTCTTCCACCACCAAAATGGATTG GTATTGGAAJGAATGAAAAACACTTGGTAGACCGGACTCGAG TGATGATTGGGAGATTCCCT SEQ ID NO: 193
Thyroid Gland Carcinoma	BRAF/ SND1	chr7	SND1	NM_014390	127,361,454	chr7	BRA F	NM_004333	140487384	TTCACCTGTCCAGCATCCGACCCAGGAGGCTGGAGGGGGAG AACACCCAGTACTTGTATTAGAGACCAAGGATTTCGTGGTG ATGGAGGATCAACCAACAGG SEQ ID NO: 194
Thyroid Gland Carcinoma	BRAF/ SND1	chr7	BRAF	NM_004333	140,487,348	chr7	SND 1	NM_014390	127724776	GTCAATATTGATGACTTGTATTAGAGACCAAGGATTTCGTGG TGATGGAGGACACCAAGTGGAGAGCTGATGGAGAACATG CGCAATGACATTGCCAGTC SEQ ID NO: 195
Thyroid Gland Carcinoma	SND1/ BRAF	chr7	SND1	NM_014390	127,361,454	chr7	BRA F	NM_004333	140487384	CACCTGTCCAGCATCCGACCCAGGCTGGAGGGGGAGA ACACCCAGGACTTGTATTAGAGACCAAGGATTTCGTGGTGAT GGAGGATCAACCAACAGGTT SEQ ID NO: 196
Thyroid Gland Carcinoma	MEMO 1/ALK	chr2	MEMO 1	NM_015955	32,168,371	chr2	ALK	NM_004304	29543748	GGCTTTCACAAAGTACAGTCTACAAAAAGACCTGCTAGAGCC ATTATTGCCCGGAAACTGCCTGTGGGTTTTTACTGCAACT TTGAAGATGGCTTCTGTG SEQ ID NO: 197
Head and Neck Squamous Cell Carcinoma	CLIP4/ ALK	chr2	CLIP4	NM_024692	29,404,563	chr2	ALK	NM_004304	29462609	GAGGGGTCTCAGGTCTCTCAGAGCTCCAATGAGATGGG TACTGTAGTGTGAAGATGCCAGCACACACGCCGTGG GACCGCATCATGGTGTCT SEQ ID NO: 198
Squamous Cell Lung Carcinoma	CLIP4/ ALK	chr2	CLIP4	NM_024692	29,404,561	chr2	ALK	NM_004304	29462607	ACGAGGGGTCTCAGGTCTCTCAGAGCTCCAATGAGATG GGTACTGTAGGTGAAGATGCCAGCACACACGCCGT GGGACCGCATCATGGTGT SEQ ID NO: 199

Table 6: breakpoint sequences for Table 3

Table 6 Fusion Name	5' Gene Chro m	5' Gene Sym bol	5' Accession	5' Gene Breakpoint	3' Gene Chro moso me	3' Gene Symbol	3' Acc essi on	3' Gene Brea kpoi nt	Breakpoint Sequence	SEQ ID NO:
SEC16A - NOTCH1	chr9	SEC 16A	NM_0148 66	139357445	chr9	NOTCH1	NM_017617	1,39 E+08	ATTGATTTCACGAATGAGG CAGTGGAGCAGGTGGAAG AGGAGGAGTCTGG CCCCG GATGCTCCAGCCCCGGTGA GACCTGCCTGAATGGCGGG AAGTGTC	28
ERC1 - RET	chr12	ERC 1	NM_1780 39	1,250,953	chr10	RET	NM_020630	4361 2032	GGACATGTTGGATGTGAAG GAGCGGAAGGTTAATGTTTC TTCAGAAAGAGG AGGATC CAAAGTGGGAATTCCCTCG GAAGAACTTGGTCTTGGA AAACT	29
ESR1/CC DC170	chr6	ESR 1	NM_0001 25	152,332,929	chr6	CCDC170	NM_025059	151,9 07,02 4	CATGGAGCACCCAGGGAA GCTACTGTTGCTCCTAACT TGCTCTTGGACA GATGGTC TCCCAGCTTGAAGCCCAA TATCTGAGCTTGTGAACA GTTGG	156
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,990,165	chr17	VMP1	NM_030938	57,91 5,656	CAGAATGTTTGTAGCTACT TCGGGTACTTGGTAAAGGG GGCTATGGAAG TGCTGTC CCCGGCATAGTCCCATCTC TGCAGAAAGCCATTTCAGGA GTACC	157
VMP1/RP S6KB1	chr17	VMP 1	NM_0309 38	57,915,758	chr17	RPS6KB1	NM_003161	57,98 7,923	GTTTCATATGGTCCAACTCC CCCATGGTCCATGCTTTTCAT TTAACTGACCC TGTGGTGT GCCCATTTGGCTTTTGTGGT GAAGCTTCTGCCGTTGAGC CTC	158

Table 6 Fusion Name	5' Gene Chro m	5' Gene Sym bol	5' Accession	5' Gene Breakpoint	3' Gene Chro moso me	3' Gene Symbol	3' Acc essi on	3' Gene Break point	Breakpoint Sequence	SEQ ID NO:
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,686	chr17	VMP1	NM_03_093 8	57,915,656	AGACCTGGACCAGCCAGAG GACGGGGCTCTGAGGATG AGCTGGAGGAGG GGTGCT GTCCCGGCATAGGTCCAT CTCTGCAGAAAGCCATTCA GGAGTA	159
VMP1/RP S6KB1	chr17	VMP1	NM_0309 38	57,915,758	chr17	RPS6KB1	NM_00316 1	57,987,923	AAGTTCATATGGTCCAACT CCCCATGGTCCATGCTTTC ATTTAACTGAC CCTGTGGT GTGCCCATTTGCGTTTGTG GTGAAGCTTCTGCCGTTGA GCC	160
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,606	chr17	VMP1	NM_03_093 8	57,915,656	GGTACTCTGAAATGGCTT CTGCAGAGATGGACCTATG CCGGGACAGCA CTTCCCT GTCTCGAAATCCGGGGCT GGTAAAAGCCGTCGCCGCC TCCTT	161
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,992,064	chr17	VMP1	NM_03_093 8	57,915,656	GTACAGGAGCAAA TACTG GGAAAATATTTGCCATGAA GGTGCTTAAAAA GTGCTGT CCCGGCATAGGTCCATCT CTGCAGAAAGCCATTTCAGG AGTAC	162
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	58,003,943	chr17	VMP1	NM_03_093 8	57,917,129	GCCTTCAGACTGGTGGAA AACTCTACCTCATCCTTGA GTATCTCAGTGG GAGAAA ACTGGTTGCTCTGGATGTTT GAAAAGTTGGTCGTTGTCA TGGTG	163
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,686	chr17	VMP1	NM_03_093 8	57,915,656	AGACCTGGACCAGCCAGAG GACGGGGCTCTGAGGATG AGCTGGAGGAGG GGTGCT GTCCCGGCATAGGTCCAT CTCTGCAGAAAGCCATTCA GGAGTA	164

Table 6 Fusion Name	5' Gene Chro m	5' Gene Sym bol	5' Accession	5' Gene Breakpoint	3' Gene Chro moso me	3' Gene Symbol	3' Acc essi on	3' Gene Break point	Breakpoint Sequence	SEQ ID NO:
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,990,165	chr17	VMP1	NM_03_093_8	57,91 5,656	CAGAAATGTTTGGAGCTACT TCGGGTACTTGGTAAAGGG GGCTATGGAAAGTGTCTGTC CCCGCATAGGTCCATCTC TGCAGAAAGCCATTTACAGGA GTACC	165
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	58,003,943	chr17	VMP1	NM_03_093_8	57,91 7,129	ATGCCCTTTCAGACTGGTGG AAACTCTACCTCATCCIT GAGTATCTCAGTGGGAGA AAACTGGTGTCTCTGGATG TTTGAAAAGTTGGTCGTTG TCATGG	166
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	58,009,009	chr17	VMP1	NM_03_093_8	57,91 7,215	ATATTATGGAAGACACTG CCTGCTTTTACTTTGGCAGA AATCTCCATGGCJACAAAGT TATGCCAAACGAAATCCAGC AGCGTTGAACTCAGAGGA GAAAA	167
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	58,009,061	chr17	VMP1	NM_03_093_8	57,89 5,132	TGGGGCATTTACATCAAAA GGGGATCATCTACAGAGAC CTGAAGCCGAGTGGTGCT GTCCCGGCATAGGTCCAT CTCTGAGAAAGCCATTCA GGAGT	168
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,625	chr17	VMP1	NM_03_093_8	57,91 5,703	TACCCAGCCCGGACTTCC GAGACAGGGAAGCTGAGG ACATGGCAGGAGTACCTG GAGGCTCAACGGCAGAAAG CTTCAACCAAAAAGCGAAA TGGGCACA	169
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,643	chr17	VMP1	NM_03_093_8	57,91 5,710	CCTGTGGTGTGCCCCATTTC GCTTTTGTGGTGAAGCTTCT GCCGTTGAGCCJTCAGGTC TATGTCAAAACACTCTCTGCC ATGTCTCAGCTTCCCTGTC TCG	170

Table 6 Fusion Name	5' Gene Chro m	5' Gene Sym bol	5' Accession	5' Gene Breakpoint	3' Gene Chro some	3' Gene Symbol	3' - Accessi on	3' - Gene Break point	Breakpoint Sequence	SEQ ID NO:
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,992,064	chr17	VMP1	NM_03_093 8	57,88 6,157	AACAGGAGCAAACTACTGG GAAAATATTGCCATGAAG GTGCTTAAAAAGGACTTTG CCTCCGGGCCAAACTGGC AGTTCAAAAACTAGTACAG AAAGTT	171
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	58,007,535	chr17	VMP1	NM_03_093 8	57,91 5,656	CTATTATGCAGTTAGAAA GAGAGGGAATATTTATGGA AGACACTGCCCTGTGCTGTC CCGGCATAAGTCCATCTC TGCAGAAGCCATTTTCAGGA GTACC	172
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,625	chr17	VMP1	NM_03_093 8	57,91 5,703	TGTGCCATTTCGCTTTGT GGTGAAGCTTCTGCCGTTG AGCCTCCAGGTACTCCTGC CATGTCTCAGCTTCCCTGT CTCGGAAGTCCGGGGCTGG GTA	173
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,990,165	chr17	VMP1	NM_03_093 8	57,91 7,129	CCAGAAATGTTTGAGCTAC TTCGGGTACTTGGTAAAGG GGCTATGGAAAGGGAGA AAACTGGTTGTCCTGGATG TTTGAAAAAGTTGGTGGTTG TCATGG	174
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,992,037	chr17	VMP1	NM_03_093 8	57,85 1,147	ATGGAAGGTTTTCACAGT ACGAAAAGTAACAGGAGC AAATACTGGGAAAJATATT CATGGCCAGAGCAGCTCGC CTCAAGGTGCTGAACCCAG ATGATG	175
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,686	chr17	VMP1	NM_03_093 8	57,88 9,031	ACCTGGACCAGCCAGAGGA CGCGGCTCTGAGGATGAG CTGGAGGAGGGGATTTCCA AATCCCTTATTGATCTGGC TGGAAATAACGTGTGGACAC TTTCT	176

Table 6 Fusion Name	5' Gene Chro m	5' Gene Sym bol	5' Accession	5' Gene Breakpoint	3' Gene Chro some	3' Gene Symbol	3' Accession	3' Gene Breakpoint	Breakpoint Sequence	SEQ ID NO:
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,686	chr17	VMP1	NM_0031 093 8	57,88 6,157	ACCTGGACCAGCCAGAGGA CGCGGGCTCTGAGGATGAG CTGGAGGAGGGG GACTTT GCCTCCCGGGCCAACTGG CAGTTCAAAAACTAGTACA GAAAGT	177
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,686	chr17	VMP1	NM_0031 093 8	57,88 9,031	GAAAGTGCCACACGTTAT TCCAGCCAGATCAAAATAA GGATTGGGAATC CCCTCCT CCAGCTCATCCTCAGAGCC CGCGTCTCTGGCTGGTCC AGGTC	178
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,686	chr17	VMP1	NM_0031 093 8	57,88 6,157	CTGGACCAGCCAGAGGACG CGGGCTCTGAGGATGAGCT GGAGGAGGGG CTTTGC CTCCCGGGCCAACTGGCA GTTCAAAAACTAGTACAGA AAGTTG	179
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,686	chr17	VMP1	NM_0031 093 8	57,88 6,157	CCTGGACCAGCCAGAGGAC GCGGGCTCTGAGGATGAGC TGGAGGAGGGG GACTTTG CTCCCGGGCCAACTGGC AGTCAAAAACTAGTACAG AAGTT	180
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,686	chr17	VMP1	NM_0031 093 8	57,88 6,157	ACCTGGACCAGCCAGAGGA CGCGGGCTCTGAGGATGAG CTGGAGGAGGGG GACTTT GCCTCCCGGGCCAACTGG CAGTTCAAAAACTAGTACA GAAAGT	181
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,970,686	chr17	VMP1	NM_0031 093 8	57,88 6,157	CGCGGGCTCTGAGGATGAG CTGGAGGAGGGG CTTT GCCTCCCGGGCCAACTGG CAGTTCAAAAACTAGTACA GAAAGT	182

Table 6 Fusion Name	5' Gene Chro m	5' Gene Sym bol	5' Accession	5' Gene Breakpoint	3' Gene Chro some	3' Gene Symbol	3' Acc essi on	3' Gene Break point	Breakpoint Sequence	SEQ ID NO:
RPS6KB1 /VMP1	chr17	RPS6 KB1	NM_0031 61	57,992,064	chr17	VMP1	NM_03 093 8	57,91 5,656	AGTAACAGGAGCAAAATACT GGGAAAATATTGCCATGA AGGTGCTTAAAAAAGTGCTG TCCCCGGCATAGGTCCATC TCTGCAGAAAGCCATTTCAG GAGTA	183

[00107] The disclosure provides novel gene fusions and gene fusion variants (ie, varying breakpoint locations on one or both of the partner genes) selected from those shown in Table 1 - Table 3, Table 19, and Table 22 of gene fusions such as TPM1/ALK, PRKAR1A/ALK, NCOA1/ALK, LPP/CASR, MDM2/EGFR, FGFR3/ELAVL3, B2M/GNAS, DOCK8/JAK2, HNF1B/NOTCH1, NFASC/NTRK1, SSBP2/NTRK1, SQSTM1/NTRK1, TBL1XR1/PIK3CA, AKAP13/RET, FKBP15/RET, TBL1XR1/RET, CEP85L/ROS1, CLCN6/RAF1, TRAK1/RAF1, PRKACA/AKT1, PRKACA/AKT2, MLL/FYN, ECHD1/FYN, TTC13/JAK2, SEC16A/NOTCH1, ERC1/RET, GTF2IRD1/ALK, HTATSF1/BRS3, CDH1/CCDC132, CCDC132/CDH1, ERBB2/SLC29A3, MET/TFG; TFG/MET, NOTCH2/MNDA, IRF2BP2/NTRK1, EIF2C2/PTK2, RARA/HOXB3, STAT3/ETV4, and GFAP/VIM; VIM/GFAP, TOP1/C17orf64, and TP53/KIAA0753. As a result of these discoveries, the disclosure provides isolated gene fusion nucleic acids and sequences complementary thereto, amplicons, transcripts, reaction mixtures, as well as probes that specifically recognize the nucleic acid sequences of the gene fusions, sequences complementary thereto, amplicons, and transcripts. The disclosure further contemplates antisense nucleotides for use in the treatment of the associated disease.

[00108] Table 1 - Table 3, Table 19, and Table 22 provide a list of the gene fusions (Gene A/Gene B) indicating the genes involved (Gene A and Gene B), the chromosome locations, the breakpoint locations, the fusion types and the distance. The gene fusions are shown with the associated TCGA disease (The Cancer Genome Atlas). The cancers are shown with 3-4 letter abbreviations which are explained in more detail in the diagnostics section.

[00109] Generally, Tables 1-3, 19, and 22 provide one or more novel gene fusions and/or associations of gene fusions with TCGA diseases. For example, Table 19 presents novel gene fusions, and Table 22 presents novel associations of gene fusions with TCGA diseases.

[00110] Tables 4-6, 20, and 23 provide the breakpoint sequences for the gene fusions in Tables 1-3, 19, and 22. The breakpoint sequences are identified as SEQ ID NO:1-289.

[00111] Assays and kits

In certain embodiments, assays and methods of detection are provided. Methods for detecting gene fusions provided herein are known in the art. As non-limiting examples,

such assays can include 5' nuclease PCR assays (Applied Biosystems, Foster City, CA), next generation sequencing assays (Ion Torrent, Carlsbad CA; Illumina, San Diego, CA), or microarray assays (Skotheim et al., *Molecular Cancer* 2009, 8:5). In at least one embodiment, the assays or methods include at least one primer or probe that is complementary to or encodes a gene fusion and/or breakpoint in Tables 1-6.

[00112] In at least one embodiment, assays and methods of quantitating the amount of expression of a gene fusion are provided. The methods may involve quantitating expression of one or more exons. For example, TaqMan™ Gene Expression Assays can be designed for a set of known fusion transcripts for quantitative analysis. Such assays can be designed such that the primers and probe span the breakpoint region, although in certain illustrative embodiments the primers and probe are not placed directly on the breakpoint.

[00113] In certain embodiments, the disclosure provides a primer, a probe or a set of probes or primers that specifically recognize one or more of the gene fusions and/or breakpoints disclosed herein.

[00114] In one embodiment, the disclosure provides a composition and a kit comprising a set of probes that specifically recognize a gene fusion selected from Tables 1-3, 19, and 22 and/or a breakpoint in Tables 4-6, 20, and 23. The set of probes can be, for example a set of amplification primers. In another embodiment, provided herein is a composition that includes a set of primers that flank a gene fusion selected from Tables 1-3, 19, and 22 in a target nucleic acid. The reaction mixture of this embodiment can further include a detector probe that binds to either side of a breakpoint in a gene fusion selected from Tables 1-3, 19, and 22, or that binds a binding region that spans the breakpoint in a gene fusion selected from Tables 1-3, 19, and 22. The reaction mixture that includes a detector probe or does not include a detector probe, can further include a polymerase, dNTPs, and/or a uracil DNA deglycosylase (UDG). The polymerase and UDG are typically not from a human origin. The reaction mixture can further include a target nucleic acid, for example a human target nucleic acid. The human target nucleic acid can be, for example, isolated from a biological sample from a person suspected of having a cancer.

[00115] In another embodiment, provided herein is a qPCR assay, such as a TaqMan™ assay or a Molecular Beacons™ assay, that specifically amplifies and detects a target nucleic acid that includes SEQ ID NOs: 1-289.

[00116] The disclosure also provides an isolated nucleic acid comprising at least one sequence selected from SEQ ID NOs: 1-289. The isolated nucleic acid can include a first primer on a 5' end. Furthermore, the nucleic acid can be single stranded or double stranded.

[00117] The disclosure, in other embodiments, provides a kit that includes a detector probe and/or a set of probes, for example, a set of amplification primers that specifically recognize a nucleic acid comprising a breakpoint for a gene fusion selected from Tables 1-3, 19, and 22. For example, in certain embodiments the detector probe or set of amplification primers are designed to amplify and/or detect a nucleic acid that includes at least one of SEQ ID NOs:1-289. The kit can further include, in a separate or in the same vessel, a component from an amplification reaction mixture, such as a polymerase, typically not from human origin, dNTPs, and/or UDG. Furthermore, the kit can include a control nucleic acid. For example the control nucleic acid can include a sequence that includes the break point in a gene fusion selected from Tables 1-3, 19, and 22.

[00118] In some embodiments there is provided a kit encompassing at least 2 primer pairs and 2 detectably labeled probes. In these non-limiting embodiments, the 2 primer pairs and/or 2 detectably labeled probes form 2 amplification detection assays.

[00119] The kits of the present invention may also comprise instructions for performing one or more methods described herein and/or a description of one or more compositions or reagents described herein. Instructions and/or descriptions may be in printed form and may be included in a kit insert. A kit also may include a written description of an Internet location that provides such instructions or descriptions.

[00120] In some embodiments, the kits and assays comprise one or more probes that specifically recognize a target, such as a gene fusion nucleic acid sequence. In at least one embodiment, the kits and assays are diagnostic kits and assays.

[00121] A kit comprising a set of probes, for example, a set of amplification primers that specifically recognize a nucleic acid comprising a break point from Tables 4-6, 20, and 23 is provided. The kit can further include, in a separate or in the same vessel, a component from an amplification reaction mixture, such as a polymerase, typically not from human origin, dNTPs, and/or UDG. Furthermore, the kit can include a control nucleic acid. For example the control nucleic acid can include a sequence that includes the break point selected from Tables 4-6, 20, and 23.

[00122] In another embodiment, a gene fusion is provided comprising at least one of the break points in Tables 4-6, 20, and 23.

[00123] In some embodiments, a reaction mixture and a kit are provided. In some embodiments, the kit encompasses a detectable probe that selectively binds a gene fusion. In some embodiments, the gene fusion is any one of the gene fusions in Table 4, Table 5, Table 6, Table 20, or Table 23.

[00124] Thus, in some embodiments are provided a kit encompassing a reaction mixture and a detectable probe that selectively binds a gene fusion, the gene fusion being any one of the gene fusions in Table 4, Table 5, Table 6, Table 20, or Table 23.

[00125] Diagnostics

[00126] Methods of diagnosing, treating, and detecting gene fusions and associated disease are contemplated herein. The methods can include detecting gene fusions in a subject sample.

[00127] A subject sample can be any bodily tissue or fluid that includes nucleic acids from the subject. In certain embodiments, the sample will be a blood sample comprising circulating tumor cells or cell free DNA. In other embodiments, the sample can be a tissue, such as a cancerous tissue. The cancerous tissue can be from a tumor tissue and may be fresh frozen or formalin-fixed, paraffin-embedded (FFPE).

[00128] The disease can be a cancer or tumor. Cancers can include, but are not limited to, melanoma, cervical cancer, pancreatic cancer, head and neck squamous cancer, lung adenocarcinoma, colon adenocarcinoma, uterine carcinoma, ovarian cancer, glioblastoma, low grade glioma, lung adenocarcinoma, thyroid cancer, and gastric cancer.

[00129] Cancers can include but are not limited to, bladder carcinoma, breast carcinoma, cervical cell carcinoma, colon adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, clear cell renal cell carcinoma, kidney renal papillary cell carcinoma, acute myeloid leukemia, brain lower grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, squamous cell lung carcinoma, ovarian serous adenocarcinoma, prostate adenocarcinoma, rectal adenocarcinoma, cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, and uterine corpus endometrioid carcinoma. As used herein, BLCA = bladder carcinoma, BRCA = breast carcinoma, CESC = cervical cell carcinoma, COAD = colon adenocarcinoma, GBM = glioblastoma multiforme, HNSC = head and neck squamous cell carcinoma, KIRK =

clear cell renal cell carcinoma, KIRP = kidney renal papillary cell carcinoma, LAML = acute myeloid leukemia, LGG = brain lower grade glioma, LIHC = liver hepatocellular carcinoma, LUAD = lung adenocarcinoma, LUSC = squamous cell lung carcinoma, OV = ovarian serous adenocarcinoma, PRAD = prostate adenocarcinoma, READ = rectal adenocarcinoma, SKCM = cutaneous melanoma, STAD = stomach adenocarcinoma, THCA = thyroid carcinoma, and UCEC = uterine corpus endometrioid carcinoma.

[00130] In some embodiments, a method of detecting novel gene variants or gene fusions is provided, the method encompassing a reaction mixture, wherein the novel gene variant or gene fusion is detected by the generation of an extension product.

[00131] In another embodiment, the disclosure provides diagnostics and treatment targets utilizing the disclosed gene fusions and gene variants. The gene fusions, gene variants and associated disease states provide targets for both diagnosis and treatment. For instance, the presence, absence, or increased or decreased expression of a gene fusion target or a gene variant can be used to diagnose a disease state or may be used to prognose or detect a disease state. In at least one embodiment, the gene fusion or gene variant can have a high prevalence (frequency) in a particular cancer, a medium prevalence or a low prevalence. In at least one embodiment, the gene fusion or gene variant can have a high frequency in one cancer or tumor and a low or medium prevalence in another. In at least one embodiment, the gene fusion or gene variant can have a medium or low frequency association with a cancer or tumor. In at least one embodiment, a low or medium frequency gene fusion or gene variant can be used in combination with one or more different high frequency biomarkers of cancers to help to diagnose, prognose or identify a predisposition for a disease. The methods can be used for screening for cancer in a patient or predicting the relative prospects of a particular outcome of a cancer. For example, the presence of BRCA1 or BRCA2 mutations can be analyzed in combination with the gene fusion JAK2/TTC13 for breast cancer.

[00132] A method of detecting a cancer is provided comprising amplifying a nucleic acid that spans a breakpoint in a gene fusion selected from Tables 1-3, 19, and 22, for example the nucleic acid can include a sequence selected from SEQ ID NOs: 1-289, and detecting the presence of the nucleic acid, wherein the presence of the nucleic acid indicates a cancer is present in the sample. In another method, provided herein is a method of detecting a cancer that includes generating an amplicon that includes a sequence selected from SEQ ID NOs: 1-289, and detecting the presence of the nucleic

acid, wherein the presence of the nucleic acid indicates the cancer or cancer cell is present in the sample. The amplicon typically includes primers that are extended to form the amplicon. The cancer is selected from bladder carcinoma, breast carcinoma, cervical cell carcinoma, colon adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, clear cell renal cell carcinoma, kidney renal papillary cell carcinoma, acute myeloid leukemia, brain lower grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, squamous cell lung carcinoma, ovarian serous adenocarcinoma, prostate adenocarcinoma, rectal adenocarcinoma, cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, and uterine corpus endometrioid carcinoma.

[00133] In another embodiment is a method to detect a cancer selected from bladder carcinoma, breast carcinoma, cervical cell carcinoma, colon adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, clear cell renal cell carcinoma, kidney renal papillary cell carcinoma, acute myeloid leukemia, brain lower grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, squamous cell lung carcinoma, ovarian serous adenocarcinoma, prostate adenocarcinoma, rectal adenocarcinoma, cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, and uterine corpus endometrioid carcinoma in a sample by detecting the presence of a gene fusion selected from Tables 1-3, 19, and 22.

[00134] New Gene fusions

[00135] Although some of the gene fusions have been previously reported, provided herein, are numerous variations of the gene fusions in which the break points differ and/or that were not previously known. Nonlimiting examples of gene fusions in which the break points differ and/or were not previously known include: TPM1/ALK, PRKAR1A/ALK, NCOA1/ALK, LPP/CASR, MDM2/EGFR, FGFR3/ELAVL3, B2M/GNAS, DOCK8/JAK2, HNF1B/NOTCH1, NFASC/NTRK1, SSBP2/NTRK1, SQSTM1/NTRK1, TBL1XR1/PIK3CA, AKAP13/RET, FKBP15/RET, TBL1XR1/RET, CEP85L/ROS1, CLCN6/RAF1, TRAK1/RAF1, PRKACA/AKT1, PRKACA/AKT2, MLL/FYN, ECHD1/FYN and TTC13/JAK2 are novel variants with the breakpoints provided in Tables 4 and 5 as SEQ ID NOs: 1-289.

[00136] Also provided herein are numerous gene fusion variants that are associated with one or more cancers.

[00137] Cancer Associations

[00138] New gene fusion associations with cancer(s) are presented herein. Some of the gene fusions may have been associated with specific cancers or disease states previously. The methods herein have identified new associations that can be used to help diagnose and/or treat the specific cancers. The gene fusions shown in Tables 1-3, 19, and 22 provide the genes involved in the fusion and the association of that gene fusion with one or more specific cancers. For example, the fusion PRKACA/AKT1 is shown to be associated with colon adenocarcinoma and endometrial endometrioid adenocarcinoma.

[00139] The gene fusions shown in Table 3 are previously known gene fusions that have been shown to be associated with new cancers. For example, SEC16A/NOTCH1 was previously identified as associated with breast cancer. Current methods identified an association of the gene fusion SEC16A/NOTCH1 with thyroid gland carcinoma. Further, ERC1/RET was previously identified as associated with thyroid cancer. Current methods identified an association of the gene fusion ERC1/RET with invasive breast carcinoma (see Tables 3 and 6).

[00140] Reaction Mixtures and Amplicons

[00141] In another embodiment, the disclosure provides a reaction mixture comprising a probe or a set of probes that specifically recognize a gene fusion selected from Table 1 – Table 3, Table 19, and Table 22. The set of probes can be, for example a set of amplification primers or a labeled probe. In another embodiment, provided herein is a reaction mixture that includes a set of primers that flank a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22 in a target nucleic acid. For example, the set of primers can each bind to a target sequence in the human genome within 1000, 750, 500, 250, 100, 90, 80, 75, 70, 65, 50, or 25 nucleotides of opposite sides of the one of the fusion breakpoints identified in Tables 4-6, 20, and 23. The reaction mixture of this embodiment can further include a detector probe that binds to either side of a breakpoint in a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22, or that binds a binding region that spans the breakpoint in a gene fusion selected from Table 1 - Table 3, Table 19, and Table 22, including specific embodiments where the breakpoint is identified in Tables 4-6, 20, and 23. In exemplary embodiments, the detector probe binds to a target sequence in the human genome within 1000, 750, 500, 250, 100, 90, 80, 75, 70, 60, 50, or 25 nucleotides of one of the fusion breakpoints identified in Tables 4-6, 20, and 23. The reaction mixture that includes a detector probe or does not include a detector probe, can further include a

polymerase, a reverse transcriptase, dNTPs, and/or a uracil DNA deglycosylase (UDG). The polymerase, the reverse transcriptase, and the UDG are typically not from human origin. The polymerase in illustrative embodiments is a thermostable polymerase such as a Taq polymerase. In certain embodiments, the dNTPs in the reaction mixture include dUTP, and the reaction mixture can in certain examples, be devoid of dTTP.

[00142] The reaction mixture can further include a target nucleic acid, for example a human target nucleic acid. The human target nucleic acid can be, for example, isolated from a biological sample, such as a tumor sample, from a person suspected of having a cancer selected from: BLCA = bladder carcinoma, BRCA = breast carcinoma, CESC = cervical cell carcinoma, COAD = colon adenocarcinoma, GBM = glioblastoma multiforme, HNSC = head and neck squamous cell carcinoma, KIRC = clear cell renal cell carcinoma, KIRP = kidney renal papillary cell carcinoma, LAML = acute myeloid leukemia, LGG = brain lower grade glioma, LIHC = liver hepatocellular carcinoma, LUAD = lung adenocarcinoma, LUSC = squamous cell lung carcinoma, OV = ovarian serous adenocarcinoma, PRAD = prostate adenocarcinoma, READ = rectal adenocarcinoma, SKCM = cutaneous melanoma, STAD = stomach adenocarcinoma, THCA = thyroid carcinoma, and UCEC = uterine corpus endometrioid carcinoma. In certain embodiments, the target nucleic acid is from a tumor, for example a tumor of one of the cancer types listed in the preceding sentence. Furthermore, the target nucleic acid can be extracted from a biological sample from a tumor such as, for example, an FFPE sample.

[00143] The reaction mixtures of the present invention can include an amplicon. The amplicon can be for example, an isolated nucleic acid. The amplicon can be between 25 and 2500, between 25 and 2000, between 25 and 1000, between 50 and 1000, between 50 and 500, between 50 and 250, between 50 and 200, between 50 and 150, between 50 and 100, or between 50 and 75 nucleotides in length, for example.

[00144] The amplicon can have a nucleotide sequence that is identical or complementary to the target sequence in the human genome within 1000, 750, 500, 250, 100, 90, 80, 75, 70, 65, 50, or 25 nucleotides of opposite sides of the one of the fusion breakpoints identified in Tables 4-6, 20, and 23. In certain embodiments, the amplicon includes 25 to 250, 25 to 100, 25 to 75, 50 to 250, 50 to 200, 50 to 150, 50 to 100, or 50 to 75 of the nucleotide sequence provided in FIGs 4-6, or a complement thereof. In certain embodiments the amplicons includes sequence variants that occur in

nature. For example, the amplicons may include variable nucleotide sequences that correspond to single nucleotide variants or naturally occurring alleles.

[00145] Amplicons of the present invention, in certain illustrative embodiments, have a chemical structure that is not found in nature, and/or not found in a mammal, such as a human. For example, certain illustrative amplicons include a base that is not found in nature or not found in a mammal or that may not be found bound to the type of sugar-phosphate backbone of the amplicon. For example, the amplicon might be a DNA amplicon that includes a uracil base bound to the sugar phosphate backbone, thus having a uridine residue at least at one position and in illustrative examples, at all positions that contain a thymidine residue in a template.

[00146] Accordingly, the amplicon in illustrative embodiments is a DNA amplicon that includes one or more deoxyuridine ("dU") residues. The dU residue can be added by including such residues in the primers used to generate the amplicon. In certain embodiments the reaction mixture includes a DNA amplicon that includes one or more dU residues for every deoxythymidine residue in the corresponding human genomic sequence. These amplicons can be generated, for example, by using a dNTP mix that includes dUTP instead of dTTP when generating the amplicon using an amplification reaction such as PCR.

[00147] In certain embodiments, the amplicon includes a segment for which a corresponding sequence is not found in the human genome, such as, for example, an oligonucleotide sequence, for example a DNA barcode sequence. The non-human segment can be for example, 5-10,000, 5-5000, 5-1000, 5-500, 5-100, 5-50, 5-25, 5-10, 10-10,000, 10-5000, 10-1000, 10-500, 10-100, 10-50, or 10-25 nucleotides in length.

[00148] In certain embodiments, the amplicon includes segment that corresponds to the region of the human genome that spans an intron, but the amplicon does not include a segment corresponding to the intron.

GENE VARIANTS (Table 7 and/or Table 11)**Table 11: Gain of Function mutations**

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Prostate Adenocarcinoma	ACOT7	1	63873	SNP	A	G	G	Oncomine	NM_007274	p.V202A	p.V202	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	ACOT7	1	63873	SNP	A	G	G	Oncomine	NM_007274	p.V202A	p.V202	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	ACOT7	1	63873	SNP	A	G	G	Oncomine	NM_007274	p.V202A	p.V202	Missense_Mutation	Hotspot
Prostate Adenocarcinoma	ANAPC1	2	11262	SNP	G	C	C	Oncomine	NM_022662	p.P222A	p.P222	Missense_Mutation	Hotspot
Medulloblastoma	ANAPC1	2	11262	SNP	G	C	C	Oncomine	NM_022662	p.P222A	p.P222	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	ANAPC1	2	11262	SNP	G	C	C	Oncomine	NM_022662	p.P222A	p.P222	Missense_Mutation	Hotspot
Lung Adenocarcinoma	ANAPC1	2	11262	SNP	G	C	C	Oncomine	NM_022662	p.P222A	p.P222	Missense_Mutation	Hotspot
Papillary Renal Cell Carcinoma	ANAPC1	2	11262	SNP	G	C	C	Oncomine	NM_022662	p.P222A	p.P222	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	C2orf69	2	20049	SNP	G	A	A	Oncomine	NM_153689	p.R119H	p.R119	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	C2orf69	2	8052	SNP	C	T	T	Oncomine	NM_153689	p.R119C	p.R119	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	C2orf69	2	20078	SNP	G	A	A	Oncomine	NM_153689	p.R119H	p.R119	Missense_Mutation	Hotspot
Cutaneous Melanoma	C4orf22	4	81791	SNP	C	T	T	Oncomine	NM_152770	p.R117C	p.R117	Missense_Mutation	Hotspot
Cutaneous Melanoma	C4orf22	4	162	SNP	C	T	T	Oncomine	NM_152770	p.R117C	p.R117	Missense_Mutation	Hotspot
Cutaneous Melanoma	C4orf22	4	81504	SNP	C	T	T	Oncomine	NM_152770	p.T96M	p.T96	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	C4orf22	4	81504	SNP	C	T	T	Oncomine	NM_152770	p.T96M	p.T96	Missense_Mutation	Hotspot
Lung Adenocarcinoma	C4orf3	4	12022	SNP	C	T	T	Oncomine	NM_001001701	p.R18Q	p.R18	Missense_Mutation	Hotspot
Ductal Breast Carcinoma	C4orf3	4	1638	SNP	C	C	G	Oncomine	NM_001001701	p.R18P	p.R18	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Lung Adenocarcinoma	C4orf3	4	12022 1638	SNP	C	T	T	Oncomine	NM_001001701	p.R18Q	p.R18	Missense_Mutation	Hotspot
Prostate Adenocarcinoma	CACNG3	16	24373 167	SNP	C	T	T	Oncomine	NM_006539	p.R311C	p.R311	Missense_Mutation	Hotspot
Cutaneous Melanoma	CACNG3	16	24372 868	SNP	C	T	T	Oncomine	NM_006539	p.S211F	p.S211	Missense_Mutation	Hotspot
Lung Adenocarcinoma	CACNG3	16	24372 930	SNP	C	T	T	Oncomine	NM_006539	p.R232W	p.R232	Missense_Mutation	Hotspot
Glioblastoma	CACNG3	16	24366 270	SNP	G	A	A	Oncomine	NM_006539	p.A138T	p.A138	Missense_Mutation	Hotspot
Astrocytoma	CACNG3	16	24373 167	SNP	C	T	T	Oncomine	NM_006539	p.R311C	p.R311	Missense_Mutation	Hotspot
Colorectal Mucinous Adenocarcinoma	CACNG3	16	24273 772	SNP	C	T	T	Oncomine	NM_006539	p.A138V	p.A138	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	CACNG3	16	24273 771	SNP	G	A	A	Oncomine	NM_006539	p.A138T	p.A138	Missense_Mutation	Hotspot
Lung Adenocarcinoma	CACNG3	16	24372 930	SNP	C	T	T	Oncomine	NM_006539	p.R232W	p.R232	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	CACNG3	16	24373 168	SNP	G	C	C	Oncomine	NM_006539	p.R311P	p.R311	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	CACNG3	16	24373 168	SNP	G	A	A	Oncomine	NM_006539	p.R311H	p.R311	Missense_Mutation	Hotspot
Cutaneous Melanoma	CACNG3	16	24372 930	SNP	C	T	T	Oncomine	NM_006539	p.R232W	p.R232	Missense_Mutation	Hotspot
Cutaneous Melanoma	CACNG3	16	24372 868	SNP	C	T	T	Oncomine	NM_006539	p.S211F	p.S211	Missense_Mutation	Hotspot
Cutaneous Melanoma	CCDC61	19	46498 687	SNP	G	A	A	Oncomine	NM_001080402	p.E29K	p.E29	Missense_Mutation	Hotspot
Cutaneous Melanoma	CCDC61	19	46498 700	SNP	C	T	T	Oncomine	NM_001080402	p.S33F	p.S33	Missense_Mutation	Hotspot
Cutaneous Melanoma	CCDC61	19	46498 687	SNP	G	A	A	Oncomine	NM_001080402	p.E29K	p.E29	Missense_Mutation	Hotspot
Prostate Carcinoma	CDC27	17	45234 367	SNP	A	A	T	Oncomine	NM_001256	p.S252T	p.S252	Missense_Mutation	Hotspot
Cutaneous Melanoma	CDC27	17	45234 366	SNP	G	A	A	Oncomine	NM_001256	p.S252F	p.S252	Missense_Mutation	Hotspot
Chromophobe Renal Cell Carcinoma	CDC27	17	45234 367	SNP	A	A	T	Oncomine	NM_001256	p.S252T	p.S252	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cutaneous Melanoma	CNTN5	11	10016 9975	SNP	G	A	A	Oncomine	NM_01436 1	p.E823K	p.E823	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	10017 0080	SNP	G	A	A	Oncomine	NM_01436 1	p.G858R	p.G858	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	99932 099	SNP	C	T	T	Oncomine	NM_01436 1	p.S379F	p.S379	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	99715 827	SNP	G	A	A	Oncomine	NM_01436 1	p.R137Q	p.R137	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	CNTN5	11	99221 037	SNP	G	T	T	Oncomine	NM_01436 1	p.R137L	p.R137	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	CNTN5	11	99221 037	SNP	G	A	A	Oncomine	NM_01436 1	p.R137Q	p.R137	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	99690 287	SNP	C	T	T	Oncomine	NM_01436 1	p.S23F	p.S23	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	10016 9975	SNP	G	A	A	Oncomine	NM_01436 1	p.E823K	p.E823	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	99932 099	SNP	C	T	T	Oncomine	NM_01436 1	p.S379F	p.S379	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	10017 0081	SNP	G	A	A	Oncomine	NM_01436 1	p.G858E	p.G858	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	99715 827	SNP	G	A	A	Oncomine	NM_01436 1	p.R137Q	p.R137	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	10012 6527	SNP	G	A	A	Oncomine	NM_01436 1	p.E681K	p.E681	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	11	10017 0080	SNP	G	A	A	Oncomine	NM_01436 1	p.G858R	p.G858	Missense_Mutation	Hotspot
Cutaneous Melanoma	CNTN5	2	21900 0407	SNP	G	C	C	Oncomine	NM_00155 7	p.A295P	p.A295	Missense_Mutation	Hotspot
Astrocytoma	CXCR2	2	21900 0407	SNP	G	C	C	Oncomine	NM_00155 7	p.A295P	p.A295	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	CXCR2	2	21899 9763	SNP	G	G	A	Oncomine	NM_00155 7	p.R80H	p.R80	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	CXCR2	2	21899 9763	SNP	G	A	A	Oncomine	NM_00155 7	p.R80H	p.R80	Missense_Mutation	Hotspot
Cutaneous Melanoma	CXCR2	2	21900 0488	SNP	C	T	T	Oncomine	NM_00155 7	p.R322C	p.R322	Missense_Mutation	Hotspot
Cutaneous Melanoma	CXCR2	2	21900 0408	SNP	C	T	T	Oncomine	NM_00155 7	p.A295V	p.A295	Missense_Mutation	Hotspot
Cutaneous Melanoma	DCD	12	55039 462	SNP	C	T	T	Oncomine	NM_05328 3	p.E43K	p.E43	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cutaneous Melanoma	DCD	12	55039 462	SNP	C	T	T	Oncomine	NM_053283	p.E43K	p.E43	Missense_Mutation	Hotspot
Cutaneous Melanoma	DSCR6	21	38390 367	SNP	G	A	A	Oncomine	NM_018962	p.E145K	p.E145	Missense_Mutation	Hotspot
Lung Adenocarcinoma	DUX4L2	10	13549 1125	SNP	G	A	A	Oncomine	NM_001127386	p.A246T	p.A246	Missense_Mutation	Hotspot
Lung Adenocarcinoma	DUX4L2	10	13549 1123	SNP	G	A	A	Oncomine	NM_001127386	p.G245D	p.G245	Missense_Mutation	Hotspot
Infiltrating Bladder Urothelial Carcinoma	DUX4L2	10	13549 1113	SNP	G	T	T	Oncomine	NM_001127386	p.A242S	p.A242	Missense_Mutation	Hotspot
Glioblastoma	DUX4L2	10	13549 1113	SNP	G	A	A	Oncomine	NM_001127386	p.A242T	p.A242	Missense_Mutation	Hotspot
Glioblastoma	DUX4L2	10	13549 1125	SNP	G	A	A	Oncomine	NM_001127386	p.A246T	p.A246	Missense_Mutation	Hotspot
Glioblastoma	DUX4L2	10	13549 1123	SNP	G	A	A	Oncomine	NM_001127386	p.G245D	p.G245	Missense_Mutation	Hotspot
Astrocytoma Head and Neck Squamous Cell Carcinoma	DUX4L2	10	13549 1112	SNP	C	A	A	Oncomine	NM_001127386	p.F241L	p.F241	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	DUX4L2	10	13549 1125	SNP	G	A	A	Oncomine	NM_001127386	p.A246T	p.A246	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	DUX4L2	10	13549 1123	SNP	G	A	A	Oncomine	NM_001127386	p.G245D	p.G245	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	DUX4L2	10	13549 1112	SNP	C	A	A	Oncomine	NM_001127386	p.F241L	p.F241	Missense_Mutation	Hotspot
Cutaneous Melanoma	DUX4L2	10	13549 1107	SNP	G	A	A	Oncomine	NM_001127386	p.A240T	p.A240	Missense_Mutation	Hotspot
Cutaneous Melanoma	DUX4L2	10	13549 1123	SNP	G	A	A	Oncomine	NM_001127386	p.G245D	p.G245	Missense_Mutation	Hotspot
Cutaneous Melanoma	DUX4L2	10	13549 1125	SNP	G	A	A	Oncomine	NM_001127386	p.A246T	p.A246	Missense_Mutation	Hotspot
Cutaneous Melanoma	DUX4L2	10	13549 1113	SNP	G	A	A	Oncomine	NM_001127386	p.A242T	p.A242	Missense_Mutation	Hotspot
Cutaneous Melanoma	DUX4L2	10	13549 1112	SNP	C	A	A	Oncomine	NM_001127386	p.F241L	p.F241	Missense_Mutation	Hotspot
Papillary Renal Cell Carcinoma	DUX4L2	10	13549 1112	SNP	C	A	A	Oncomine	NM_001127386	p.F241L	p.F241	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Thyroid Gland Papillary Carcinoma	DUX4L2	10	13549	SNP	G	A	A	Oncomine	NM_00112	p.A246T	p.A246	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	DUX4L2	10	13549	SNP	G	A	A	Oncomine	NM_00112	p.A240T	p.A240	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	DUX4L2	10	1113	SNP	G	A	A	Oncomine	NM_00112	p.A242T	p.A242	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	DUX4L2	10	13549	SNP	G	A	A	Oncomine	NM_00112	p.G245D	p.G245	Missense_Mutation	Hotspot
Cutaneous Melanoma	EDDM3A	14	21216	SNP	G	A	A	Oncomine	NM_00668	p.R88Q	p.R88	Missense_Mutation	Hotspot
Glioblastoma	EDDM3A	14	21216	SNP	G	A	A	Oncomine	NM_00668	p.R88Q	p.R88	Missense_Mutation	Hotspot
Colorectal Mucinous Adenocarcinoma	EDDM3A	14	20285	SNP	G	G	A	Oncomine	NM_00668	p.R88Q	p.R88	Missense_Mutation	Hotspot
Cutaneous Melanoma	EDDM3A	14	21216	SNP	G	A	A	Oncomine	NM_00668	p.R88Q	p.R88	Missense_Mutation	Hotspot
Ductal Breast Carcinoma	ENDOU	12	48110	SNP	G	G	A	Oncomine	NM_00602	p.P130L	p.P130	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	ENDOU	12	48110	SNP	G	G	A	Oncomine	NM_00602	p.P130L	p.P130	Missense_Mutation	Hotspot
Cutaneous Melanoma	ENDOU	12	48110	SNP	G	C	C	Oncomine	NM_00602	p.P130A	p.P130	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	ERAS	X	48572	SNP	C	T	T	Oncomine	NM_18153	p.A97V	p.A97	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	ERAS	X	48687	SNP	G	G	A	Oncomine	NM_18153	p.A97T	p.A97	Missense_Mutation	Hotspot
Lung Adenocarcinoma	ERAS	X	48687	SNP	G	A	A	Oncomine	NM_18153	p.A97T	p.A97	Missense_Mutation	Hotspot
Cutaneous Melanoma	FABP1	2	88425	SNP	C	T	T	Oncomine	NM_00144	p.E62K	p.E62	Missense_Mutation	Hotspot
Cutaneous Melanoma	FABP1	2	88425	SNP	C	T	T	Oncomine	NM_00144	p.E62K	p.E62	Missense_Mutation	Hotspot
Medulloblastoma	FAM22F	9	97080	DEL	AGA	*	*	Oncomine	NM_01756	p.S691_in_frame_del	p.S691_in_frame_del	In_Frame_Deletion	Hotspot
Cervical Squamous Cell Carcinoma	FAM22F	9	97082	SNP	C	G	G	Oncomine	NM_01756	p.K355N	p.K355	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Colorectal Adenocarcinoma	FAM22F	9	96122 614	SNP	C	G	G	Oncomine	NM_017561	p.K355N	p.K355	Missense_Mutation	Hotspot
Cutaneous Melanoma	FAM22F	9	97080 945	DEL	AGA	-	-	Oncomine	NM_017561	p.S691_inframe del	p.S691_inframe del	In_Frame_Deletion	Hotspot
Prostate Adenocarcinoma	FAM22F	9	97080 945	DEL	AGA	-	-	Oncomine	NM_017561	p.S691_inframe del	p.S691_inframe del	In_Frame_Deletion	Hotspot
Thyroid Gland Carcinoma, NOS	FAM22F	9	97080 945	DEL	AGA	-	-	Oncomine	NM_017561	p.S691_inframe del	p.S691_inframe del	In_Frame_Deletion	Hotspot
Ductal Breast Carcinoma	FBXW8	12	11746 5850	SNP	G	G	A	Oncomine	NM_012174	p.R491H	p.R491	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	FBXW8	12	11595 0233	SNP	G	A	A	Oncomine	NM_012174	p.R491H	p.R491	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	FBXW8	12	11746 5849	SNP	C	T	T	Oncomine	NM_012174	p.R491C	p.R491	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	FBXW8	12	11746 5849	SNP	C	T	T	Oncomine	NM_012174	p.R491C	p.R491	Missense_Mutation	Hotspot
Cutaneous Melanoma	FBXW8	12	11746 5849	SNP	C	T	T	Oncomine	NM_012174	p.R491C	p.R491	Missense_Mutation	Hotspot
Glioblastoma	FHL3	1	38463 709	SNP	G	A	A	Oncomine	NM_004468	p.P143S	p.P143	Missense_Mutation	Hotspot
Lung Adenocarcinoma	FHL3	1	38463 709	SNP	G	A	A	Oncomine	NM_004468	p.P143S	p.P143	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	FHL3	1	38463 709	SNP	G	C	C	Oncomine	NM_004468	p.P143A	p.P143	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	GGT1	22	23340 828	SNP	G	A	A	Oncomine	NM_005265	p.G84S	p.G84	Missense_Mutation	Hotspot
Cutaneous Melanoma	GK2	4	80327 859	SNP	C	G	G	Oncomine	NM_033214	p.R499P	p.R499	Missense_Mutation	Hotspot
Cutaneous Melanoma	GK2	4	80328 367	SNP	G	A	A	Oncomine	NM_033214	p.R330C	p.R330	Missense_Mutation	Hotspot
Cutaneous Melanoma	GK2	4	80327 860	SNP	G	A	A	Oncomine	NM_033214	p.R499C	p.R499	Missense_Mutation	Hotspot
Lung Adenocarcinoma	GK2	4	80328 367	SNP	G	A	A	Oncomine	NM_033214	p.R330C	p.R330	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Glioblastoma	GK2	4	80328 891	SNP	C	A	A	Oncomine	NM_03321 4	p.R155L	p.R155	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	GK2	4	80547 121	SNP	G	A	A	Oncomine	NM_03321 4	p.R420C	p.R420	Missense_Mutation	Hotspot
Endometrial Adenocarcinoma	GK2	4	80328 892	SNP	G	G	A	Oncomine	NM_03321 4	p.R155C	p.R155	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	GK2	4	80327 860	SNP	G	A	A	Oncomine	NM_03321 4	p.R499C	p.R499	Missense_Mutation	Hotspot
Lung Adenocarcinoma	GK2	4	80328 679	SNP	G	A	A	Oncomine	NM_03321 4	p.P226S	p.P226	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	GK2	4	80328 678	SNP	G	A	A	Oncomine	NM_03321 4	p.P226L	p.P226	Missense_Mutation	Hotspot
Cutaneous Melanoma	GK2	4	80328 892	SNP	G	A	A	Oncomine	NM_03321 4	p.R155C	p.R155	Missense_Mutation	Hotspot
Cutaneous Melanoma	GK2	4	80328 367	SNP	G	A	A	Oncomine	NM_03321 4	p.R330C	p.R330	Missense_Mutation	Hotspot
Cutaneous Melanoma	GK2	4	80327 860	SNP	G	A	A	Oncomine	NM_03321 4	p.R499C	p.R499	Missense_Mutation	Hotspot
Cutaneous Melanoma	GK2	4	80328 097	SNP	G	A	A	Oncomine	NM_03321 4	p.R420C	p.R420	Missense_Mutation	Hotspot
Cutaneous Melanoma	GK2	4	80328 679	SNP	G	A	A	Oncomine	NM_03321 4	p.P226S	p.P226	Missense_Mutation	Hotspot
Glioblastoma	GOLGA6 L10	15	83014 132	SNP	C	G	G	Oncomine	NM_00116 4465	p.E151Q	p.E151	Missense_Mutation	Hotspot
Ductal Breast Carcinoma	GOLGA6 L10	15	83014 132	SNP	C	C	G	Oncomine	NM_00116 4465	p.E151Q	p.E151	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	GOLGA6 L10	15	83014 132	SNP	C	G	G	Oncomine	NM_00116 4465	p.E151Q	p.E151	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	GOLGA6 L10	15	83014 132	SNP	C	G	G	Oncomine	NM_00116 4465	p.E151Q	p.E151	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	GOLGA6 L10	15	83014 132	SNP	C	G	G	Oncomine	NM_00116 4465	p.E151Q	p.E151	Missense_Mutation	Hotspot
Cutaneous Melanoma	GPX7	1	53072 530	SNP	C	T	T	Oncomine	NM_01569 6	p.R105C	p.R105	Missense_Mutation	Hotspot
Head and Neck Squamous Cell	GPX7	1	53072 531	SNP	G	T	T	Oncomine	NM_01569 6	p.R105L	p.R105	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Carcinoma													
Lung Adenocarcinoma	GPX7	1	53072	SNP	G	A	A	Oncomine	NM_015696	p.R105H	p.R105	Missense_Mutation	Hotspot
Cutaneous Melanoma	GTSF1	12	54858	SNP	G	A	A	Oncomine	NM_144594	p.P31S	p.P31	Missense_Mutation	Hotspot
Cutaneous Melanoma	GTSF1	12	54858	SNP	G	A	A	Oncomine	NM_144594	p.P31S	p.P31	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	H3F3A	1	22625	SNP	C	T	T	Oncomine	NM_002107	p.R3C	p.R3	Missense_Mutation	Hotspot
Astrocytoma	H3F3A	1	22625	SNP	C	T	T	Oncomine	NM_002107	p.R3C	p.R3	Missense_Mutation	Hotspot
Cervical Squamous Cell Carcinoma	H3F3A	1	22625	SNP	C	T	T	Oncomine	NM_002107	p.R3C	p.R3	Missense_Mutation	Hotspot
Small Cell Lung Carcinoma	HDDC2	6	12566	SNP	C	G	G	Oncomine	NM_016063	p.R101P	p.R101	Missense_Mutation	Hotspot
Small Cell Lung Carcinoma	HDDC2	6	12561	SNP	C	G	G	Oncomine	NM_016063	p.R101P	p.R101	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	HDDC2	6	9867	SNP	C	T	T	Oncomine	NM_016063	p.R101Q	p.R101	Missense_Mutation	Hotspot
Cutaneous Melanoma	HEXDC	17	80400	SNP	A	C	C	Oncomine	NM_173620	p.T482P	p.T482	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	HEXDC	17	80400	SNP	A	C	C	Oncomine	NM_173620	p.T482P	p.T482	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	HEXDC	17	80400	SNP	A	C	C	Oncomine	NM_173620	p.T482P	p.T482	Missense_Mutation	Hotspot
Small Cell Lung Carcinoma	HIST1H4C	6	26212	SNP	G	C	C	Oncomine	NM_003542	p.R68P	p.R68	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	HIST1H4C	6	378	SNP	G	C	C	Oncomine	NM_003542	p.R68P	p.R68	Missense_Mutation	Hotspot
Cutaneous Melanoma	HNRNPC	1	12907	SNP	C	T	T	Oncomine	NM_001013	p.D58N	p.D58	Missense_Mutation	Hotspot
Melanoma	HNRNPC	1	12907	SNP	C	T	T	Oncomine	NM_001013	p.R99Q	p.R99	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	HNRNPC	1	12830	SNP	G	A	A	Oncomine	NM_001013	p.R167W	p.R167	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Endometrial Adenocarcinoma	HNRNPC L1	1	12907 644	SNP	G	G	A	Oncomine	NM_001013631	p.R167W	p.R167	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	HNRNPC L1	1	12907 847	SNP	C	T	T	Oncomine	NM_001013631	p.R99Q	p.R99	Missense_Mutation	Hotspot
Lung Adenocarcinoma	HNRNPC L1	1	12907 643	SNP	C	A	A	Oncomine	NM_001013631	p.R167L	p.R167	Missense_Mutation	Hotspot
Cutaneous Melanoma	HNRNPC L1	1	12907 847	SNP	C	T	T	Oncomine	NM_001013631	p.R99Q	p.R99	Missense_Mutation	Hotspot
Cutaneous Melanoma	HNRNPC L1	1	12907 865	SNP	C	T	T	Oncomine	NM_001013631	p.G93E	p.G93	Missense_Mutation	Hotspot
Cutaneous Melanoma	HNRNPC L1	1	12907 971	SNP	C	T	T	Oncomine	NM_001013631	p.D58N	p.D58	Missense_Mutation	Hotspot
Prostate Adenocarcinoma	HRCTI	9	35906 559	SNP	A	C	C	Oncomine	NM_001039792	p.H92P	p.H92	Missense_Mutation	Hotspot
Glioblastoma	HRCTI	9	35906 348	DEL	CTG	-	-	Oncomine	NM_001039792	p.L22_in_frame_del	p.L22_in_frame_del	In_Frame_Deletion	Hotspot
Ductal Breast Carcinoma	HRCTI	9	35906 348	DEL	CTG	CTG	-	Oncomine	NM_001039792	p.L22_in_frame_del	p.L22_in_frame_del	In_Frame_Deletion	Hotspot
Cervical Squamous Cell Carcinoma	HRCTI	9	35906 559	SNP	A	C	C	Oncomine	NM_001039792	p.H92P	p.H92	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	HRCTI	9	35906 584	DEL	CCA	-	-	Oncomine	NM_001039792	p.L100_in_frame_del	p.L100_in_frame_del	In_Frame_Deletion	Hotspot
Cutaneous Melanoma	HRCTI	9	35906 348	DEL	CTG	-	-	Oncomine	NM_001039792	p.L22_in_frame_del	p.L22_in_frame_del	In_Frame_Deletion	Hotspot
Cutaneous Melanoma	HRCTI	9	35906 559	SNP	A	C	C	Oncomine	NM_001039792	p.H92P	p.H92	Missense_Mutation	Hotspot
Papillary Renal Cell Carcinoma	HRCTI	9	35906 584	DEL	CCA	-	-	Oncomine	NM_001039792	p.L100_in_frame_del	p.L100_in_frame_del	In_Frame_Deletion	Hotspot
Papillary Renal Cell Carcinoma	HRCTI	9	35906 559	SNP	A	C	C	Oncomine	NM_001039792	p.H92P	p.H92	Missense_Mutation	Hotspot
Thyroid Gland Carcinoma, NOS	HRCTI	9	35906 584	DEL	CCA	-	-	Oncomine	NM_001039792	p.L100_in_frame_del	p.L100_in_frame_del	In_Frame_Deletion	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Colorectal Adenocarcinoma	IL3	5	13142	SNP	G	A	A	Oncomine	NM_000588	p.A90T	p.A90	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	IL3	5	13139	SNP	G	A	A	Oncomine	NM_000588	p.A90T	p.A90	Missense_Mutation	Hotspot
Pancreatic Ductal Adenocarcinoma	JAM3	11	13401	SNP	G	A	G	Oncomine	NM_032801	p.R191H	p.R191	Missense_Mutation	Hotspot
Lobular Breast Carcinoma	JAM3	11	13401	SNP	G	G	A	Oncomine	NM_032801	p.R191H	p.R191	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	JAM3	11	13401	SNP	C	T	T	Oncomine	NM_032801	p.R191C	p.R191	Missense_Mutation	Hotspot
Cutaneous Melanoma	KCNK9	8	14063	SNP	C	T	T	Oncomine	NM_016601	p.D104N	p.D104	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	KCNK9	8	14063	SNP	C	C	T	Oncomine	NM_016601	p.A265T	p.A265	Missense_Mutation	Hotspot
Lung Adenocarcinoma	KCNK9	8	14063	SNP	G	A	A	Oncomine	NM_016601	p.A265V	p.A265	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	KCNK9	8	14063	SNP	C	T	T	Oncomine	NM_016601	p.A265T	p.A265	Missense_Mutation	Hotspot
Cutaneous Melanoma	KCNK9	8	14063	SNP	C	T	T	Oncomine	NM_016601	p.D104N	p.D104	Missense_Mutation	Hotspot
Glioblastoma	KLK6	19	51466	SNP	C	T	T	Oncomine	NM_002774	p.R111H	p.R111	Missense_Mutation	Hotspot
Colorectal Mucinous Adenocarcinoma	KLK6	19	56158	SNP	G	A	A	Oncomine	NM_002774	p.R111C	p.R111	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	KLK6	19	51466	SNP	C	C	T	Oncomine	NM_002774	p.R111H	p.R111	Missense_Mutation	Hotspot
Cutaneous Melanoma	KLK6	19	51462	SNP	G	A	A	Oncomine	NM_002774	p.P200L	p.P200	Missense_Mutation	Hotspot
Prostate Adenocarcinoma	KLK6	19	51462	SNP	G	A	A	Oncomine	NM_002774	p.P200L	p.P200	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	KLK6	19	51462	SNP	G	A	A	Oncomine	NM_002774	p.P200L	p.P200	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	KRTAP12-4	21	44898	SNP	T	G	G	Oncomine	NM_198698	p.T4P	p.T4	Missense_Mutation	Hotspot
Ovarian Serous Adenocarcinoma	KRTAP12-4	21	44898	SNP	G	G	A	Oncomine	NM_198698	p.T4I	p.T4	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cutaneous Melanoma	KRTAP4-11	17	39274 424	SNP	G	C	C	Oncomine	NM_033059	p.S48R	p.S48	Missense_Mutation	Hotspot
Cutaneous Melanoma	KRTAP4-11	17	39274 206	SNP	C	T	T	Oncomine	NM_033059	p.R121K	p.R121	Missense_Mutation	Hotspot
Lung Adenocarcinoma	KRTAP4-11	17	39274 150	SNP	T	A	A	Oncomine	NM_033059	p.S140C	p.S140	Missense_Mutation	Hotspot
Lung Adenocarcinoma	KRTAP4-11	17	39274 206	SNP	C	T	T	Oncomine	NM_033059	p.R121K	p.R121	Missense_Mutation	Hotspot
Glioblastoma	KRTAP4-11	17	39274 424	SNP	G	C	C	Oncomine	NM_033059	p.S48R	p.S48	Missense_Mutation	Hotspot
Oligodendroglioma	KRTAP4-11	17	39274 087	SNP	G	C	C	Oncomine	NM_033059	p.L161V	p.L161	Missense_Mutation	Hotspot
Ductal Breast Carcinoma	KRTAP4-11	17	39274 087	SNP	G	G	C	Oncomine	NM_033059	p.L161V	p.L161	Missense_Mutation	Hotspot
Cervical Squamous Cell Carcinoma	KRTAP4-11	17	39274 424	SNP	G	C	C	Oncomine	NM_033059	p.S48R	p.S48	Missense_Mutation	Hotspot
Cervical Squamous Cell Carcinoma	KRTAP4-11	17	39274 087	SNP	G	C	C	Oncomine	NM_033059	p.L161V	p.L161	Missense_Mutation	Hotspot
Cervical Squamous Cell Carcinoma	KRTAP4-11	17	39274 150	SNP	T	A	A	Oncomine	NM_033059	p.S140C	p.S140	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	KRTAP4-11	17	39274 424	SNP	G	C	C	Oncomine	NM_033059	p.S48R	p.S48	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	KRTAP4-11	17	39274 206	SNP	C	T	T	Oncomine	NM_033059	p.R121K	p.R121	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	KRTAP4-11	17	39274 087	SNP	G	C	C	Oncomine	NM_033059	p.L161V	p.L161	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	KRTAP4-11	17	39274 150	SNP	T	A	A	Oncomine	NM_033059	p.S140C	p.S140	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	KRTAP4-11	17	39274 291	SNP	T	C	C	Oncomine	NM_033059	p.M93V	p.M93	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	KRTAP4-11	17	39274 416	SNP	C	T	T	Oncomine	NM_033059	p.R51K	p.R51	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Lung Adenocarcinoma	KRTAP4-11	17	39274 424	SNP	G	C	C	Oncomine	NM_033059	p.S48R	p.S48	Missense_Mutation	Hotspot
Lung Adenocarcinoma	KRTAP4-11	17	39274 206	SNP	C	T	T	Oncomine	NM_033059	p.R121K	p.R121	Missense_Mutation	Hotspot
Cutaneous Melanoma	KRTAP4-11	17	39274 150	SNP	T	A	A	Oncomine	NM_033059	p.S140C	p.S140	Missense_Mutation	Hotspot
Cutaneous Melanoma	KRTAP4-11	17	39274 206	SNP	C	T	T	Oncomine	NM_033059	p.R121K	p.R121	Missense_Mutation	Hotspot
Cutaneous Melanoma	KRTAP4-11	17	39274 424	SNP	G	C	C	Oncomine	NM_033059	p.S48R	p.S48	Missense_Mutation	Hotspot
Cutaneous Melanoma	KRTAP4-11	17	39274 087	SNP	G	C	C	Oncomine	NM_033059	p.L161V	p.L161	Missense_Mutation	Hotspot
Cutaneous Melanoma	KRTAP4-11	17	39274 416	SNP	C	T	T	Oncomine	NM_033059	p.R51K	p.R51	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	KRTAP4-11	17	39274 291	SNP	T	C	C	Oncomine	NM_033059	p.M93V	p.M93	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	KRTAP4-11	17	39274 206	SNP	C	T	T	Oncomine	NM_033059	p.R121K	p.R121	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	KRTAP4-11	17	39274 150	SNP	T	A	A	Oncomine	NM_033059	p.S140C	p.S140	Missense_Mutation	Hotspot
Papillary Renal Cell Carcinoma	KRTAP4-11	17	39274 087	SNP	G	C	C	Oncomine	NM_033059	p.L161V	p.L161	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	KRTAP4-11	17	39274 424	SNP	G	C	C	Oncomine	NM_033059	p.S48R	p.S48	Missense_Mutation	Hotspot
Papillary Renal Cell Carcinoma	KRTAP4-7	17	39240 900	SNP	T	G	G	Oncomine	NM_033061	p.L148V	p.L148	Missense_Mutation	Hotspot
Cutaneous Melanoma	LAD1	1	20135 4881	SNP	C	T	T	Oncomine	NM_005558	p.R360Q	p.R360	Missense_Mutation	Hotspot
Cutaneous Melanoma	LAD1	1	20135 2246	SNP	C	T	T	Oncomine	NM_005558	p.E448K	p.E448	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	LAD1	1	20135 4881	SNP	C	A	A	Oncomine	NM_005558	p.R360L	p.R360	Missense_Mutation	Hotspot
Melanoma	LELP1	1	15317 7244	SNP	C	T	T	Oncomine	NM_001010857	p.P21S	p.P21	Missense_Mutation	Hotspot
Cutaneous Melanoma	LELP1	1	15317 7437	SNP	C	T	T	Oncomine	NM_001010857	p.S85F	p.S85	Missense_Mutation	Hotspot
Cutaneous Melanoma	LELP1	1	15317 7245	SNP	C	T	T	Oncomine	NM_001010857	p.P21L	p.P21	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cutaneous Melanoma	LELP1	1	15317 7244	SNP	C	T	T	Oncomine	NM_00101 0857	p.P21S	p.P21	Missense_Mutation	Hotspot
Cutaneous Melanoma	LOC100509575	X	47972	SNP	G	A	A	Oncomine	NM_00120 5103	p.R96H	p.R96	Missense_Mutation	Hotspot
Lobular Breast Carcinoma	LOC100509575	X	47972 582	SNP	G	G	A	Oncomine	NM_00120 5103	p.R96H	p.R96	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	LOC100509575	X	47972	SNP	C	C	T	Oncomine	NM_00120 5103	p.R96C	p.R96	Missense_Mutation	Hotspot
Glioblastoma	MUC4	3	19551 6064	SNP	C	T	T	Oncomine	NM_01840 6	p.R796Q	p.R796	Missense_Mutation	Hotspot
Ductal Breast Carcinoma	MUC4	3	19551 6064	SNP	C	C	T	Oncomine	NM_01840 6	p.R796Q	p.R796	Missense_Mutation	Hotspot
Lung Adenocarcinoma	MUC4	3	19551 6064	SNP	C	T	T	Oncomine	NM_00596 7	p.F208L	p.F208	Missense_Mutation	Hotspot
Glioblastoma	NAB2	12	57485 446	SNP	T	C	C	Oncomine	NM_00596 7	p.F208L	p.F208	Missense_Mutation	Hotspot
Oligodendroglioma	NAB2	12	57485 446	SNP	T	C	C	Oncomine	NM_00596 7	p.F208L	p.F208	Missense_Mutation	Hotspot
Lung Adenocarcinoma	NAB2	12	57485 446	SNP	T	C	C	Oncomine	NM_00596 7	p.F208L	p.F208	Missense_Mutation	Hotspot
Cutaneous Melanoma	NAB2	12	57485 446	SNP	T	C	C	Oncomine	NM_00596 7	p.F208L	p.F208	Missense_Mutation	Hotspot
Glioblastoma	NBPF10	1	14532 4371	SNP	T	C	C	Oncomine	NM_00103 9703	p.V1189A	p.V1189	Missense_Mutation	Hotspot
Astrocytoma	NBPF10	1	14536 0584	SNP	G	A	A	Oncomine	NM_00103 9703	p.G3070E	p.G3070	Missense_Mutation	Hotspot
Cutaneous Melanoma	NBPF10	1	14536 0584	SNP	G	A	A	Oncomine	NM_00103 9703	p.G3070E	p.G3070	Missense_Mutation	Hotspot
Cutaneous Melanoma	NSFL1C	20	14263 60	SNP	G	A	A	Oncomine	NM_01614 3	p.R301W	p.R301	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	NSFL1C	20	13743 60	SNP	G	A	A	Oncomine	NM_01614 3	p.R301W	p.R301	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	NSFL1C	20	14263 60	SNP	G	G	A	Oncomine	NM_01614 3	p.R301W	p.R301	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Head and Neck Squamous Cell Carcinoma	NSFL1C	20	1426360	SNP	G	A	A	Oncomine	NM_016143	p.R301W	p.R301	Missense_Mutation	Hotspot
Medulloblastoma	OBP2B	9	136081795	SNP	A	G	G	Oncomine	NM_014581	p.S133P	p.S133	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	OBP2B	9	136081795	SNP	A	G	G	Oncomine	NM_014581	p.S133P	p.S133	Missense_Mutation	Hotspot
Lung Adenocarcinoma	OBP2B	9	136081795	SNP	A	G	G	Oncomine	NM_014581	p.S133P	p.S133	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2L13	1	248262729	SNP	C	A	A	Oncomine	NM_175911	p.P18T	p.P18	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2L13	1	248263173	SNP	C	T	T	Oncomine	NM_175911	p.P166S	p.P166	Missense_Mutation	Hotspot
Lung Adenocarcinoma	OR2L13	1	248263401	SNP	A	G	G	Oncomine	NM_175911	p.T242A	p.T242	Missense_Mutation	Hotspot
Small Cell Lung Carcinoma	OR2L13	1	248262832	SNP	C	A	A	Oncomine	NM_175911	p.P52H	p.P52	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	OR2L13	1	248262831	SNP	C	T	T	Oncomine	NM_175911	p.P52S	p.P52	Missense_Mutation	Hotspot
Lung Adenocarcinoma	OR2L13	1	248263401	SNP	A	T	T	Oncomine	NM_175911	p.T242S	p.T242	Missense_Mutation	Hotspot
Lung Adenocarcinoma	OR2L13	1	248263401	SNP	A	G	G	Oncomine	NM_175911	p.T242A	p.T242	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	OR2L13	1	248262831	SNP	C	T	T	Oncomine	NM_175911	p.P52S	p.P52	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2L13	1	248263371	SNP	G	A	A	Oncomine	NM_175911	p.G232R	p.G232	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2L13	1	248263174	SNP	C	T	T	Oncomine	NM_175911	p.P166L	p.P166	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2L13	1	248263173	SNP	C	T	T	Oncomine	NM_175911	p.P166S	p.P166	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2L13	1	248262826	SNP	C	T	T	Oncomine	NM_175911	p.P18L	p.P18	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2L13	1	248262730	SNP	C	A	A	Oncomine	NM_175911	p.P18T	p.P18	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2L13	1	248262729	SNP	C	A	A	Oncomine	NM_175911	p.P18T	p.P18	Missense_Mutation	Hotspot
Ovarian Serous Adenocarcinoma	OR2L13	1	246329995	SNP	G	G	A	Oncomine	NM_175911	p.G232E	p.G232	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Colorectal Adenocarcinoma	OR2T27	1	24688	SNP	C	T	T	Oncomine	NM_001001824	p.D11N	p.D11	Missense_Mutation	Hotspot
Endometrial Serous Adenocarcinoma	OR2T27	1	24881	SNP	G	G	A	Oncomine	NM_001001824	p.R122C	p.R122	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	OR2T27	1	24881	SNP	C	T	T	Oncomine	NM_001001824	p.R122H	p.R122	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	OR2T27	1	24881	SNP	C	T	T	Oncomine	NM_001001824	p.R138H	p.R138	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	OR2T27	1	24881	SNP	C	G	G	Oncomine	NM_001001824	p.R138P	p.R138	Missense_Mutation	Hotspot
Lung Adenocarcinoma	OR2T27	1	24881	SNP	C	A	A	Oncomine	NM_001001824	p.D11Y	p.D11	Missense_Mutation	Hotspot
Lung Adenocarcinoma	OR2T27	1	24881	SNP	C	A	A	Oncomine	NM_001001824	p.R138L	p.R138	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2Z1	19	88414	SNP	C	T	T	Oncomine	NM_001004699	p.S23L	p.S23	Missense_Mutation	Hotspot
Glioblastoma	OR2Z1	19	88418	SNP	C	T	T	Oncomine	NM_001004699	p.R138C	p.R138	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	OR2Z1	19	88418	SNP	C	T	T	Oncomine	NM_001004699	p.R138C	p.R138	Missense_Mutation	Hotspot
Lung Adenocarcinoma	OR2Z1	19	88414	SNP	C	T	T	Oncomine	NM_001004699	p.S23L	p.S23	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR2Z1	19	88418	SNP	C	T	T	Oncomine	NM_001004699	p.P167S	p.P167	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR4E2	14	22133	SNP	G	A	A	Oncomine	NM_001001912	p.G151E	p.G151	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR4E2	14	22133	SNP	G	A	A	Oncomine	NM_001001912	p.R226Q	p.R226	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR4E2	14	22133	SNP	G	A	A	Oncomine	NM_001001912	p.G151R	p.G151	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR4E2	14	22133	SNP	G	A	A	Oncomine	NM_001001912	p.G151E	p.G151	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR51B2	11	53452	SNP	C	T	T	Oncomine	NM_033180	p.E89K	p.E89	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR51B2	11	53450	SNP	G	A	A	Oncomine	NM_033180	p.S163L	p.S163	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Lung Adenocarcinoma	OR51B2	11	53447	SNP	G	T	T	Oncomine	NM_033180	p.T252K	p.T252	Missense_Mutation	Hotspot
Glioblastoma	OR51B2	11	53447	SNP	G	A	A	Oncomine	NM_033180	p.T252I	p.T252	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	OR51B2	11	53447	SNP	T	C	C	Oncomine	NM_033180	p.T252A	p.T252	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	OR51B2	11	53451	SNP	C	T	T	Oncomine	NM_033180	p.G143R	p.G143	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	OR51B2	11	53451	SNP	C	A	A	Oncomine	NM_033180	p.G143V	p.G143	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR51B2	11	53452	SNP	C	T	T	Oncomine	NM_033180	p.E89K	p.E89	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR51B2	11	53450	SNP	G	A	A	Oncomine	NM_033180	p.S163L	p.S163	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR51B2	11	53451	SNP	C	T	T	Oncomine	NM_012375	p.G143E	p.G143	Missense_Mutation	Hotspot
Glioblastoma	OR52A1	11	51726	SNP	C	T	T	Oncomine	NM_012375	p.R303H	p.R303	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR52A1	11	51729	SNP	G	A	A	Oncomine	NM_012375	p.R230C	p.R230	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR52A1	11	51726	SNP	G	A	A	Oncomine	NM_012375	p.R303C	p.R303	Missense_Mutation	Hotspot
Prostate Adenocarcinoma	OR52A1	11	51729	SNP	C	T	T	Oncomine	NM_001005	p.R230H	p.R230	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR5AN1	11	59132	SNP	C	T	T	Oncomine	NM_001004729	p.S218F	p.S218	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR6T1	11	12381	SNP	G	A	A	Oncomine	NM_001005187	p.R122C	p.R122	Missense_Mutation	Hotspot
Ductal Breast Carcinoma	OR6T1	11	12381	SNP	G	G	T	Oncomine	NM_001005187	p.S217Y	p.S217	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	OR6T1	11	12331	SNP	C	C	T	Oncomine	NM_001005187	p.R261H	p.R261	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	OR6T1	11	12331	SNP	G	A	A	Oncomine	NM_001005187	p.R179C	p.R179	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	OR6T1	11	12331	SNP	G	T	T	Oncomine	NM_001005187	p.S217Y	p.S217	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR6T1	11	12381	SNP	G	A	A	Oncomine	NM_001005187	p.R261C	p.R261	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cutaneous Melanoma	OR6T1	11	12381 4011	SNP	G	A	A	Oncomine	NM_00100 5187	p.R179C	p.R179	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR6T1	11	12381 3896	SNP	G	A	A	Oncomine	NM_00100 5187	p.S217F	p.S217	Missense_Mutation	Hotspot
Cutaneous Melanoma	OR6T1	11	12381 4182	SNP	G	A	A	Oncomine	NM_00100 5187	p.R122C	p.R122	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	OR6T1	11	12381 4182	SNP	G	T	T	Oncomine	NM_00100 5187	p.R122S	p.R122	Missense_Mutation	Hotspot
Cutaneous Melanoma	OTUD5	X	48792 073	SNP	C	T	T	Oncomine	NM_01760 2	p.R274Q	p.R274	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	OTUD5	X	48668 111	SNP	G	A	A	Oncomine	NM_01760 2	p.R412W	p.R412	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	OTUD5	X	48677 018	SNP	G	A	A	Oncomine	NM_01760 2	p.R274W	p.R274	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	OTUD5	X	48792 074	SNP	G	G	A	Oncomine	NM_01760 2	p.R274W	p.R274	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	OTUD5	X	48783 167	SNP	G	G	A	Oncomine	NM_01760 2	p.R412W	p.R412	Missense_Mutation	Hotspot
Lung Adenocarcinoma	OTUD5	X	48783 166	SNP	C	A	A	Oncomine	NM_01760 2	p.R412L	p.R412	Missense_Mutation	Hotspot
Cutaneous Melanoma	OXA1L	14	23235 902	SNP	C	T	T	Oncomine	NM_00501 5	p.P58S	p.P58	Missense_Mutation	Hotspot
Cutaneous Melanoma	OXA1L	14	23235 899	SNP	C	T	T	Oncomine	NM_00501 5	p.L57F	p.L57	Missense_Mutation	Hotspot
Cutaneous Melanoma	OXA1L	14	23235 902	SNP	C	T	T	Oncomine	NM_00501 5	p.P58S	p.P58	Missense_Mutation	Hotspot
Cervical Squamous Cell Carcinoma	PBX2	6	32155 509	SNP	T	A	A	Oncomine	NM_00258 6	p.Y262F	p.Y262	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	PBX2	6	32155 509	SNP	T	A	A	Oncomine	NM_00258 6	p.Y262F	p.Y262	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	PBX2	6	32155 509	SNP	T	A	A	Oncomine	NM_00258 6	p.Y262F	p.Y262	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	PBX2	6	32155 509	SNP	T	A	A	Oncomine	NM_00258 6	p.Y262F	p.Y262	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	PBX2	6	32155 509	SNP	T	A	A	Oncomine	NM_00258 6	p.Y262F	p.Y262	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Prostate Adenocarcinoma	PDHA2	4	96761 513	SNP	G	A	A	Oncomine	NM_005390	p.R71H	p.R71	Missense_Mutation	Hotspot
Melanoma	PDHA2	4	96761 738	SNP	G	A	A	Oncomine	NM_005390	p.G146E	p.G146	Missense_Mutation	Hotspot
Cutaneous Melanoma	PDHA2	4	96761 737	SNP	G	A	A	Oncomine	NM_005390	p.G146R	p.G146	Missense_Mutation	Hotspot
Glioblastoma	PDHA2	4	96761 557	SNP	C	T	T	Oncomine	NM_005390	p.R86C	p.R86	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	PDHA2	4	96980 580	SNP	C	T	T	Oncomine	NM_005390	p.R86C	p.R86	Missense_Mutation	Hotspot
Endometrial Serous Adenocarcinoma	PDHA2	4	96761 738	SNP	G	G	A	Oncomine	NM_005390	p.G146E	p.G146	Missense_Mutation	Hotspot
Lung Adenocarcinoma	PDHA2	4	96761 513	SNP	G	A	A	Oncomine	NM_005390	p.R71H	p.R71	Missense_Mutation	Hotspot
Cutaneous Melanoma	PDHA2	4	96761 854	SNP	G	A	A	Oncomine	NM_005390	p.D185N	p.D185	Missense_Mutation	Hotspot
Cutaneous Melanoma	PDHA2	4	96761 738	SNP	G	A	A	Oncomine	NM_005390	p.G146E	p.G146	Missense_Mutation	Hotspot
Thyroid Gland Carcinoma, NOS	PDHA2	4	96761 513	SNP	G	A	A	Oncomine	NM_005390	p.R71H	p.R71	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	PDHA2	4	96761 557	SNP	C	T	T	Oncomine	NM_005390	p.R86C	p.R86	Missense_Mutation	Hotspot
Lung Adenocarcinoma	POTEC	18	14543 019	SNP	T	C	C	Oncomine	NM_001137671	p.M43V	p.M43	Missense_Mutation	Hotspot
Glioblastoma	POTEC	18	14543 019	SNP	T	C	C	Oncomine	NM_001137671	p.M43V	p.M43	Missense_Mutation	Hotspot
Astrocytoma	POTEC	18	14543 734	SNP	C	T	T	Oncomine	NM_001137671	p.G487E	p.G487	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	POTEC	18	14513 734	SNP	C	T	T	Oncomine	NM_001137671	p.G487E	p.G487	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	POTEC	18	14543 019	SNP	T	C	C	Oncomine	NM_001137671	p.M43V	p.M43	Missense_Mutation	Hotspot
Lung Adenocarcinoma	POTEC	18	14513 734	SNP	C	T	T	Oncomine	NM_001137671	p.G487E	p.G487	Missense_Mutation	Hotspot
Cutaneous Melanoma	POTEC	18	14543 019	SNP	T	C	C	Oncomine	NM_001137671	p.M43V	p.M43	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cutaneous Melanoma	POTEC	18	14542791	SNP	C	T	T	Oncomine	NM_001137671	p.A119T	p.A119	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	POTEC	18	14542791	SNP	C	T	T	Oncomine	NM_001137671	p.A119T	p.A119	Missense_Mutation	Hotspot
Glioblastoma	POTEM	14	20010235	SNP	A	G	G	Oncomine	NM_001145442	p.V308A	p.V308	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	POTEM	14	20010235	SNP	A	G	G	Oncomine	NM_001145442	p.V308A	p.V308	Missense_Mutation	Hotspot
Cutaneous Melanoma	POTEM	14	20019948	SNP	C	T	T	Oncomine	NM_001145442	p.M91I	p.M91	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	PPIL1	6	36842542	SNP	C	T	T	Oncomine	NM_016059	p.A3T	p.A3	Missense_Mutation	Hotspot
Ovarian Serous Adenocarcinoma	PPIL1	6	36950519	SNP	G	G	A	Oncomine	NM_016059	p.A3V	p.A3	Missense_Mutation	Hotspot
Papillary Renal Cell Carcinoma	PPIL1	6	36842542	SNP	C	T	T	Oncomine	NM_001099852	p.A3T	p.A3	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRAMEF20	1	13743091	SNP	C	T	T	Oncomine	NM_001099852	p.R94C	p.R94	Missense_Mutation	Hotspot
Glioblastoma	PRAMEF20	1	13743092	SNP	G	A	A	Oncomine	NM_006249	p.R94H	p.R94	Missense_Mutation	Hotspot
Melanoma	PRB3	12	11420548	SNP	C	T	T	Oncomine	NM_006249	p.G212E	p.G212	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRB3	12	11420963	SNP	G	A	A	Oncomine	NM_006249	p.R74C	p.R74	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	PRB3	12	11420963	SNP	G	A	A	Oncomine	NM_006249	p.R74C	p.R74	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRB3	12	11420548	SNP	C	T	T	Oncomine	NM_006249	p.G212E	p.G212	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRB4	12	11461597	SNP	C	T	T	Oncomine	NM_002723	p.G107E	p.G107	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRB4	12	11461475	SNP	C	T	T	Oncomine	NM_002723	p.G148R	p.G148	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRB4	12	11461474	SNP	C	T	T	Oncomine	NM_002723	p.G148E	p.G148	Missense_Mutation	Hotspot
Cutaneous Melanoma	PROL1	4	71275418	SNP	C	T	T	Oncomine	NM_021225	p.P125S	p.P125	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cutaneous Melanoma	PROL1	4	71275 418	SNP	C	T	T	Oncomine	NM_02122 5	p.P125S	p.P125	Missense_Mutation	Hotspot
Cutaneous Melanoma	PROL1	4	71275 428	SNP	C	A	A	Oncomine	NM_02122 5	p.P128H	p.P128	Missense_Mutation	Hotspot
Cutaneous Melanoma	PROL1	4	71275 427	SNP	C	T	T	Oncomine	NM_02122 5	p.P128S	p.P128	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRSS37	7	14153 6973	SNP	C	T	T	Oncomine	NM_00100 8270	p.G169E	p.G169	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRSS37	7	14154 0847	SNP	C	T	T	Oncomine	NM_00100 8270	p.M1I	p.M1	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRSS37	7	14153 6973	SNP	C	T	T	Oncomine	NM_00100 8270	p.G169E	p.G169	Missense_Mutation	Hotspot
Cutaneous Melanoma	PRSS37	7	14154 0847	SNP	C	T	T	Oncomine	NM_00100 8270	p.M1I	p.M1	Missense_Mutation	Hotspot
Cutaneous Melanoma	RAB39A	11	10783 2799	SNP	C	G	G	Oncomine	NM_01751 6	p.R119G	p.R119	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	RAB39A	11	10733 8009	SNP	C	T	T	Oncomine	NM_01751 6	p.R119W	p.R119	Missense_Mutation	Hotspot
Cutaneous Melanoma	RAB39A	11	10783 2799	SNP	C	T	T	Oncomine	NM_01751 6	p.R119W	p.R119	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	RALB	2	12103 6297	SNP	G	A	A	Oncomine	NM_00288 1	p.M19I	p.M19	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	RALB	2	12103 6296	SNP	T	C	C	Oncomine	NM_00288 1	p.M19T	p.M19	Missense_Mutation	Hotspot
Cutaneous Melanoma	RALB	2	12103 6296	SNP	T	A	A	Oncomine	NM_00288 1	p.M19K	p.M19	Missense_Mutation	Hotspot
Medulloblastoma	RANGAP1	22	41652 800	SNP	A	C	C	Oncomine	NM_00288 3	p.V268G	p.V268	Missense_Mutation	Hotspot
Cervical Squamous Cell Carcinoma	RANGAP1	22	41652 800	SNP	A	C	C	Oncomine	NM_00288 3	p.V268G	p.V268	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	RANGAP1	22	41652 800	SNP	A	C	C	Oncomine	NM_00288 3	p.V268G	p.V268	Missense_Mutation	Hotspot
Lung Adenocarcinoma	RANGAP1	22	41652 800	SNP	A	C	C	Oncomine	NM_00288 3	p.V268G	p.V268	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	RANGAP1	22	41652 800	SNP	A	C	C	Oncomine	NM_00288 3	p.V268G	p.V268	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Gastric Adenocarcinoma	RAP1B	12	69042 539	SNP	G	A	A	Oncomine	NM_01564 6	p.G12E	p.G12	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	RAP1B	12	69042 539	SNP	G	A	A	Oncomine	NM_01564 6	p.G12E	p.G12	Missense_Mutation	Hotspot
Acute Myeloid Leukemia	RAP1B	12	67328 806	SNP	G	G	A	Oncomine	NM_01564 6	p.G12E	p.G12	Missense_Mutation	Hotspot
Cutaneous Melanoma	RBM1D	Y	23702 641	SNP	C	T	T	Oncomine	NM_00100 6120	p.P124L	p.P124	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	RBM1D	Y	23702 641	SNP	C	A	A	Oncomine	NM_00100 6120	p.P124H	p.P124	Missense_Mutation	Hotspot
Cutaneous Melanoma	RBM1D	Y	23702 640	SNP	C	T	T	Oncomine	NM_00100 6120	p.P124S	p.P124	Missense_Mutation	Hotspot
Prostate Adenocarcinoma	RQCD1	2	21944 7749	SNP	C	G	G	Oncomine	NM_00544 4	p.S87C	p.S87	Missense_Mutation	Hotspot
Melanoma	RQCD1	2	21944 7749	SNP	C	G	G	Oncomine	NM_00544 4	p.S87C	p.S87	Missense_Mutation	Hotspot
Cutaneous Melanoma	RQCD1	2	21944 9406	SNP	C	T	T	Oncomine	NM_00544 4	p.P131L	p.P131	Missense_Mutation	Hotspot
Cutaneous Melanoma	RQCD1	2	21944 7748	SNP	T	C	C	Oncomine	NM_00544 4	p.S87P	p.S87	Missense_Mutation	Hotspot
Cutaneous Melanoma	S100A7L 2	1	15340 9566	SNP	C	T	T	Oncomine	NM_00104 5479	p.G103R	p.G103	Missense_Mutation	Hotspot
Cutaneous Melanoma	S100A7L 2	1	15340 9566	SNP	C	T	T	Oncomine	NM_00104 5479	p.G103R	p.G103	Missense_Mutation	Hotspot
Cutaneous Melanoma	S100A7L 2	1	15340 9565	SNP	C	T	T	Oncomine	NM_00104 5479	p.G103E	p.G103	Missense_Mutation	Hotspot
Non-Small Cell Lung Carcinoma, NOS	S100A8	1	15336 2715	SNP	T	C	C	Oncomine	NM_00296 4	p.K49R	p.K49	Missense_Mutation	Hotspot
Glioblastoma	S100A8	1	15336 2715	SNP	T	C	C	Oncomine	NM_00296 4	p.K49R	p.K49	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	S100A8	1	15336 2715	SNP	T	C	C	Oncomine	NM_00296 4	p.K49R	p.K49	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	S100A8	1	15336 2715	SNP	T	C	C	Oncomine	NM_00296 4	p.K49R	p.K49	Missense_Mutation	Hotspot
Oligodendroglioma	SAA2	11	18269 491	SNP	G	A	A	Oncomine	NM_03075 4	p.S23L	p.S23	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Lung Adenocarcinoma	SDR16C5	8	57228	SNP	C	A	A	Oncomine	NM_13896_9	p.A94S	p.A94	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	SDR16C5	8	57228	SNP	G	T	T	Oncomine	NM_13896_9	p.A94D	p.A94	Missense_Mutation	Hotspot
Cutaneous Melanoma	SDR16C5	8	57228	SNP	C	T	T	Oncomine	NM_13896_9	p.M26I	p.M26	Missense_Mutation	Hotspot
Cutaneous Melanoma	SDR16C5	8	57228	SNP	G	A	A	Oncomine	NM_13896_9	p.S18L	p.S18	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	SDR16C5	8	57228	SNP	C	G	G	Oncomine	NM_13896_9	p.A94P	p.A94	Missense_Mutation	Hotspot
Cutaneous Melanoma	SHH	7	15559	SNP	G	A	A	Oncomine	NM_00019_3	p.R244C	p.R244	Missense_Mutation	Hotspot
Lung Adenocarcinoma	SHH	7	15559	SNP	G	A	A	Oncomine	NM_00019_3	p.R244C	p.R244	Missense_Mutation	Hotspot
Cutaneous Melanoma	SHH	7	15559	SNP	G	A	A	Oncomine	NM_00019_3	p.R244C	p.R244	Missense_Mutation	Hotspot
Prostate Adenocarcinoma	SLC35G3	17	33520	SNP	C	T	T	Oncomine	NM_15246_2	p.R335K	p.R335	Missense_Mutation	Hotspot
Infiltrating Bladder Urothelial Carcinoma	SLC35G3	17	33520	SNP	C	T	T	Oncomine	NM_15246_2	p.R335K	p.R335	Missense_Mutation	Hotspot
Glioblastoma	SLC35G3	17	33520	SNP	C	T	T	Oncomine	NM_15246_2	p.R335K	p.R335	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	SLC35G3	17	33520	SNP	C	T	T	Oncomine	NM_15246_2	p.R335K	p.R335	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	SLC35G3	17	33520	SNP	G	C	C	Oncomine	NM_15246_2	p.A312G	p.A312	Missense_Mutation	Hotspot
Lung Adenocarcinoma	SLC35G3	17	33520	SNP	C	T	T	Oncomine	NM_15246_2	p.R335K	p.R335	Missense_Mutation	Hotspot
Cutaneous Melanoma	SLC35G3	17	33520	SNP	G	C	C	Oncomine	NM_15246_2	p.A312G	p.A312	Missense_Mutation	Hotspot
Cutaneous Melanoma	SLC35G3	17	33520	SNP	C	T	T	Oncomine	NM_15246_2	p.R335K	p.R335	Missense_Mutation	Hotspot
Cutaneous Melanoma	SPATA8	15	97326	SNP	G	A	A	Oncomine	NM_17349_9	p.E18K	p.E18	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	SPATA8	15	97326	SNP	G	A	A	Oncomine	NM_17349_9	p.E18K	p.E18	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cutaneous Melanoma	SPATA8	15	97326	SNP	G	A	A	Oncomine	NM_173499	p.E18K	p.E18	Missense_Mutation	Hotspot
Cervical Squamous Cell Carcinoma	SPINK13	5	14766	SNP	G	A	A	Oncomine	NM_001040129	p.R84H	p.R84	Missense_Mutation	Hotspot
Cutaneous Melanoma	SPINK13	5	14766	SNP	C	T	T	Oncomine	NM_001040129	p.R84C	p.R84	Missense_Mutation	Hotspot
Cutaneous Melanoma	ST6GAL2	2	10745	SNP	C	T	T	Oncomine	NM_032528	p.E313K	p.E313	Missense_Mutation	Hotspot
Colorectal Adenocarcinoma	ST6GAL2	2	10681	SNP	G	A	A	Oncomine	NM_032528	p.S346L	p.S346	Missense_Mutation	Hotspot
Endometrial Endometrioid Adenocarcinoma	ST6GAL2	2	10746	SNP	C	C	T	Oncomine	NM_032528	p.R11Q	p.R11	Missense_Mutation	Hotspot
Lung Adenocarcinoma	ST6GAL2	2	10745	SNP	C	A	A	Oncomine	NM_032528	p.G235V	p.G235	Missense_Mutation	Hotspot
Lung Adenocarcinoma	ST6GAL2	2	10746	SNP	G	A	A	Oncomine	NM_032528	p.P53L	p.P53	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	ST6GAL2	2	10745	SNP	G	A	A	Oncomine	NM_032528	p.H90Y	p.H90	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	ST6GAL2	2	10745	SNP	C	A	A	Oncomine	NM_032528	p.G235W	p.G235	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	ST6GAL2	2	10742	SNP	C	T	T	Oncomine	NM_032528	p.E455K	p.E455	Missense_Mutation	Hotspot
Cutaneous Melanoma	ST6GAL2	2	10745	SNP	C	T	T	Oncomine	NM_032528	p.E313K	p.E313	Missense_Mutation	Hotspot
Cutaneous Melanoma	ST6GAL2	2	10746	SNP	C	T	T	Oncomine	NM_032528	p.R11Q	p.R11	Missense_Mutation	Hotspot
Cutaneous Melanoma	ST6GAL2	2	10745	SNP	G	A	A	Oncomine	NM_032528	p.S346L	p.S346	Missense_Mutation	Hotspot
Cutaneous Melanoma	ST6GAL2	2	10742	SNP	C	T	T	Oncomine	NM_032528	p.E455K	p.E455	Missense_Mutation	Hotspot
Cutaneous Melanoma	ST6GAL2	2	10745	SNP	T	A	A	Oncomine	NM_032528	p.E313V	p.E313	Missense_Mutation	Hotspot
Cutaneous Melanoma	ST6GAL2	2	10745	SNP	C	T	T	Oncomine	NM_032528	p.G235R	p.G235	Missense_Mutation	Hotspot
Cutaneous Melanoma	ST6GAL2	2	10746	SNP	G	A	A	Oncomine	NM_032528	p.H90Y	p.H90	Missense_Mutation	Hotspot
Cutaneous Melanoma	SYPL1	7	10573	SNP	G	A	A	Oncomine	NM_006754	p.P81S	p.P81	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cutaneous Melanoma	SYPL1	7	10573 9611	SNP	G	A	A	Oncomine	NM_006754	p.P81S	p.P81	Missense_Mutation	Hotspot
Melanoma	SYT1	12	79689 912	SNP	C	T	T	Oncomine	NM_005639	p.P180S	p.P180	Missense_Mutation	Hotspot
Melanoma	SYT1	12	79679 683	SNP	G	A	A	Oncomine	NM_005639	p.E95K	p.E95	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	SYT1	12	79611 355	SNP	C	T	T	Oncomine	NM_005639	p.A19V	p.A19	Missense_Mutation	Hotspot
Acute Myeloid Leukemia	SYT1	12	78135 485	SNP	G	G	A	Oncomine	NM_005639	p.A19T	p.A19	Missense_Mutation	Hotspot
Cutaneous Melanoma	SYT1	12	79689 912	SNP	C	T	T	Oncomine	NM_005639	p.P180S	p.P180	Missense_Mutation	Hotspot
Cutaneous Melanoma	SYT1	12	79679 683	SNP	G	A	A	Oncomine	NM_005639	p.E95K	p.E95	Missense_Mutation	Hotspot
Prostate Adenocarcinoma	SYT1	12	79611 355	SNP	C	T	T	Oncomine	NM_005639	p.A19V	p.A19	Missense_Mutation	Hotspot
Lung Adenocarcinoma	TCEAL8	X	10250 8844	SNP	G	T	T	Oncomine	NM_153333	p.R22S	p.R22	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	TCEAL8	X	10250 8843	SNP	C	T	T	Oncomine	NM_153333	p.R22H	p.R22	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	TCEAL8	X	10250 8844	SNP	G	A	A	Oncomine	NM_153333	p.R22C	p.R22	Missense_Mutation	Hotspot
Prostate Adenocarcinoma	TMEM147	19	36037 641	SNP	C	T	T	Oncomine	NM_032635	p.A92V	p.A92	Missense_Mutation	Hotspot
Glioblastoma	TMEM147	19	36037 641	SNP	C	T	T	Oncomine	NM_032635	p.A92V	p.A92	Missense_Mutation	Hotspot
Cutaneous Melanoma	WFDC5	20	43739 300	SNP	G	A	A	Oncomine	NM_145652	p.R68C	p.R68	Missense_Mutation	Hotspot
Ductal Breast Carcinoma	WFDC5	20	43739 300	SNP	G	G	A	Oncomine	NM_145652	p.R68C	p.R68	Missense_Mutation	Hotspot
Chromophobe Renal Cell Carcinoma	WFDC5	20	43739 299	SNP	C	C	T	Oncomine	NM_145652	p.R68H	p.R68	Missense_Mutation	Hotspot
Clear Cell Renal Cell Carcinoma	ZFAND2B	2	22007 2989	SNP	T	C	C	Oncomine	NM_138802	p.I149T	p.I149	Missense_Mutation	Hotspot
Papillary Renal Cell Carcinoma	ZFAND2B	2	22007 2989	SNP	T	G	G	Oncomine	NM_138802	p.I149S	p.I149	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Non-Small Cell Lung Carcinoma, NOS	ZNF780A	19	40581 109	SNP	T	C	C	Oncomine	NM_001010880	p.I414V	p.I414	Missense_Mutation	Hotspot
Lung Adenocarcinoma	ZNF780A	19	40581 529	SNP	C	T	T	Oncomine	NM_001010880	p.V274I	p.V274	Missense_Mutation	Hotspot
Lung Adenocarcinoma	ZNF780A	19	40581 535	SNP	A	C	C	Oncomine	NM_001010880	p.S272A	p.S272	Missense_Mutation	Hotspot
Oligoastrocytoma	ZNF780A	19	40580 552	SNP	T	G	G	Oncomine	NM_001010880	p.Q599H	p.Q599	Missense_Mutation	Hotspot
Cervical Squamous Cell Carcinoma	ZNF780A	19	40580 552	SNP	T	G	G	Oncomine	NM_001010880	p.Q599H	p.Q599	Missense_Mutation	Hotspot
Gastric Adenocarcinoma	ZNF780A	19	40581 529	SNP	C	T	T	Oncomine	NM_001010880	p.V274I	p.V274	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	ZNF780A	19	40581 109	SNP	T	C	C	Oncomine	NM_001010880	p.I414V	p.I414	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	ZNF780A	19	40580 552	SNP	T	G	G	Oncomine	NM_001010880	p.Q599H	p.Q599	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	ZNF780A	19	40581 529	SNP	C	T	T	Oncomine	NM_001010880	p.V274I	p.V274	Missense_Mutation	Hotspot
Lung Adenocarcinoma	ZNF780A	19	40581 109	SNP	C	T	T	Oncomine	NM_001010880	p.V274I	p.V274	Missense_Mutation	Hotspot
Squamous Cell Lung Carcinoma	ZNF780A	19	40581 535	SNP	A	C	C	Oncomine	NM_001010880	p.S272A	p.S272	Missense_Mutation	Hotspot
Cutaneous Melanoma	ZNF780A	19	40581 535	SNP	A	C	C	Oncomine	NM_001010880	p.S272A	p.S272	Missense_Mutation	Hotspot
Cutaneous Melanoma	ZNF780A	19	40581 109	SNP	T	C	C	Oncomine	NM_001010880	p.I414V	p.I414	Missense_Mutation	Hotspot
Thyroid Gland Follicular Carcinoma	ZNF780A	19	40581 535	SNP	A	C	C	Oncomine	NM_001010880	p.S272A	p.S272	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	ZNF780A	19	40580 552	SNP	T	G	G	Oncomine	NM_001010880	p.Q599H	p.Q599	Missense_Mutation	Hotspot
Lung Adenocarcinoma	ZNF844	19	12187 394	SNP	T	C	C	Oncomine	NM_001113650I	p.F487L	p.F487	Missense_Mutation	Hotspot
Glioblastoma	ZNF844	19	12187 394	SNP	T	C	C	Oncomine	NM_001113650I	p.F487L	p.F487	Missense_Mutation	Hotspot
Glioblastoma	ZNF844	19	12187 275	SNP	G	C	C	Oncomine	NM_001113650I	p.R447P	p.R447	Missense_Mutation	Hotspot

Table 11: Gain of Function mutations

Cancer Type	Gene Symbol	Chromosome	Start Position	Variant Type	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	CBI Annotation Source	Transcript	Variant Change	Variant Position	Variant Classification	Pan-Disease CBI Variant Category
Cervical Squamous Cell Carcinoma	ZNF844	19	12187394	SNP	T	C	C	Oncomine	NM_001136501	p.F487L	p.F487	Missense_Mutation	Hotspot
Head and Neck Squamous Cell Carcinoma	ZNF844	19	12187275	SNP	G	C	C	Oncomine	NM_001136501	p.R447P	p.R447	Missense_Mutation	Hotspot
Lung Adenocarcinoma	ZNF844	19	12187275	SNP	G	C	C	Oncomine	NM_001136501	p.R447P	p.R447	Missense_Mutation	Hotspot
Cutaneous Melanoma	ZNF844	19	12187275	SNP	G	C	C	Oncomine	NM_001136501	p.R447P	p.R447	Missense_Mutation	Hotspot
Cutaneous Melanoma	ZNF844	19	12187394	SNP	T	C	C	Oncomine	NM_001136501	p.F487L	p.F487	Missense_Mutation	Hotspot
Oligodendroglioma	ZNF845	19	53855196	SNP	T	C	C	Oncomine	NM_138374	p.M423T	p.M423	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	ZNF845	19	53855196	SNP	T	C	C	Oncomine	NM_138374	p.M423T	p.M423	Missense_Mutation	Hotspot
Thyroid Gland Papillary Carcinoma	ZNF845	19	53855197	SNP	G	A	A	Oncomine	NM_138374	p.M423I	p.M423	Missense_Mutation	Hotspot

[00149] The disclosure provides novel gene variants and gene variant-disease state associations. The gene variants can have one or more mutations that result in a variant protein. The gene variants provided herein are associated with certain cancers. The gene variants result in protein variants. The disclosure further provides probes, such as amplification primer sets and detection probes, as well as methods of detection, diagnosis, and treatment and kits that include or detect the gene variants disclosed herein.

[00150] The variants are shown as amino acid variants in Tables 7 and 11 with the accession no. or the Entrez nucleotide and/or protein sequence of the parent or wildtype gene provided. The associations with various cancers are shown in Tables 7 and 11. Tables 7 and 11 provide a list of more than 99 genes that were identified using the methods outlined in Example 2. The variations or mutations were not found in the corresponding normal tissue. This is important because in a typical patient, a tumor sample can have 10's-100's of tumor specific variations. However, variations that occur at the same place in multiple patients (and not in the normal tissue) are more significant. 4445 samples (from 4445 patients) were analyzed and list of hotspots was prepared. A number of recurrent mutations were found at the same position in 15-20 different cancer types.

[00151] Diagnostics and kits

[00152] Methods of diagnosing, treating, and detecting gene variants and associated disease are contemplated herein. The methods can include detecting gene fusions and/or gene variants in a subject sample. Any number and combination of gene fusions and/or gene variants can be detected in any of the reaction mixtures, compositions, and kits disclosed herein.

[00153] In one embodiment, the disclosure provides a composition and a kit comprising a set of probes that specifically recognize the nucleotide sequence that encodes a gene variant selected from Table 7 and/or Table 11. The set of probes can be, for example a set of amplification primers. In another embodiment, provided herein is a composition that includes a set of primers that flank a gene variant that encodes one or more variants in Table 7 and/or Table 11. The reaction mixture of this embodiment can further include a detector probe that binds to a nucleotide sequence including a gene variant selected from Table 7 and/or Table 11. The reaction mixture that includes a detector probe or does not include a detector probe, can further include

a polymerase, dNTPs, and/or a uracil DNA deglycosylase (UDG). The polymerase and UDG are typically not from a human origin. The reaction mixture can further include a target nucleic acid, for example a human target nucleic acid. The human target nucleic acid can be, for example, isolated from a biological sample from a person suspected of having a cancer. The cancer can be selected from: BLCA = bladder carcinoma, BRCA = breast carcinoma, CESC = cervical cell carcinoma, COAD = colon adenocarcinoma, GBM = glioblastoma multiforme, HNSC = head and neck squamous cell carcinoma, KIRC = clear cell renal cell carcinoma, KIRP = kidney renal papillary cell carcinoma, LAML = acute myeloid leukemia, LGG = brain lower grade glioma, LIHC = liver hepatocellular carcinoma, LUAD = lung adenocarcinoma, LUSC = squamous cell lung carcinoma, OV = ovarian serous adenocarcinoma, PRAD = prostate adenocarcinoma, READ = rectal adenocarcinoma, SKCM = cutaneous melanoma, STAD = stomach adenocarcinoma, THCA = thyroid carcinoma, and UCEC = uterine corpus endometrioid carcinoma.

[00154] In some embodiments a kit is provided, wherein the kit encompasses one or more probes. In some embodiments, the kit encompasses probes for 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 50, 75, 100, 150, 200, 250, 500 or more fusion genes. In some embodiments the probe is detectably labeled. In some embodiments the probe hybridizes to the breakpoint present in the gene fusion.

[00155] In some embodiments the detection of any one of the gene variants disclosed in Tables 7 and 11 can be combined with the detection of another of the gene variants disclosed in those tables or any of the gene fusions disclosed herein. That is, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 50, 75, 100, 150, 200, 250, 500 or more of the gene variants can be detected in the same reaction. In some embodiments the detected gene variants are those disclosed in Tables 4-6, 7 and 11, 20, and 23 and can be combined with the detection of another of the gene fusion disclosed in those tables. That is, 2, 3, such that 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 50, 75, 100, 150, 200, 250, 500 or more of the gene fusions of can be detected in the same reaction.

[00156] The nucleotide sequence that encodes one or more gene variants in Table 7 and/or Table 11 can be any size that encompasses the variation. For example, the nucleotide sequence can be any size that can be easily copied using a primer and/or detected using a probe.

[00157] In another embodiment, a set of probes that specifically recognize a nucleic acid coding for a gene variant selected from Table 7 and/or Table 11 (gene variants) is provided. In another embodiment, provided herein is a set of primers that specifically amplify a target nucleic acid that codes for a gene variant selected from Table 7 and/or Table 11. In another embodiment, provided herein is a qPCR assay, such as a TaqMan™ assay or a Molecular Beacons™ assay that specifically amplifies and detects a target nucleic acid that codes for a gene variant selected from Table 7 and/or Table 11.

[00158] The disclosure also provides an isolated nucleic acid comprising at least one sequence that includes the variation found in one or more gene variants selected from Table 7 and/or Table 11. The isolated nucleic acid can include a first primer on a 5' end. Furthermore, the nucleic acid can be single stranded or double stranded.

[00159] The disclosure, in other embodiments, provides a kit that includes a detector probe and/or a set of probes, for example, a set of amplification primers that specifically recognize a nucleic acid that codes for a gene variant selected from Table 7 and/or Table 11. For example, in certain embodiments the detector probe or set of amplification primers are designed to amplify and/or detect a nucleic acid that includes at least one of a nucleic acid coding for a gene variant in Table 7 and/or Table 11. The kit can further include, in a separate or in the same vessel, a component from an amplification reaction mixture, such as a polymerase, typically not from human origin, dNTPs, and/or UDG. Furthermore, the kit can include a control nucleic acid. For example the control nucleic acid can include a sequence that includes the gene variant selected from Table 7 and/or Table 11.

[00160] A method of detecting a cancer is provided comprising amplifying a nucleic acid that encodes a gene variant selected from Table 7 and/or Table 11, for example the nucleic acid can include a sequence from one of the accession numbers in Table 7 and/or Table 11 except that the sequence contains the variant that codes for the gene variants in Table 7 and/or Table 11, and detecting the presence of the nucleic acid, wherein the presence of the nucleic acid indicates a cancer is present in the sample. In another method, provided herein is a method of detecting a cancer that includes generating an amplicon that includes a sequence selected from a sequence coding for a gene variant in Table 7 and/or Table 11, and detecting the presence of the nucleic acid, wherein the presence of the nucleic acid indicates bladder, head and neck, or lung squamous cell carcinoma is present in the sample. The amplicon typically includes

primers that are extended to form the amplicon. The cancer is selected from bladder carcinoma, breast carcinoma, cervical cell carcinoma, colon adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, clear cell renal cell carcinoma, kidney renal papillary cell carcinoma, acute myeloid leukemia, brain lower grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, squamous cell lung carcinoma, ovarian serous adenocarcinoma, prostate adenocarcinoma, rectal adenocarcinoma, cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, and uterine corpus endometrioid carcinoma.

[00161] A kit comprising a set of probes, for example, a set of amplification primers that specifically recognize a nucleic acid comprising a gene variant from Table 7 and/or Table 11 is provided. The kit can further include, in a separate or in the same vessel, a component from an amplification reaction mixture, such as a polymerase, typically not from human origin, dNTPs, and/or UDG. Furthermore, the kit can include a control nucleic acid. For example the control nucleic acid can include a sequence that includes the gene variant from Table 7 and/or Table 11. In certain embodiments, a set of probes that specifically recognize a nucleic acid comprising a gene variant from Table 7 and/or Table 11 is provided.

[00162] In another embodiment, a gene variant is provided comprising at least one of the gene variants in Table 7 and/or Table 11.

[00163] In another embodiment is a method to detect a cancer selected from bladder carcinoma, breast carcinoma, cervical cell carcinoma, colon adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, clear cell renal cell carcinoma, kidney renal papillary cell carcinoma, acute myeloid leukemia, brain lower grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, squamous cell lung carcinoma, ovarian serous adenocarcinoma, prostate adenocarcinoma, rectal adenocarcinoma, cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, and uterine corpus endometrioid carcinoma in a sample by detecting the presence of a gene variant selected from Table 7 and/or Table 11. Gene variants, for example, can include, but are not limited to ZNF479 variants R11Q, R295K, R295T, R295I, R345I, R345T, K438T, and T466K (see Table 8).

Table 18

Cancer Type	Gene Symbol	Druggability	KM evidence
Astrocytoma	CXCR2	Y	
Endometrial Endometrioid Adenocarcinoma	CXCR2	Y	

Cancer Type	Gene Symbol	Druggability	KM evidence
Squamous Cell Lung Carcinoma	CXCR2	Y	
Cutaneous Melanoma	CXCR2	Y	
Cutaneous Melanoma	CXCR2	Y	
Colorectal Adenocarcinoma	IL3	Y	
Gastric Adenocarcinoma	IL3	Y	
Cutaneous Melanoma	KCNK9	Y	favorable outcome
Endometrial Endometrioid Adenocarcinoma	KCNK9	Y	
Lung Adenocarcinoma	KCNK9	Y	
Squamous Cell Lung Carcinoma	KCNK9	Y	poor outcome
Non-Small Cell Lung Carcinoma, NOS	S100A8	Y	
Glioblastoma	S100A8	Y	
Head and Neck Squamous Cell Carcinoma	S100A8	Y	
Thyroid Gland Papillary Carcinoma	S100A8	Y	
Cutaneous Melanoma	SHH	Y	
Lung Adenocarcinoma	SHH	Y	
Cutaneous Melanoma	CCDC61		poor outcome
Cutaneous Melanoma	CCDC61		poor outcome
Cutaneous Melanoma	CNTN5		poor outcome
Cutaneous Melanoma	CNTN5		poor outcome
Cutaneous Melanoma	CNTN5		poor outcome
Cutaneous Melanoma	CNTN5		poor outcome
Colorectal Adenocarcinoma	CNTN5		poor outcome
Colorectal Adenocarcinoma	CNTN5		poor outcome
Cutaneous Melanoma	CNTN5		poor outcome
Cutaneous Melanoma	CNTN5		poor outcome
Cutaneous Melanoma	CNTN5		poor outcome
Cutaneous Melanoma	EDDM3A		poor outcome
Cutaneous Melanoma	FABP1		poor outcome
Lung Adenocarcinoma	OR2L13		poor outcome
Cutaneous Melanoma	OR4E2		poor outcome
Cutaneous Melanoma	OR4E2		poor outcome
Cutaneous Melanoma	OR4E2		poor outcome
Cutaneous Melanoma	PRSS37		poor outcome
Cutaneous Melanoma	PRSS37		poor outcome
Cutaneous Melanoma	SPINK13		poor outcome
Endometrial Endometrioid Adenocarcinoma	ST6GAL2		poor outcome

[00164] Table 18 provides druggability or prognostic associations that were filtered from Table 11. Table 18 provides the cancer type, gene symbol, druggability (Y=yes), and KM evidence for the genes identified in Table 11 as druggable. The KM Evidence column provides the Kaplan-Meier evidence. The KM evidence indicates if the event type supports good or poor prognosis in the particular cancer type.

Targeted treatment

[00165] In at least one embodiment, the gene fusions and/or gene variants can be used to identify targeted therapies. Targeted therapies can include the identification of agents that specifically interact with the gene fusion and/or gene variant. Targeted therapies can include, but are not limited to, antibody therapies, antisense therapies and small molecule therapies. Antisense therapies are discussed in more detail under the heading “antisense.”

[00166] Compositions and methods for inactivating nucleic acid molecules involve, in part, the use of molecules with nucleic acid regions with sequence complementarity to the nucleic acid molecule which is the subject of desired inactivation (*i.e.*, a target nucleic acid molecule). Methods of the invention can be used for inactivation of gene fusions and/or gene variants associated with specific cancers. Thus, antisense molecules can be identified that are complementary to any of the gene fusions or gene variants identified herein.

[00167] Small molecules are low molecular weight (<800 Daltons) organic compounds that may serve as enzyme substrates or regulators of biological processes, with a size on the order of 10^{-9} m. In pharmacology, the term is usually used for a molecule that binds to a protein or nucleic acid, and acts as an effector, altering the activity or function of the protein or nucleic acid. Small molecules can be tested for effector functions by expressing a gene fusion or variant in a cellular assay and identifying small molecules that inhibit expression or activity of the gene fusion or variant.

[00168] Druggability is a term used in drug discovery to describe a biological target such as a protein that is known to bind or is predicted to bind with high affinity to a drug. Furthermore, the binding of the drug to a druggable target alters the function of the target with a therapeutic benefit to the patient. The term “drug” herein includes small molecules (low molecular weight organic substances) but also has been extended to include biologic medical products such as therapeutic monoclonal antibodies. In at least one embodiment, the gene fusion or gene variant can be used to identify a druggable target. Table 8 provides a list of druggable targets that have been identified from Tables 1-3 and 7. For example, the TPM1/ALK gene fusion is a druggable target because, as shown in Table 8, diseases for which ALK is involved can be treated with crizotinib. Thus, if a gene fusion includes ALK, the cancer may be treatable with

crizotinib. Further if a gene variant includes a mutation in ALK, the cancer may be treatable with crizotinib.

[00169] Similarly, Table 21 provides a list of druggable targets that have been identified from Table 19 and Table 24 a list of druggable targets that have been identified from Table 22.

Table 8: Druggable genes from Table 1

Druggable Gene	Approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	Preclinical
ALK	crizotinib	N	N	AP-26113; RG-7853; LDK-378; TSR-011	X-396; ASP-3026	NMS-E628; aurora kinase + ALK inhibitor (Sareum, AstraZeneca); ALK inhibitors (AstraZeneca, Cephalon, Aurigene); ARN-5032; DLX-521
CASR	cincacalcet hydrochloride	N	N	N	N	N
EGFR	erlotinib; panitumumab; cetuximab; nepidermin; gefitinib; nimotuzumab; vandetanib; lapatinib ditosylate; icotinib hydrochloride;	Afatinib	zalutumumab; neratinib; dovitinib lactate; XL-647; rindopepimut; necitumumab; dacomitinib	BMS-690514; varlitinib; AC-480; AZD-8931; Sym-004; imgatuzumab; AVL-301; AVL-301; poziotinib; MEHD-7945A; PR-610;	marizomib; CUDC-101; MM-151; AL-6802; S-222611; ABT-806; antroquinonol; GT-MAB 5.2-GEX; epitinib; thelatinib; cipatinib; AMG-595	STP-503; SN-29966; MT-062; STP-801
FGFR3	ponatinib	Masitinib	dovitinib lactate	ENMD-2076; AZD-4547	JNJ-42756493; BGJ-398; LY-2874455; S-49076	N
GNAS	N	N	N	N	N	N
JAK2	ruxolitinib (for idiopathic myelofibrosis)	N	SAR-302503; pacritinib	AT-9283; momelotinib; gandotinib; BMS-911543; NS-018	AC-430; SB-1317	ON-044580; INCB-16562; NVP-BSK805; TP-0413; MRLB-11055; CPL-407-22
NOTCH1	N	N	N	N	OMP-52M51	Debio-0826; TR-4; Notch antibody (AVEO); Notch1 inhibitors (Interprotein)
NTRK1	N	N	N	miliciclib maleate	N	tyrosine kinase inhibitors (Bristol-Myers Squibb); PLX-7486
PIK3CA	N	N	perifosine; buparlisib;	ZSTK-474; PX-866; pictilisib; XL-765; XL-147; BEZ-235; PKI-587; PF-04691502; PF-	INK-1117; GSK-2126458; CUDC-907; GDC-0032; PWT-33597; DS-7423;	LOR-220; AEZS-129; SB-2343; WX-037; PI3/Mnk kinase inhibitors (Progenics); AEZS-132; CLR-1401;

Druggable Gene	Approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	Preclinical
				04691502; BAY-80-6946; BYL-719;	GDC-0084; BAY-1082439; PI3 kinase/mTOR inhibitor (Lilly)	PI3/mTOR kinase inhibitors (Amgen); AEZS-136; HM-032; AMG-511; anticancer therapy (Sphaera Pharma); HMPL-518; GNE-317; mTOR inhibitor/PI3 kinase inhibitor (Lilly); CUDC908; PF-06465603; AEZS-134;
RET	sorafenib; vandetanib; sunitinib malate; cabozantinib; regorafenib	N	motesanib diphosphate; SAR-302503; apatinib		N	MG-516; RET kinase inhibitor; NMS-173
ROS1	crizotinib	N	N	N	N	N
ALK	crizotinib	N	N	AP-26113; RG-7853; LDK-378; TSR-011; NMS-E628	X-396; ASP-3026	NMS-E628; aurora kinase + ALK inhibitor (Sareum, AstraZeneca); ALK inhibitors (AstraZeneca, Cephalon, Aurigene); ARN-5032; DLX-521
NTRK1	N	N	N	milciclib maleate	N	tyrosine kinase inhibitors (Bristol-Myers Squibb); PLX-7486
VIM	N	N	N	pratumumab	N	N
PTK2				PF-04554878	GSK-2256098; BI-853520; VS-4718	CFAK-C4; FAK inhibitors (Varastem, Takeda); CTX-0294945; CTX-0294945
BRS3	N	N	N	N	N	N
TP53	Gendicine	N	N	quinacrine; APR-246; ISA-102	RG-7388; SGT-53; CBLC-137; SAR-405838	PXN-527; ORCA-010; TR-2; ALT-802; OBP-702
STAT3	N	N	N	brivudine; OPB-31121; anatabine citrate; ISIS-STAT3Rx	OPB-51602	CLT-005; GLG-101; GLG-202; GLG-302; GLG-401; PNT-500
NOTCH2	N	N	N	OMP-59R5	N	N
MET	cabozantinib; crizotinib	N	tivantinib; rilotumumab; onartuzumab;	MGCD-265; foretinib; ficlatuzumab; BMS-777607; golvatinib; INCB-028060; LY-2875358	AMG-208; TAS-115; volitinib; SAR-125844; S-49076	X-379; metatatinib; PRS-110; ASP-08001; ARGX-111; DCC-2701; DCC-2721; MG-516; AL-2846; CG-206481; T-1840383; cMet-

Druggable Gene	Approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	Preclinical
						EGFR dual inhibitors(CrystalGenomics); bispecific antibodies (Hoffmann-La Roche)
CDH1	N	N	N	N	N	N
TOP1	belotecan hydrochloride; irinotecan hydrochloride; topotecan	N	cositecan; irinotecan, HyACT; irinotecan, PharmaEngine; etirinotecan pegol	gimatecan; camptothecin, Calando; irinotecan HCl + floxuridine, Celator; firtecane pegol; TLC-388 hydrochloride; hRS7-SN-38; irinotecan bead, Biocompatibles	irinotecan, liposomal, Yakult; HM-30181A; namitecan; camptothecin prodrug, Mersana; labetuzumab-SN-38; Genz-644282; simmitemcan hydrochloride prodrug	camptothecin (Aphios); irinotecan (BioAlliance); cisplatin + irinotecan (Celator); APH-0804; irinotecan (Champions); SER-203; SN-38; topotecan + vincristine (LipoCure); topotecan (EnduRx Pharmaceuticals)
RARA	tamibarotene	N	N	IRX-5183	N	N
ERBB2	trastuzumab; trastuzumab emtansine; pertuzumab; lapatinib ditosylate; catumaxomab; afatinib	trastuzumab, Enhance	neratinib; XL-647; dacomitinib; nelipepimut-S; trastuzumab (Celltrion, Biocad, Biocon, Synthron, Harvest Moon, Aryogen)	lapuleucel-T; AVX-901; AE-37; BMS-690514; MVA-BN-HER2; varlitinib; MM-111; AC-480; ovarian cancer vaccine (Generex); margetuximab; poztotinib; PR-610	Her-VAXX; VM-206; ARRY-380; JNJ-26483327; S-222611; doxorubicin (Merrimack); cipatinib; TrasGEX; trastuzumab (Hanwha Chemical); trastuzumab (Pfizer); IDN-6439	Lovaxin B; TH-1 (Algeta); trastuzumab-antibody conjugates (Synthon); CUDC-101; Her-2/neu Stradobody (Gliknik); ARX-788; Etbx-021; SN-34003; IBI-302; NT-004; ICT-140; ONS-1050; Sym-013; anti-HER2 X anti-CD3 (Emergent Biosolutions); Z-650; breast cancer vaccine (Cel-Sci); JNJ-28871063; trastuzumab (PlantForm, BioXpress, biOasis Technologies, Stada, Natco, Curaxys, Oncobiologics, Alteogen, Mabion)
ALK	crizotinib	N	N	AP-26113; RG-7853; LDK-378; TSR-011; NMS-E628	X-396; ASP-3026	NMS-E628; aurora kinase + ALK inhibitor (Sareum, AstraZeneca); ALK inhibitors (AstraZeneca, Cephalon, Aurigene); ARN-5032; DLX-521

Druggable Gene	Approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	Preclinical
NTRK1	N	N	N	milciclib maleate	N	tyrosine kinase inhibitors (Bristol-Myers Squibb); PLX-7486
LTK	crizotinib	N	N	N	N	N
BRAF	pazopanib; vemurafenib; dabrafenib	N	N	RAF-265; XL-281; LGX-818	ARQ-761; ARQ-736	AB-024; b-raf inhibitors (Sareum); BRAF kinase inhibitor (Selexagen Therapeutics); BeiGene-283; DP-4978; TL-241

[00170] Table 8 provides a list of 11 druggable targets that were identified in the gene fusions in Tables 1-3 or gene variants in Tables 7 and 11. Tables 16 and 17 provide an analysis of other druggable targets within Tables 1-3 or gene variants in Tables 7 and 11. Tables 8, 16 and 17 provide information about druggable targets including the gene name whether the drug has been approved (N=no) by the U.S. Food and Drug Administration (FDA), if the drug has not been approved, which phase the clinical trial is in (Pre-registration, Phase III, Phase II, Phase I, and preclinical). For example, the drug associated with the NOTCH1 gene has not been approved, but is in Phase 1 of clinical trials (see OMP-52M51) as of this writing.

[00171] Approved drugs include, but are not limited to, crizotinib for diseases having ALK gene fusions and cinacalcet hydrochloride for diseases having CASR gene fusions. A number of approved drugs have been identified for gene fusions having EGFR, including, but not limited to, erlotinib; panitumumab; cetuximab; nepidermin; gefitinib; nimotuzumab; vandetanib; lapatinib ditosylate; and icotinib hydrochloride. The approved drug ponatinib has been identified for diseases having FGFR3, ruxolitinib has been identified for diseases having JAK2 gene fusions. A number of approved drugs have been identified for gene fusions having RET, including but not limited to, sorafenib; vandetanib; sunitinib malate; cabozantinib; and regorafenib. The approved drug crizotinib has been identified for diseases having ROS1. Additional drugs that may prove useful include, but are not limited to, crizotinib, afatinib, masitinib, zalutumumab, neratinib, dovitinib lactate, XL647, rindopepimut, nectumumab, dacomitinib, SAR-302503, pacritinib, perifosine, buparlisib, motesinib diphosphate, and apatinib.

[00172] Methods provided herein can include delivering a drug to a subject or a patient. The drug can be an approved drug according to a governmental drug regulatory authority, such as the FDA, or the drug can be in any of the stages before the approved stage. In illustrative aspects, the drug is an FDA-approved drug. In other aspects the drug can be in a pre-clinical, Phase I, Phase II, Phase III, or pre-approval stage. In certain aspects, the methods provided herein include delivering one or more than one of the drugs listed in Tables 8, 16 and 17 to a subject. Where genetic events are identified in a subject that involve more than one gene listed in Tables 8, 16 and 17, methods provided herein can include delivering more than one drug, particularly delivering drugs associated with the different genes affected by the identified genetic events.

[00173] Antisense

[00174] Antisense technology has been applied to inhibit the expression of various oncogenes. For example, Craf-1 cDNA fragments in an antisense orientation, brought under the control of an adenovirus 2 late promoter introduced into a human squamous carcinoma resulted in a greatly reduced tumorigenic potential relative to cells transfected with control sense transfectants. Similarly, a Cmyc antisense construct accelerated differentiation and inhibited G₁ progression in Friend Murine Erythroleukemia cells. Antisense methodology takes advantage of the fact that nucleic acids tend to pair with "complementary" sequences. Complementary sequences are those polynucleotides which are capable of base-pairing according to the standard Watson-Crick complementarity rules. Targeting double-stranded (ds) DNA with polynucleotides leads to triple-helix formation; targeting RNA will lead to double-helix formation. Antisense polynucleotides, when introduced into a target cell, specifically bind to their target polynucleotide and interfere with transcription, RNA processing, transport, translation and/or stability. Antisense RNA constructs, or DNA encoding such antisense RNA's, may be employed to inhibit gene transcription or translation or both within a host cell, either *in vitro* or *in vivo*, such as within a host animal, including a human subject.

[00175] Antisense can be under transcriptional control of a promoter. A "promoter" refers to a DNA sequence recognized by the synthetic machinery of the cell, or introduced synthetic machinery, required to initiate the specific transcription of a gene. The phrase "under transcriptional control" means that the promoter is in the correct location and orientation in relation to the nucleic acid to control RNA polymerase initiation and expression of the gene.

[00176] In certain instances, an antisense expression construct will comprise a virus or engineered construct derived from a viral genome. Where a cDNA insert is employed, a polyadenylation signal to effect proper polyadenylation of the gene transcript may be included. The nature of the polyadenylation signal is not believed to be crucial and any such sequence may be employed. A terminator can be used to enhance message levels and to minimize read through from the cassette into other sequences.

[00177] Antisense constructs may be designed to bind to the promoter and other control regions, exons, introns or even exon-intron boundaries of a gene fusion or gene variant disclosed herein. The most effective antisense constructs include regions complementary to intron/exon splice junctions. One embodiment includes an antisense construct with complementarity to regions within 50-200 bases of an intron-exon splice junction. It has been observed that some exon sequences can be included in the construct without seriously affecting the target selectivity thereof. The amount of exonic material included will vary, depending on the particular exon and intron sequences used. One can readily test whether too much exon DNA is included simply by testing the constructs *in vitro* to determine whether normal cellular function is affected or whether the expression of related genes having complementary sequences is affected.

[00178] The word "complementary" with respect to antisense means polynucleotide sequences that are substantially complementary over their entire length and have very few base mismatches. For example, sequences of fifteen bases in length may be termed complementary when they have complementary nucleotides at thirteen or fourteen positions. Naturally, sequences which are completely complementary will be sequences which are entirely complementary throughout their entire length and have no base mismatches. Other sequences with lower degrees of homology also are contemplated. For example, an antisense construct which has limited regions of high homology, but also contains a non-homologous region (e.g., ribozyme) could be designed. These molecules, though having less than 50% homology, would bind to target sequences under appropriate conditions.

[00179] *In vivo*, *ex vivo* or *in vitro* delivery of antisense can involve the use of vectors. One effective vector for antisense delivery is an adenovirus expression vector. "Adenovirus expression vector" is meant to include those constructs containing adenovirus sequences sufficient to support packaging of the construct and to express an antisense polynucleotide that has been cloned therein. The expression vector can

include a genetically engineered form of adenovirus. Adenovirus is particularly suitable for use as a gene transfer vector because of its mid-sized genome, ease of manipulation, high titer, wide target-cell range and high infectivity. Both ends of the viral genome contain 100-200 base pair inverted repeats (ITRs), which are cis elements necessary for viral DNA replication and packaging.

[00180] Adenovirus is easy to grow and manipulate and exhibits broad host range *in vitro* and *in vivo*. This group of viruses can be obtained in high titers, e.g., 10^9 - 10^{11} plaque-forming units per ml, and they are highly infective. The life cycle of adenovirus does not require integration into the host cell genome. The foreign genes delivered by adenovirus vectors are episomal and, therefore, have low genotoxicity to host cells. No side effects have been reported in studies of vaccination with wild-type adenovirus, demonstrating their safety and therapeutic potential as *in vivo* gene transfer vectors.

[00181] The retroviruses are a group of single-stranded RNA viruses characterized by an ability to convert their RNA to double-stranded DNA in infected cells by a process of reverse-transcription. The resulting DNA then stably integrates into cellular chromosomes as a provirus and directs synthesis of viral proteins. The integration results in the retention of the viral gene sequences in the recipient cell and its descendants. The retroviral genome contains three genes, gag, pol, and env that code for capsid proteins, polymerase enzyme, and envelope components, respectively. A sequence found upstream from the gag gene contains a signal for packaging of the genome into virions. Two long terminal repeat (LTR) sequences are present at the 5' and 3' ends of the viral genome. These contain strong promoter and enhancer sequences and are also required for integration in the host cell genome.

[00182] In order to construct a retroviral vector, a nucleic acid encoding a gene of interest is inserted into the viral genome in the place of certain viral sequences to produce a virus that is replication-defective. In order to produce virions, a packaging cell line containing the gag, pol, and env genes but without the LTR and packaging components is constructed. When a recombinant plasmid containing a human cDNA, together with the retroviral LTR and packaging sequences is introduced into a cell line (by calcium phosphate precipitation for example), the packaging sequence allows the RNA transcript of the recombinant plasmid to be packaged into viral particles, which are then secreted into the culture media. The recombinant retrovirus is then collected, optionally concentrated, and used for gene transfer. Retroviral vectors are able to infect

a broad variety of cell types. However, integration and stable expression require the division of host cells.

[00183] A novel approach designed to allow specific targeting of retrovirus vectors was recently developed based on the chemical modification of a retrovirus by the chemical addition of lactose residues to the viral envelope. This modification could permit the specific infection of hepatocytes via sialoglycoprotein receptors.

[00184] Other viral vectors may be employed as expression vectors. Vectors derived from viruses such as vaccinia virus, adeno-associated virus (AAV) and herpes viruses may be employed.

[00185] In order to effect expression of sense or antisense gene constructs, the expression vector may be delivered into a cell. This delivery may be accomplished *in vitro*, as in laboratory procedures for transforming cells lines, or *in vivo* or *ex vivo*, as in the treatment of certain disease states. As described above, one mechanism for delivery is via viral infection where the expression construct is encapsidated in an infectious viral particle.

[00186] Several non-viral methods for the transfer of expression constructs into cultured mammalian cells also are contemplated. These include calcium phosphate precipitation DEAE-dextran, electroporation, direct microinjection, DNA-loaded liposomes, and lipofectamine-DNA complexes, cell sonication, gene bombardment using high velocity microprojectiles, and receptor-mediated transfection. Some of these techniques may be successfully adapted for *in vivo* or *ex vivo* use.

[00187] Pharmaceutical Compositions - Where clinical applications are contemplated, pharmaceutical compositions can be produced--either gene delivery vectors or engineered cells--in a form appropriate for the intended application. Generally, this entails preparing compositions that are essentially free of pyrogens, as well as other impurities that could be harmful to humans or animals.

[00188] Appropriate salts and buffers are used to render delivery vectors stable and allow for uptake by target cells. Buffers also will be employed when recombinant cells are introduced into a patient. Aqueous compositions of the present invention comprise an effective amount of the vector to cells, dissolved or dispersed in a pharmaceutically acceptable carrier or aqueous medium. The phrase "pharmaceutically or pharmacologically acceptable" refer to molecular entities and compositions that do not produce adverse, allergic, or other untoward reactions when administered to an animal

or a human. As used herein, "pharmaceutically acceptable carrier" includes any and all solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonic and absorption delaying agents and the like. The use of such media and agents for pharmaceutically active substances is well known in the art. Except insofar as any conventional media or agent is incompatible with the vectors or cells of the present invention, its use in therapeutic compositions is contemplated.

[00189] The expression vectors and delivery vehicles may be administered via any common route so long as the target tissue is available via that route. This includes oral, nasal, buccal, rectal, vaginal or topical. Alternatively, administration may be by orthotopic, intradermal, subcutaneous, intramuscular, intraperitoneal or intravenous injection. Such compositions would normally be administered as pharmaceutically acceptable compositions.

[00190] An effective amount of the therapeutic agent is determined based on the intended goal. The term "unit dose" refers to a physically discrete unit suitable for use in a subject, each unit containing a predetermined quantity of the therapeutic composition calculated to produce the desired response in association with its administration, i.e., the appropriate route and treatment regimen. The quantity to be administered, both according to number of treatments and unit dose, depends on the subject to be treated, the state of the subject and the protection desired. Precise amounts of the therapeutic composition also depend on the judgment of the practitioner and are peculiar to each individual.

[00191] Therapeutic Kits - All the essential materials and reagents required for inhibiting tumor cell proliferation may be assembled together in a kit. This generally will comprise selected expression vectors, viruses or cells. Also included may be various media for replication of the expression vectors and host cells for such replication. Such kits will comprise distinct containers for each individual reagent. The kits may also include an instruction sheet defining (i) administration of the antisense expression vector construct; (ii) the antisense expressing viruses; and (iii) the antisense expressing cells.

[00192] In some embodiments, an interfering (iRNA or siRNA) is provided. In some embodiments the iRNA is complementary to the breakpoint of a fusion gene.

Methods associated with clinical outcome discoveries

[00193] Tables 15 and 39 provided herein, contain more than 100 genetic events, including gain-of-function mutations, loss-of-function mutations, in-peak gene amplification/deletions, and fusion events for various cancer types that are associated with a clinical outcome with high statistical significance ($q < 0.1$). Accordingly, provided herein are methods for delivering a treatment to a subject, methods for determining whether a subject receives a treatment, methods for determining whether to deliver a treatment, and methods for delivering a report. The treatment, in certain illustrative embodiments, is a drug. As non-limiting examples, the drug can be a drug listed in Tables 8, 16 and 17, especially where the method involves a genetic event that affects the gene listed for the drug in Tables 8, 16 and 17. In other examples, the drug can be any drug approved by a regulatory agency, or any drug in a stage of development before approval, as discussed herein.

[00194] Accordingly, in another embodiment, a method of delivering a treatment to a subject is provided, wherein the method includes detecting a genetic event identified in Table 15, and treating the subject, wherein the treatment is believed to positively affect the clinical outcome of cancer patients having the genetic event and/or is believed to affect a biological pathway associated with the genetic event. This embodiment can be considered a method for determining if a subject receives a treatment or a method for determining whether to deliver or perform a treatment to or on a subject. Thus, provided herein is a method for determining if a subject receives a drug, the method includes detecting a genetic event identified in Table 15 and/or 39, and then delivering a drug to the subject if the detected genetic event is listed in Table 15 and/or 39, wherein the drug is believed to positively affect the clinical outcome of patients having the genetic event. In illustrative aspects of these embodiments, the genetic event is associated with a gene found in Tables 8, 16 and 17, and the drug is listed in Tables 8, 16 and 17, as a companion for that gene. The subject is typically a subject that has a cancer of the type listed in Table 15 and/or 39. In illustrative aspects of this embodiment the genetic event is associated with a poor prognosis for the subject, who is afflicted with a cancer, typically the cancer listed in Table 15 and/or 39 for which the poor prognosis is associated with that genetic event.

[00195] In another embodiment, provided herein is a method of delivering a report, wherein the method includes detecting a genetic event identified in Table 15 and/or 39 and delivering to a medical professional, a report that provides a predicted clinical outcome associated with that genetic event for a cancer of the subject. The medical

professional can be, as non-limiting examples, a physician, genetic counselor, or other medical professional. Typically, the physician, genetic counselor, or other medical professional have a professional relationship with the subject, such as a patient/doctor relationship. The report can be a paper report or can be an electronic report delivered to the medical professional over a computer network. The method and report can include one or more of the genetic events and associated prognosis identified in Table 15 and/or 39.

[00196] In another embodiment, provided herein is a method for determining which treatment to administer to a subject, the method includes detecting a genetic event listed in Table 15, and administering the treatment depending on the genetic event that is detected. In illustrative embodiments, the treatment is an aggressive treatment, such as a treatment that will involve more pain and suffering for the patient as a result of the treatment, if the detected genetic event is associated with a poor prognosis. In related embodiments the treatment is a more aggressive treatment if the detected genetic event is associated with a poor prognosis and a less aggressive treatment if the detected genetic event is another genetic event, especially if the detected genetic event is identified in Table 15 and/or 39 as indicating a good prognosis. For example, if a AADAC gene deletion, an amplification of the CHD1L gene, the FMO5 gene, or the PRKAB2 gene, or a combination thereof, is detected in a lung cancer adenocarcinoma patient, the patient may be treated with an aggressive chemotherapeutic drug regimen. If these genetic events are not detected in the patient, then the patient may be monitored but the chemotherapeutic drug may not be administered.

[00197] In another embodiment, provided herein is a method for determining whether to treat a cancer patient, the method includes detecting a genetic event listed in Table 15 and/or 39, and treating the subject if a genetic event is detected that is associated in Table 15 with a poor prognosis. In another embodiment, provided herein is a method for determining whether to treat a cancer patient, the method includes detecting a genetic event listed in Table 15 and/or 39, and not treating the subject if a genetic event is detected that is associated in Table 15 and/or 23 with a good prognosis. In another embodiment, provided herein is a method for determining whether to treat or monitor a cancer patient, the method includes detecting a genetic event listed in Table 15 and/or 39, and monitoring, but not treating the subject if a genetic event is detected that is associated in Table 15 and/or 39 with a good prognosis. Treatment may

be administered at a later time if the monitoring detects recurrence or progression of the cancer.

[00198] In certain aspects of these embodiments of the invention that relate to methods provided herein based on the clinical outcomes associated with genetic events in Table 15 and/or 39, for example methods for delivering a treatment to a subject or determining whether to deliver a treatment to a subject, or determining which treatment to administer or deliver, or methods for delivering a report, the subject can be identified as having any of the types of genetic events and any of the specific genetic events listed in Table 15 and/or 39. For example, the genetic event can be a gain-of-function mutation, loss-of-function mutation, a gene amplification or deletion, typically an in-peak gene amplification/deletion, or a fusion event. In certain illustrative embodiments the genetic event is identified in Table 15 and/or 39 of having a q-value of 1×10^{-3} or less, 1×10^{-4} or less, or 1×10^{-5} or less. In certain aspects, the genetic event is listed in Table 15 and/or 39 as involving a druggable gene. For example, the genetic event can be a genetic event listed in Table 15 and/or 39 associated with a gene that is a preclinical drug target. As a non-limiting example, provided herein is a method for determining which treatment or course of treatment to administer to a patient who has ovarian cancer, for example ovarian serous cystadenocarcinoma, wherein the method includes detecting or otherwise determining an amplification of the ID1 or BCL2L1 gene and administering the treatment. The treatment in illustrative embodiments, is an approved treatment for BCL2L1, such as a currently FDA-approved BCL2L1 treatment, wherein a BCL2L1 amplification is detected.

[00199] Methods are known to skilled artisans for detecting the types of genetic events listed in Table 15 and/or 39. Those methods can include nucleic acid sequencing methods or amplification methods, such as PCR or isothermal amplification methods, or combinations thereof. Those methods can include providing a primer that is designed to bind to a gene identified in Table 15 and/or 39 or bind upstream of a gene identified in Table 15 and/or 39. Thus, provided herein are reaction mixtures and kits that include a nucleic acid sample for a subject and one or more primers that bind to, or upstream from, a gene identified in Table 15 and/or 39. Typically, the gene is associated with a genetic event in Table 15 and/or 39, and the subject has a cancer identified in Table 15 and/or 39 as having a prognosis associated with the genetic event. The kit can also include a control nucleic acid that is bound by the primer as disclosed herein for various embodiments of the invention. The reaction mixture can

also include a polymerase as disclosed herein for various embodiments of the invention.

[00200] In certain aspects of these embodiments of the invention that relate to methods provided herein based on the clinical outcomes associated with genetic events in Table 15 and/or 39, for example methods for delivering a treatment to a subject or determining whether to deliver a treatment to a subject, methods for determining which treatment to deliver, or methods for delivering a report to a medical professional, the genetic event can include more than one of the genetic events identified in Table 15 and/or 39. In certain aspects, a method according to this embodiment detects 2, 3, 4, 5, 6, 7, 8, 9, 10, or more of the genetic events identified in Table 15, especially those identified with the same prognosis for a given cancer type. For example, the method can include detecting a genetic event in a breast cancer patient and administering a treatment to the patient, where the detected genetic event includes a gene amplification of two or more of the BRF2, ERLIN2, GPR124, PROSC, and TAB11FI genes. In another example, the method includes detecting two or more genetic events in a subject afflicted with a lower grade glioma and administering a treatment to the subject, wherein the genetic event is at least two of an amplification of the EGFR or SEC61G gene, an amplification of the CDK4, CYP27B1, MARCH9, TSPAN31, or AGAP2 gene, a gain of function mutation in the EGFR gene, or a deletion of the CDKN2A, CDKN2B, or MTAP gene. In another aspect, the method includes detecting a genetic event associated with a poor prognosis and the genetic event is identified in Tables 8, 16, 17, Table 15 and/or 39 as being a target for a current drug in pre-clinical trials or an approved drug, such as an FDA approved drug.

[00201] In certain aspects of these embodiments of the invention that relate to methods provided herein based on the clinical outcomes associated with genetic events in Table 15 and/or 39, for example methods for delivering a treatment to a subject or determining whether to deliver a treatment to a subject, or determining which treatment to administer or deliver, or methods for delivering a report, the genetic event can be a specific genetic event identified in one of the other tables herein. A skilled artisan can identify which general type of genetic event in Table 15 and/or 39 a specific genetic event in one of the other tables will fall under.

[00202] Computer Implemented Systems

[00203] Computer systems can be utilized to in certain embodiments of the disclosure. In various embodiments, computer system can include a bus or other

communication mechanism for communicating information, and a processor coupled with bus for processing information. In various embodiments, computer system 100 can also include a memory, which can be a random access memory (RAM) or other dynamic storage device, coupled to bus for determining base calls, and instructions to be executed by processor. Memory also can be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor. In various embodiments, computer system can further include a read only memory (ROM) or other static storage device coupled to bus for storing static information and instructions for processor. A storage device, such as a magnetic disk or optical disk, can be provided and coupled to bus for storing information and instructions.

[00204] In various embodiments, computer system can be coupled via bus to a display, such as a cathode ray tube (CRT) or liquid crystal display (LCD), for displaying information to a computer user. An input device, including alphanumeric and other keys, can be coupled to bus for communicating information and command selections to processor. Another type of user input device is a cursor control, such as a mouse, a trackball or cursor direction keys for communicating direction information and command selections to processor and for controlling cursor movement on display. This input device typically has two degrees of freedom in two axes, a first axis (i.e., x) and a second axis (i.e., y), that allows the device to specify positions in a plane.

[00205] A computer system can perform the present teachings. Consistent with certain implementations of the present teachings, results can be provided by computer system 100 in response to processor executing one or more sequences of one or more instructions contained in memory. Such instructions can be read into memory from another computer-readable medium, such as storage device. Execution of the sequences of instructions contained in memory can cause processor to perform the processes described herein. Alternatively hard-wired circuitry can be used in place of or in combination with software instructions to implement the present teachings. Thus implementations of the present teachings are not limited to any specific combination of hardware circuitry and software.

[00206] In various embodiments, the term "computer-readable medium" as used herein refers to any media that participates in providing instructions to processor for execution. Such a medium can take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Examples of non-volatile media

can include, but are not limited to, optical or magnetic disks, such as storage device. Examples of volatile media can include, but are not limited to, dynamic memory, such as memory. Examples of transmission media can include, but are not limited to, coaxial cables, copper wire, and fiber optics, including the wires that comprise bus.

[00207] Common forms of non-transitory computer-readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, or any other magnetic medium, a CD-ROM, any other optical medium, punch cards, paper tape, any other physical medium with patterns of holes, a RAM, PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, or any other tangible medium from which a computer can read.

[00208] In accordance with various embodiments, instructions configured to be executed by a processor to perform a method are stored on a computer-readable medium. The computer-readable medium can be a device that stores digital information. For example, a computer-readable medium includes a compact disc read-only memory (CD-ROM) as is known in the art for storing software. The computer-readable medium is accessed by a processor suitable for executing instructions configured to be executed.

[00209] In accordance with the teachings and principles embodied in this application, methods, systems, and computer readable media that can efficiently collect, analyze, store, transfer, retrieve, and/or distribute information across multiple sites and/or entities, including genomic and/or patient information, are provided.

[00210] In one embodiment, a system is provided for determining whether one or more gene fusion and/or variant is present in a sample. The system can further determine identify a disease state, such as cancer, associated with the one or more gene fusion and/or gene variant, as well as an appropriate treatment in accordance with the mutation status. In certain embodiments, the system comprises a processor in communication with a sequencing instrument that receives sequencing data.

[00211] In some embodiments, the processor can execute one or more variant calls. In some embodiments, the processor can provide, filter, and/or annotate predictions.

EXAMPLES

[00212] In the following examples, methods were used to identify gene fusions and gene variants associated with a panel of 19 cancers in 4,225 cancer patient samples. The gene fusions and gene variants are then used to produce diagnostic methods to identify a predisposition for cancer, to diagnose cancer, to stage cancer, to provide a prognosis and to identify a druggable cancer. Methods are provided to provide targeted therapy for the cancer based on the identification of gene fusions.

[00213] EXAMPLE 1: High-throughput systematic analysis of paired-end next-generation sequencing data to characterize the gene fusion landscape in cancer

[00214] 4,225 cancer patient samples across 19 diseases were processed with deFuse McPherson et al. "deFuse: an algorithm for gene fusion discovery in tumor RNASeq data" *PLoS Comp. Bio.* 2011. and TopHat (Kim et al. "TopHat-Fusion: an algorithm for discovery of novel fusion transcripts" *Genome Biology* 2011) gene fusion calling software using a cloud-based computation infrastructure. Filtering criteria were identified for gene fusion events that enriched for high confidence, chemically validated gene fusion events.

[00215] Gene fusions encode oncogenic drivers in hematological and solid tumors and are often associated with dramatic clinical responses with the appropriate targeted agents. Massively parallel paired-end sequencing can identify structural rearrangements in tumor genomes and transcriptomes. However, computational methods to identify gene fusions are varied, still evolving and largely trained on cell line data. Systematic methods were developed to characterize known oncogenic gene fusions and to discover novel gene fusions in cancer. RNASeq data for approximately 3,400 clinical cases from 16 cancer types was obtained from the Cancer Genomics Hub (CGHub) of the Cancer Genome Atlas (TCGA). The performance of several gene fusion callers was surveyed and two were chosen (deFuse and TopHat) for further method development with the goal of supporting both single and paired end data. An analysis pipeline was developed and executed in parallel on a high-performance computing cluster. Filtering and annotation was conducted on aggregated data as a post-processing step to enable exploratory analyses of various filters. Filtering approaches were optimized on datasets that included known standards (e.g., TMPRSS2.ERG in prostate adenocarcinoma, PML.RARA in acute myeloid leukemia, etc.) to enrich for these and other gene fusions with correct 5'-3' orientation while excluding cases with ambiguous breakpoints and spanning reads, alignment errors, and

read through transcripts from adjacent genes. Predicted fusions were summarized based on the occurrence of unique genes participating in fusion with multiple partners and of unique gene pairs, each within specific diseases. Elevated expression was observed after the predicted breakpoint of the 3' gene in cases positive for predicted fusions and added important confirmatory evidence. Pan-disease fusions and multi-partner fusion events broadened the clinical population scope of gene fusion events.

[00216] All single-end data was processed using TopHat and all paired-end data was processed using deFuse. TopHat has been shown to be effective with longer 75 bp single-end data. The deFuse algorithm is not compatible with single-end data and has been designed to leverage read pairs. The pre-processing data and Detect fusions: deFuse TopHat steps were executed in parallel for all samples on a high-performance computing cluster. The filtering and annotation was conducted on the aggregated data as a post-processing step to enable filtering criteria to minimize false positive fusions. The list of priority fusions was validated with RNASeq Exon Expression data.

[00217] TCGA Data Source: All RNASeq data for gene fusion analysis was obtained from the Cancer Genomics Hub (CGHub), the current repository for TCGA genomic data – [hypertext transfer protocol secure ://cghub.ucsc.edu/](https://cghub.ucsc.edu/). Table 9 lists the TCGA sample counts downloaded and processed for M2 and M3:

Table 9: TCGA samples processed

Cancer Type	Cancer Type Abbreviation	Samples	Center	Instrument
Bladder Urothelial Carcinoma	BLCA	122	UNC-LCCC	Illumina HiSeq 2000
Breast invasive carcinoma	BRCA	841	UNC-LCCC	Illumina HiSeq 2000
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	88	UNC-LCCC	Illumina HiSeq 2000
Colon adenocarcinoma	COAD*	196	UNC-LCCC	Illumina GA Iix
Glioblastoma multiforme	GBM	167	UNC-LCCC	Illumina HiSeq 2000
Head and Neck squamous cell carcinoma	HNSC	302	UNC-LCCC	Illumina HiSeq 2000
Kidney Chromophobe	KICH	66	UNC-LCCC	Illumina HiSeq 2000
Kidney renal clear cell carcinoma	KIRC	480	UNC-	Illumina HiSeq

Cancer Type	Cancer Type Abbreviation	Samples	Center	Instrument
			LCCC	2000
Kidney renal papillary cell carcinoma	KIRP	76	UNC-LCCC	Illumina HiSeq 2000
Acute Myeloid Leukemia	LAML	179	BCCAGSC	Illumina GA IIx
Brain Lower Grade Glioma	LGG	184	UNC-LCCC	Illumina HiSeq 2000
Liver hepatocellular carcinoma	LIHC	34	UNC-LCCC	Illumina HiSeq 2000
Lung adenocarcinoma	LUAD	345	UNC-LCCC	Illumina HiSeq 2000
Lung squamous cell carcinoma	LUSC	221	UNC-LCCC	Illumina HiSeq 2000
Ovarian serous cystadenocarcinoma	OV	417	BCCAGSC	Illumina HiSeq 2000
Pancreatic adenocarcinoma	PAAD	31	UNC-LCCC	Illumina HiSeq 2000
Prostate adenocarcinoma	PRAD	140	UNC-LCCC	Illumina HiSeq 2000
Rectum adenocarcinoma	READ*	71	UNC-LCCC	Illumina GA IIx
Skin Cutaneous Melanoma	SKCM	267	UNC-LCCC	Illumina HiSeq 2000
Stomach adenocarcinoma	STAD	41	BCCAGSC	Illumina HiSeq 2000
Thyroid carcinoma	THCA	373	UNC-LCCC	Illumina HiSeq 2000
Uterine Corpus Endometrioid Carcinoma	UCEC*	317	UNC-LCCC	Illumina GA IIx

* Single-end TCGA disease

BAM files were downloaded from CGHub using its Gene Torrent Software

[00218] With the goal of supporting both single and paired-end data, 4,374 paired-end samples were processed with deFuse and 584 single-end samples with TopHat.

[00219] Broadly, the analysis pipeline consisted of 5 main steps: 1. Pre-process the raw data to obtain FASTQ files 2. Run fusion callers 3. Filter breakpoints to gene regions of interest 4. Annotate the breakpoints with the Oncomine transcript set and 5. Summarize and prioritize potentially interesting novel fusions.

[00220] The input to the fusion callers consisted of RNASeq reads in FASTQ format, which required conversion of the BAM file provided by TCGA to one or two FASTQ files for single or paired end data (respectively).

[00221] A custom SamToFastq converter was developed to generate FASTQ files from a TCGA BAM file. In addition to allowing conversion of all paired-end RNASeq TCGA BAMs systematically, the SamToFASTQ converter had other advantages over other conversion tools. First, it was written in C and compiled to run faster and reduce processing time. Second, it incorporated several validation steps to ensure proper mate pairing and consistent mate pair ordering in the output FASTQ files, both of which are input requirements for the fusion callers.

[00222] There were 3 cancer types (COAD, READ, UCEC) only available as single-end RNASeq data. For single-end BAM file conversion the program BamTools ([hyper text transfer protocol secure://github.com/pezmaster31/bamtools](https://github.com/pezmaster31/bamtools)) was used to generate FASTQ files.

[00223] Integration – FIG. 1 diagrams the relative levels of result filtering done by both callers. As part of the analysis “Level I” data was integrated – the output from TopHat-Fusion Post’s *potential_fusion.txt* file and the output from deFuse’s *results.classify.tsv* file. The integration steps involved converting the reported breakpoints to ones based on the genomic coordinate system and consolidation into a common file format.

[00224] Breakpoint Filtering - The ~5.5 million predictions from the “Level I” output of the callers were filtered to only retain those calls where each breakpoint was either in the 5’UTR or CDS region of a RefSeq transcript (refGene circa July18, 2012, obtained from UCSC). This was done to enrich the predicted fusions for those containing functional gene regions. Breakpoints predicted to occur in intronic sequences were also excluded, resulting in a set of 423,587 predicted chimeras.

[00225] Breakpoint Annotation - For each pair of breakpoints, only one transcript per Entrez ID was retained. This ensured consistency in annotating breakpoints at the same location. However, predicted breakpoints at different locations for the same gene partners may still result in multiple transcripts representing a pair of genes – possible evidence of alternative transcripts.

[00226] Basic annotation from the callers was discarded, as it was based on the default annotation source of each respective caller. However, certain output fields from

both TopHat and deFuse were retained to help prioritize the predicted fusions. Additionally, certain annotation properties that weren't explicitly reported by the callers were inferred from other caller properties.

[00227] Inferred Properties - Supporting and Spanning read counts were obtained from each caller and summarized in to **Reads Span** and **Reads Span Support**. The latter is a sum of reads spanning the fusion and those supporting the fusion. The breakpoint sequence reported by the callers was trimmed to include 50 bases on each side of the fusion and consolidated as Breakpoint Sequence. The fusion breakpoint is delineated by a "|". Since neither of the callers provides a definitive '5-prime' or '3-prime' flag, the relative 5'-3' orientation of the fusion partners was inferred by combining a caller parameter with the gene strand annotation. A Valid Orientation field was labeled as "Y" if there was an inferred 5' and 3' partner for a given gene fusion call.

[00228] RepeatMasker Annotation - Each predicted breakpoint location was also annotated with RepeatMasker features in the neighborhood of the breakpoint. This was done to identify breakpoints in highly repetitive genomic regions, where alignment errors were likely to affect the prediction of the chimeric transcript. For each fusion prediction, a RepeatMasker Overlap field was set to 1 if either of the breakpoint flank sequences overlaps with a RepeatMasker element by 12 or more bases. The frequency of overlapping fusion calls is used in the Oncomine Prioritization described below such that gene fusion predictions with a lower frequency of overlap are considered higher quality.

[00229] Fusion Exon Expression Imbalance - Recurrent Oncomine Priority Fusions were vizualized using RNASeq exon expression data downloaded using the GDAC Firehose tool to provide secondary evidence of true positive fusion events by searching for exon expression imbalance before and after the breakpoint call. Specifically, if the 3' partner's expression is impacted by the 5' partner's promoter region, then exon expression should increase post the predicted breakpoint. This effect is especially visible when viewing fused versus non-fused patient samples.

[00230] RPKM RNASeq values are listed for each patient as Gene Annotation Format (GAF) features corresponding to a composite of UCSC exons from several different gene definitions including Refseq. Compendia processed fusion breakpoints were mapped to the GAF features. 80.8% of the 396,298 Refseq exons map perfectly to GAF features in the plot shown below. The Refseq exon and GAF feature pair that resulted in the largest overlap was selected and reported on.

[00231] A value called *rg_pct* provides a metric of the mapping quality of a given Refseq exon with a GAF feature based on the following formula: $rg_pct = \frac{overlap}{length_{refseq} * overlap / length_{GAF\ feature}}$

[00232] Mappings with an *rg_pct* value of 1 overlap perfectly, while values less than 1 indicate the refseq exon or GAF feature did not map to the exact same genomic regions and the RPKM value may be suspect. RNASeq V2 data was selected for all diseases except OV, STAD, and LAML due to disease coverage shown in the barplot.

[00233] Fusion exon expression was manually reviewed for expression imbalance of a subset of Oncomine Priority fusions meeting the following criteria: 1. Recurrent Oncomine Priority Fusions 2. Oncomine Priority Fusions that are listed in the Mitelman Database 3. One fusion partner is an Oncomine Gain of Function Oncogene and involved in at least 3 Oncomine Priority Fusions and 4. One fusion partner is listed in the Sanger Cancer Gene Census (hypertext transfer protocol:www.sanger.ac.uk/genetics/CGP/Census/) and involved in at least 3 Oncomine Priority Fusions.

[00234] A total of 994 gene fusions meet these criteria and were manually reviewed for exon expression imbalance by assigning a “supported”, “refuted”, “neutral” or “not tested” rating to the gene fusion call.

[00235] Experts used the following criteria to assign ratings: Supported: Fused samples had a highly expressed 3' fusion partner post-breakpoint such that fused samples were outliers of the patient population. Prior to the breakpoint, the 3' partner's expression should be low compared to post-breakpoint. Refuted: Extremely low average expression of the 5' partner (< 5 RPKM) or average expression of one partner is much lower than the other (~1/10). Neutral: Neither Support or Refute criteria are met. Fusions that were not manually reviewed were assigned a rating of Not Tested.

[00236] Fusion Summarization - Fusions were summarized within a disease based on the occurrence of unique gene pairs, and based on the occurrence of individual genes, possibly with multiple partners.

[00237] Fusion-Level Summary - For a unique fusion pair (unique by Entrez ID pair), the number of samples within a disease with at least one prediction of that fusion by either caller is the Fused Sample Count. Since multiple breakpoints for the same pair of genes may be reported in one sample and across the samples, the number of unique fusion pairs within each disease represented by the 424K+ fusion calls was 49,588.

Table 10 shows the properties that were summarized for a given fusion partner pair across the individual predictions:

Table 10:

Property	Summary Method
DEFUSE_EVERSION	% of total fusion calls = 'Y'
DEFUSE_VALID_ORIENTATION	% of total fusion calls = 'Y'
DEFUSE_NUM_MULTI_MAP	% of total fusion calls > 0
TOPHAT_VALID_ORIENTATION	% of total fusion calls = 'Y'
3P/5P_REPEATMASKER_OVERLAP	% of total fusion calls = 1

The **Adjacent** flag is set for a fusion if the genes are <1Mb apart on the genome and the *defuse_eversion* flag is set in $\leq 75\%$ of the individual fusion prediction for these fusion partners.

[00238] Mitelman Cross-reference - Individual unique fusion pairs were cross-referenced to the Mitelman database of genomic aberrations (hypertext transfer protocol://cgap.nci.nih.gov/Chromosomes/Mitelman downloaded February 25, 2013). The match was done based on gene names and not disease type. Therefore, gene fusions reported in Mitelman in a certain disease may have occurred in a different disease type in the TCGA datasets. Gene fusions summarized at the gene level were cross-referenced to the Mitelman database based on gene name. Thus, there is more potential for the gene as reported in Mitelman to be of different histology or altogether different aberration type (for example a large chromosome-level deletion instead of a fusion) than the predicted unique fusion pairs.

[00239] Normal Sample Fusion Blacklist - To reduce the number of false positive fusions, 344 paired-end normal samples were processed across 10 diseases using the same deFuse pipeline described above. A total of 56,579 total fusion calls consisting of 6,024 unique fusions were observed. Of the 49,588 unique gene fusion events, 11,801 of these calls were observed in normal samples. These normal sample fusion calls were used to generate a blacklist and thereby remove these false positives from the OncoPrint gene fusions.

[00240] Paralogous Fusion Partner Blacklist - A blacklist of fusions between paralogous gene family members was assembled using two strategies: 1) manually inspecting high frequency fusion partner gene names and 2) comparing the first 3 characters of all Priority Fusion partner gene names. In the latter strategy, fusion partners were verified to be "paralogous" using HomoloGene, Ensembl, and SIMAP before inclusion in the final blacklist. This blacklist consists of 375 unique paralogous

gene fusions and was used to remove false positives from the Oncomine Priority gene fusions.

[00241] EXAMPLE 2: NGS mutation methods for identifying gene variants associated with cancer

[00242] The goal of the data integration for gene variants was to create the most complete set of mutation data currently available from the TCGA.

[00243] Data Sources- For this release, the following were integrated: TCGA mutation data from the Broad GDAC Mutation_Packager 2013_02_22 stddata build, Level 2 (public, experimentally un-validated) data available from the TCGA DCC as of March 1st 2013, and, for prostate adenocarcinoma, mutation data generated by Compendia from TCGA primary data.

[00244] Compendia (CBI) Mutation Calls – There was concern that the prostate adenocarcinoma mutation calls available from TCGA were of low quality and resulted in false-positive 'Gain of Function' predictions. Therefore, all calls for this disease were sourced from Compendia's own mutation calling pipeline, which closely parallels the process used by the TCGA cancer type working groups to generate the publically-available mutation calls.

Table 12: Data Source Selection

Cancer Type	TCGA Disease	Mutation Packager (2013_02_22)	DCC (20130301)	Compendia	TOTAL
Bladder Urothelial Carcinoma	BLCA	28			
Breast Invasive Carcinoma	BRCA	772			
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	CESC	39			
Colon Adenocarcinoma	COAD	153			
Glioblastoma, Multiforme	GBM	290			
Head and Neck Squamous Cell Carcinoma	HNSC		306		
Kidney Renal Clear Cell Carcinoma	KIRC	293			
Kidney Renal Papillary Cell	KIRP	100			

Cancer Type	TCGA Disease	Mutation Packager (2013_02_22)	DCC (20130301)	Compendia	TOTAL
Carcinoma					
Acute Myeloid Leukemia	LAML	196			
Brain Lower Grade Glioma	LGG	169			
Lung Adenocarcinoma	LUAD		379		
Lung Squamous Cell Carcinoma	LUSC	178			
Ovarian Serous Cystadenocarcinoma	OV	316			
Pancreatic Adenocarcinoma	PAAD	34			
Prostate Adenocarcinoma	PRAD			170	
Rectal Adenocarcinoma	READ	68			
Skin Cutaneous Melanoma	SKCM	252			
Stomach Adenocarcinoma	STAD		136		
Thyroid Carcinoma	THCA	323			
Uterine Corpus Endometrioid Carcinoma	UCEC	235			
					4,437

[00245] Data Cleaning - some simple clean-up operations were performed to remove duplicate mutation records present in the source data. Duplicate mutations from various tumor/normal aliquots pairs of the same patient sample were removed. A total of 25 “ultra-mutator” samples (mutation count of > 5,000 per sample) were also excluded from the downstream analysis pipelines. In certain diseases, such as uterine corpus endometrioid carcinoma, several highly-mutated samples may dominate the overall mutation counts and dilute the results of mutation recurrence analysis necessary for the Compendia mutation and gene classification scheme.

[00246] Mutation Annotation: A. Compendia Annotation - Compendia’s approach to defining mutations relied on accurate variant annotation hence; the mutations were re-annotated using a standard annotation pipeline which ensured that mutations across disease types were evaluated consistently and were subject to common interpretation during the nomination of potential oncogenes or tumor suppressor genes.

[00247] Mutations obtained from TCGA were processed by Compendia according to the following general steps: 1. Each mutation was first re-annotated using the Compendia transcript set. Successfully annotated mutations received Compendia-derived annotation, while the rest retain annotation obtained from the TCGA. Annotation includes: Variant classification, Variant position, Variant change. 2. Redundant annotations of a mutation in multiple transcripts were removed. 3. Mutations located outside of gene regions of interest were removed. 4. Mutations without a valid gene Entrez ID were removed.

[00248] "Mutation" is defined herein as a specific change at a genomic location, i.e.: Chromosome, start, stop, reference base, alternate base, variant type (SNP, INS, DEL) etc.

[00249] "Annotation" is defined herein as a transcript-specific set of properties that describe the effect of the mutation, i.e.: Gene, transcript, variant classification, variant change, variant codon position, etc.

[00250] In the Mutation Annotation step, the mutations obtained from TCGA were re-annotated against a standard transcript set compiled by Compendia. This transcript set included RefGene transcripts from hg18 and hg19 genome builds, obtained from UCSC.

[00251] Each mutation was individually mapped against a contig in the CBI Transcript Set within the specified genome build. SNP mutations were mapped directly to their start location, while for small insertion (INS) and deletion (DEL) mutations a position of interest is selected for mapping.

[00252] For a mutation successfully mapped to a transcript, the CBI mutation annotation was inferred with respect to that transcript. For mutations that fail to map, the more limited TCGA annotation was retained, and a variant position for Hotspot calculations was constructed based on the genomic coordinate.

[00253] Below is a description of the criteria used in annotating the mutations that map to the CBI Transcript Set:

[00254] Variant Classification: For each mutation successfully mapped to a transcript, the variant classification was inferred using the location and the sequence variant type of the mutation. This approach identified the following main mutation variant classifications:

Table 13: main mutation variant classifications:

Variant Classification	Transcript Region
Splice_Site	exon or intron
3'UTR, 5'UTR	UTR exon
Intron	intron
Missense, Nonsense, Nonstop, Silent	coding exon
Frame_Shift_Ins/Del	coding exon
In_Frame_Ins/Del	coding exon
Non_Coding_Exon	exon of a non-coding gene

[00255] Variant Position: The variant position of a mutation is the location used to identify genes with Hotspot mutations, which are mutations of a certain classification that are observed at the same location in multiple tumor samples. To effectively identify recurrence and define a hotspot for each mutation, a mutation spot identifier was constructed that encompassed the mutation position, the identity of the amino acid or base affected, and the variant classification. Mutations that occurred at the same location irrespective of the specific base change they generated were aggregated. Therefore, only the reference base or amino acid was used to define the variant position. This ensured that mutations affecting the same codon or genomic position would be counted towards a possible hotspot, even if the alternate alleles they generated were different. For example, for a given gene, missense mutations V600E, V600F and V600G would all have a variant position of V600 and would thus be aggregated together when identifying hotspot mutations. When the amino-acid level position was not available, the RNA-level or genomic-level position was utilized.

[00256] For mutations that do not map to the CBI Transcript Set, and hence do not have a transcript-based location, the genomic location (start position) and the reference nucleotide (reference allele) was used as the variant position irrespective of the coding region or splice site proximity. The TCGA-annotated variant classification was then added as a suffix. The variant change (see below) for these mutations was not defined.

[00257] Variant Change: The variant change provides HGVS-like information about the alternate allele change of the mutation (e.g. V600E). For SNP mutations in the coding region, the variant change was a full HGVS protein-level sequence variant description, indicating the alternate amino acid. For SNPs outside of the coding region, the alternate allele nucleotide base was provided. For mutations that do not map to the CBI Transcript Set, the variant classification from TCGA was retained.

[00258] Transcript Filtering: To avoid retrieving multiple transcripts, and hence, multiple annotations for a single mutation within a gene, only one transcript per mutation per gene (unique Entrez ID) were kept. If a mutation mapped to several transcripts of a gene, only one was chosen. However, if a mutation mapped to several genes, then only one transcript per gene was selected. It was thus possible for a mutation to receive two different annotations, but only if they stemmed from transcripts with different Entrez IDs. In effect, any mutation of the same variant classification at the same genomic location was always assigned to the same transcript, and hence would be in the same frame of reference when computing recurrence for hotspot identification.

[00259] Gene Region Filtering: All mutations were further filtered by variant type and class to avoid including mutations of minor interest to gene function analysis. Mutations were filtered out that were not resolved to a gene region, either because they fell significantly far outside of a transcript, or because they were in a location not associated with a RefSeq gene. These mutations were evident either by their lack of gene identifier, or membership in the following variant classes: *Intron*, *5'Flank*, *IGR*, and *miRNA*. Mutations were also filtered out with variant type of *DNP*, *TNP*, *ONP*, *Complex_substitution*, and *Indel*, as their annotation was not supported by the pipeline

[00260] Classifying Mutations as Hotspot, Deleterious, or Other - The next step in the analysis pipeline identified recurring mutations in multiple samples based on their variant position, and categorized them into Hotspot, Deleterious or Other variant categories. For this step, and the subsequent frequency calculations, mutations for each disease type were processed independently. Only mutations of the same variant classification were tallied together, so, for example, a missense mutation and a silent mutation at the same position was counted separately.

[00261] To identify driver events, each mutation for a given Entrez Gene Id was categorized as "Deleterious" or "Hotspot". A mutation was deemed 'recurrent' if it was observed in the same variant position in 3 or more tumor samples. A mutation belonged to the "Hotspot" variant category if it was recurrent and was annotated with one of the following variant classifications: In-frame insertion/deletion, Nonstop, Missense, Non_Coding_Exon. A mutation belonged to the "Deleterious" category if it was annotated with one of the following variant classifications: Frame shift insertion/deletion, Nonsense. A mutation was considered in the "Other" variant category if it did not fit the above criteria.

[00262] Nominating "Gain of Function" and "Loss of Function" Genes - Individual genes were classified into predicted functional classes, namely "Gain of Function", "Recurrent Other", and "Loss of Function" to reflect their relative enrichment in potential activating or deleterious mutations.

[00263] **Frequency of Mutations:** Mutation frequencies for each gene were calculated with respect to a given variant classification and variant category across all samples within a disease type. Overall mutation frequency for a gene within a disease was calculated by combining all the mutations.

[00264] **Mutation Significance:** The Hotspot p-values for each gene within a disease were calculated by selecting the most recurrent mutation m and using sampling to determine the probability p of observing r or more mutations at that position. More specifically: $p = \frac{100,000 - \sum_{m=1}^{r-1} c_m}{100,000}$, where c_m is the count of replicates with maximum multiplicity m . P-values for transcripts with a maximum multiplicity of one are defined as 1.0. P-value for transcripts with a maximum multiplicity that is never observed is defined as $1e-5$.

[00265] Hotspot Q-values were calculated within each disease by counting the number of transcripts mutated at least once (N) and calculating the rank of each p-value. The q-value for a given p-value is then $Q = p * N / \text{rank}$.

[00266] To assess whether a gene was significantly enriched for deleterious mutations compared with other genes, given the background mutation rate, Fisher's exact test was performed comparing the deleterious mutation frequency of the gene in question to that of other genes. Nonsense mutations, frame shift insertions and frame shift deletions were classified as deleterious mutations, while mutations of any other type (missense, etc., but non-intergenic) counted as others.

[00267] Deleterious Q-values were calculated within each disease, by counting the number of genes with deleterious mutations (N), and calculating the rank of each association. The q-value for a given p-value was then $Q = p * N / \text{rank}$.

[00268] **Gene Classification:** Once the mutations were classified, individual genes were nominated to one of three classes – "Gain of Function," "Loss of Function," and "Recurrent Other." The classification is based on the combination of relative frequencies and the significance of the mutations observed in the gene. The significance of the mutations per gene is assessed by a p-value. The classification scheme in FIG. 2 specifies the criteria for Gain of Function and Loss of Function genes.

[00269] A "Gain of Function" gene will have a relatively high frequency of Hotspot Missense mutations and a low frequency of Deleterious mutations, while a "Loss of Function" gene contains a large fraction of Deleterious mutations. "Recurrent Other" tend to contain recurrent insertion/deletion mutations, some of which - for example recurrent frame shift indels of 1 base - exhibit signs of potential false-positive calls that may arise from local alignment errors.

[00270] Pan-Cancer Analysis - To summarize mutations across diseases identical calculations were performed as for within-disease analyses, but without stratifying the mutation records by disease. For the pan-disease gene classification, the genes (unique by Entrez ID) were summarized across all cancer types.

[00271] **EXAMPLE 3: Diagnostic assay for the identification of gene fusions and/or gene variants in cancer**

[00272] Library Preparation

[00273] PCR Amplify Genomic DNA Targets

[00274] The disclosed variant and fusion polynucleotides can be detected by the sequencing of nucleic acids. This can be accomplished by next generation sequencing, the description of which follows. The source of the nucleic acid for next generation sequencing can include a Fresh-Frozen Paraffin-Embedded (FFPE) sample.

[00275] A multiplex polymerase chain reaction is performed to amplify 384 individual amplicons across a genomic DNA sample. A pool of greater than 32,000 primers is developed covering more than 100 gene variants or fusion polynucleotides. Each primer in the primer pool was designed to contain at least one uridine nucleotide near the terminus of each primer. Each primer is also designed to selectively hybridize to, and promote amplification, by forming a primer pair, with a specific gene, gene variant, or fusion polypeptide of a nucleic acid sample.

[00276] To a single well of a 96-well PCR plate is added 5 microliters of the Primer Pool containing 384 primer pairs at a concentration of 15 μ M in TE, 10-50 ng genomic DNA and 10 microliters of an amplification reaction mixture (2x AmpliSeq HiFi Master Mix) that can include glycerol, dNTPs, and Platinum® Taq High Fidelity DNA Polymerase (Invitrogen, Catalog No. 11304) to a final volume of 20 microliters with DNase/RNase Free Water (Life Technologies, CA, Part No. 600004).

[00277] The PCR plate is sealed and loaded into a thermal cycler (GeneAmp® PCR system 9700 Dual 96-well thermal cycler (Life Technologies, CA, Part No. N8050200 and 4314445)) and run using the following temperate profile to generate the preamplified amplicon library.

[00278] An initial holding stage is performed at 98°C for 2 minutes, followed by 16 cycles of denaturing at 98°C for 15 seconds and an annealing and extending stage at 60°C for 4 minutes. After cycling, the preamplified amplicon library is held at 4°C until proceeding to the purification step outlined below.

[00279] Purify the Amplicons from input DNA and Primers

[00280] Two rounds of Agencourt® AMPure® XP Reagent (Beckman Coulter, CA) binding, wash, and elution at 0.6x and 1.2x volume ratios are found to remove genomic DNA and unbound or excess primers. The amplification and purification step outlined herein produces amplicons of about 100 bp to about 600 bp in length.

[00281] In a 1.5ml LoBind tube (Eppendorf, Part No. 022431021), the preamplified amplicon library (20 microliters) is combined with 12 microliters (0.6x volumes) of Agencourt® AMPure® XP reagent (Beckman Coulter, CA). The bead suspension is pipetted up and down to thoroughly mix the bead suspension with the preamplified amplicon library. The sample is then pulse-spin and incubated for 5 minutes at room temperature.

[00282] The tube containing the sample is placed on a magnetic rack such as a DynaMag™-2 spin magnet (Life Technologies, CA, Part No. 123-21D) for 2 minutes to capture the beads. Once the solution cleared, the supernatant is transferred to a new tube, where 24 microliters (1.2x volume) of AgenCourt® AMPure® XP beads (Beckman Coulter, CA) is added to the supernatant. The mixture is pipetted to ensure that the bead suspension is mixed with the preamplified amplicon library. The sample is then pulse-spun and incubated at room temperature for 5 minutes. The tube containing the sample is placed on a magnetic rack for 2 minutes to capture the beads. Once the solution clears, the supernatant is carefully discarded without disturbing the bead pellet. The desired preamplified amplicon library is then bound to the beads. Without removing the tube from the magnetic rack, 200 microliters of freshly prepared 70% ethanol is introduced into the sample. The sample is incubated for 30 seconds while gently rotating the tube on the magnetic rack. After the solution clears, the supernatant is discarded without disturbing the pellet. A second ethanol wash is performed and the

supernatant discarded. Any remaining ethanol is removed by pulse-spinning the tube and carefully removing residual ethanol while not disturbing the pellet. The pellet is air-dried for about 5 minutes at room temperature.

[00283] Once the tube is dry, the tube is removed from the magnetic rack and 20 microliters of DNase/RNase Free Water is added (Life Technologies, CA, Part No. 600004). The tube is vortexed and pipetted to ensure the sample is mixed thoroughly. The sample is pulse-spun and placed on the magnetic rack for two minutes. After the solution clears, the supernatant containing the eluted DNA is transferred to a new tube.

[00284] Phosphorylate the Amplicons

[00285] To the eluted DNA (~ 20 microliters), 3 microliters of DNA ligase buffer (Invitrogen, Catalog No. 15224041), 2 microliters dNTP mix, and 2 microliters of FuP reagent are added. The reaction mixture is mixed thoroughly to ensure uniformity and incubated at 37°C for 10 minutes.

[00286] Ligate Adapters to the Amplicons and Purify the Ligated Amplicons

[00287] After incubation, the reaction mixture proceeds directly to a ligation step. Here, the reaction mixture now containing the phosphorylated amplicon library is combined with 1 microliter of A/P1 Adapters (20 µm each)(sold as a component of the Ion Fragment Library Kit, Life Technologies, Part No. 4466464) and 1 microliter of DNA ligase (sold as a component of the Ion Fragment Library Kit, Life Technologies, Part No. 4466464), and incubated at room temperature for 30 minutes.

[00288] After the incubation step, 52 microliters (1.8x sample volume) of AgenCourt® AMPure® Reagent (Beckman Coulter, CA) is added to the ligated DNA. The mixture is pipetted thoroughly to mix the bead suspension with the ligated DNA. The mixture is pulse-spun and incubated at room temperature for 5 minutes. The samples undergo another pulse-spin and are placed on a magnetic rack such as a DynaMag™-2 spin magnet (Life Technologies, CA, Part No. 123-21D) for two minutes. After the solution clears, the supernatant is discarded. Without removing the tube from the magnetic rack, 200 microliters of freshly prepared 70% ethanol is introduced into the sample. The sample is incubated for 30 seconds while gently rotating the tube on the magnetic rack. After the solution clears, the supernatant is discarded without disturbing the pellet. A second ethanol wash is performed and the supernatant is discarded. Any remaining ethanol is removed by pulse-spinning the tube and carefully removing

residual ethanol while not disturbing the pellet. The pellet is air-dried for about 5 minutes at room temperature.

[00289] The pellet is resuspended in 20 microliters of DNase/RNase Free Water (Life Technologies, CA, Part No. 600004) and vortexed to ensure the sample is mixed thoroughly. The sample is pulse-spun and placed on the magnetic rack for two minutes. After the solution clears, the supernatant containing the ligated DNA is transferred to a new LoBind tube (Eppendorf, Part No. 022431021).

[00290] Nick Translate and Amplify the Amplicon Library and Purify the Library

[00291] The ligated DNA (~20 microliters) is combined with 76 microliters of Platinum® PCR SuperMix High Fidelity (Life Technologies, CA, Part No. 12532-016, sold as a component of the Ion Fragment Library Kit, Life Technologies, Part No. 4466464) and 4 microliters of Library Amplification Primer Mix (5 µM each) (Life Technologies, CA, Part No. 602-1068-01, sold as a component of the Ion Fragment Library Kit, Life Technologies, Part No. 4466464), the mixture is pipetted thoroughly to ensure a uniformed solution. The solution is applied to a single well of a 96-well PCR plate and sealed. The plate is loaded into a thermal cycler (GeneAmp® PCR system 9700 Dual 96-well thermal cycler (Life Technologies, CA, Part No. N8050200 and 4314445)) and run on the following temperate profile to generate the final amplicon library.

[00292] A nick-translation is performed at 72°C for 1 minute, followed by an enzyme activation stage at 98°C for 2 minutes, followed by 5-10 cycles of denaturing at 98°C for 15 seconds and an annealing and extending stage at 60°C for 1 minute. After cycling, the final amplicon library is held at 4°C until proceeding to the final purification step outlined below.

[00293] In a 1.5ml LoBind tube (Eppendorf, Part No. 022431021), the final amplicon library (~100 microliters) is combined with 180 microliters (1.8x sample volume) of Agencourt® AMPure® XP reagent (Beckman Coulter, CA). The bead suspension is pipetted up and down to thoroughly mix the bead suspension with the final amplicon library. The sample is then pulse-spun and incubated for 5 minutes at room temperature.

[00294] The tube containing the final amplicon library is placed on a magnetic rack such as a DynaMag™-2 spin magnet (Life Technologies, CA, Part No. 123-21D) for 2 minutes to capture the beads. Once the solution clears, the supernatant is carefully

discarded without disturbing the bead pellet. Without removing the tube from the magnetic rack, 400 microliters of freshly prepared 70% ethanol is introduced into the sample. The sample is incubated for 30 seconds while gently rotating the tube on the magnetic rack. After the solution clears, the supernatant is discarded without disturbing the pellet. A second ethanol wash is performed and the supernatant is discarded. Any remaining ethanol is removed by pulse-spinning the tube and carefully removing residual ethanol while not disturbing the pellet. The pellet is air-dried for about 5 minutes at room temperature.

[00295] Once the tube is dry, the tube is removed from the magnetic rack and 20 microliters of Low TE was added (Life Technologies, CA, Part No. 602-1066-01). The tube is pipetted and vortexed to ensure the sample is mixed thoroughly. The sample is pulse-spin and placed on the magnetic rack for two minutes. After the solution clears, the supernatant containing the final amplicon library is transferred to a new Lobind tube (Eppendorf, Part No. 022431021).

[00296] Assess the Library Size Distribution and Determine the Template Dilution Factor

[00297] The final amplicon library is quantitated to determine the library dilution (Template Dilution Factor) that results in a concentration within the optimized target range for Template Preparation (e.g., PCR-mediated addition of library molecules onto Ion Sphere™ Particles). The final amplicon library is typically quantitated for downstream Template Preparation procedure using an Ion Library Quantitation Kit (qPCR) (Life Technologies, Part No. 4468802) and/or a Bioanalyzer™ (Agilent Technologies, Agilent 2100 Bioanalyzer) to determine the molar concentration of the amplicon library, from which the Template Dilution Factor is calculated. For example, instructions to determine the Template Dilution Factor by quantitative real-time PCR (qPCR) can be found in the Ion Library Quantitation Kit User Guide (Life Technologies, Part No. 4468986), hereby incorporated by reference in its entirety.

[00298] In this example, 1 microliter of the final amplicon library preparation is analyzed on the 2100 Bioanalyzer™ with an Agilent High Sensitivity DNA Kit (Agilent Technologies, Part No. 5067-4626) to generate peaks in the 135-205 bp size range and at a concentration of about 5×10^9 copies per microliter.

[00299] Proceed to Template Preparation

[00300] An aliquot of the final library is used to prepare DNA templates that are clonally amplified on Ion Sphere™ Particles using emulsion PCR (emPCR). The preparation of template in the instant example is prepared according to the manufacturer's instructions using an Ion Xpress Template Kit (Life Technologies, Part No. 4466457), hereby incorporated by reference in its entirety. Once template-positive Ion Sphere Particles are enriched, an aliquot of the Ion Spheres are loaded onto an Ion 314™ Chip (Life Technologies, Part No. 4462923) as described in the Ion Sequencing User Guide (Part No. 4467391), hereby incorporated in its entirety, and subjected to analysis and sequencing as described in the Ion Torrent PGM Sequencer User Guide (Life Technologies, Part No. 4462917), hereby incorporated in its entirety.

[00301] EXAMPLE 4: Oncomine NGS Integrative Analysis Methods to Identify Genetic Events Associated with Clinical Outcomes

[00302] The Oncomine NGS Integrative Analysis was designed to bring together the largest possible set of core NGS data to enable scientific workflows that interrogate relationships across data types and diseases, summarizing the analyses at multiple biological levels of abstraction, such as genes and pathways.

[00303] Data Sources (Oncomine is available from Life Technologies/Compendia Biosciences – Ann Arbor, Michigan and <http://www.oncomine.org>)

[00304] The data for the Integrative Analysis was taken from the below sources:

[00305] Fusions: Oncomine driver fusions

[00306] Mutations: Oncomine pan-cancer driver mutations

[00307] CNVs: Peak amplification and deletion data derived from Oncomine-processed copy number data

[00308] DNA: Oncomine-processed DNA-seq continuous data

[00309] RNA: Normalized gene-level RNAseq continuous data

[00310] Clinical: Oncomine-curated clinical and outcome metadata

[00311] Pathways: Oncomine pathway definitions

[00312] Fusions Data and Filtering

[00313] Fusion data for integrative analysis was obtained from Oncomine NGS Fusion data.. Oncomine Prioritized Fusion is a priority scheme developed at

Compendia to capture attributes of known true positive fusion events and characterize a subset of observed gene fusions as high-confidence priority fusions. Criteria used to define priority fusions include: valid 5' to 3' orientation, non-adjacent fusion partners, uniquely mapping spanning reads, non-paralogous fusion partners, not observed in normal tissue, and non-overlapping with redundant regions in the genome.

[00314] Fusions were included and considered driver fusions if they were called by deFuse or Tophat, had exon expression evidence that was “supported” or “neutral” and met one of the following 4 criteria:

[00315] Oncomine Prioritized Fusion + Recurrent

[00316] Oncomine Prioritized Fusion + Mitelman Annotated

[00317] Oncomine Prioritized Fusion + One partner is an Oncomine Gain of Function gene involved in 3 or more Pan-Disease Priority Fusions

[00318] Oncomine Prioritized Fusion + One partner is a Sanger Oncogene (<http://goo.gl/JQBw9>) involved in 3 or more Pan-Disease Priority Fusions

[00319] Mutations Data and Filtering

[00320] Mutation data for Integrative Analysis was obtained from Oncomine NGS Mutation data. Individual genes are classified into predicted functional classes, namely “Gain of Function” and “Loss of Function” to reflect their relative enrichment in potential activating or deleterious mutations. This classification is based on the combination of relative frequencies and the significance of the mutations observed in the gene assessed by a p-value. A “Gain of Function” gene will have a relatively high frequency of Hotspot Missense mutations and a low frequency of Deleterious mutations, while a “Loss of Function” gene contains a large fraction of Deleterious mutations.

[00321] Copy Number Segmentation and Quantification

[00322] DNA copy number data for each TCGA sample was obtained from Oncomine. Measurements from multiple reporters for a single gene were averaged.

[00323] Minimum Common Region (MCR) Peak Generation

[00324] In genes that were recurrently amplified (4 or more copies) or deleted (1 or less copy), peaks were identified independently in 25 cancer types by applying MCR analysis on Oncomine clinical samples. To define peaks, contiguous genomic regions with multiple genes that were significantly aberrant (common regions) were identified first. In every common region, a Peak is defined as one or more genes whose aberrant

sample count meets or exceeds a peak threshold. In every cancer, common regions are defined as regions whose aberrant sample count meet or exceed a common region threshold. The baseline, average number of aberrant samples observed across all genes, is calculated for every arm of every chromosome in every cancer.

[00325] mRNA Gene Expression

[00326] Expression data was obtained from the Broad GDAC's TCGA Standard Data.

[00327] Clinical Data Curation

[00328] Patient clinical data was obtained from TCGA and curated by Compendia. Curated data types included demographics, major clinical and histological disease subtypes, and clinical outcome data. All properties were standardized to be consistent across the diseases.

[00329] Construction of Clinically Relevant Subsets

[00330] Curated clinical data obtained from TCGA and Oncomine NGS data was used and the rules in Table 14 were applied to define the Clinical Subsets:

Table 14 – Rules to define the Clinical Subsets

Disease	Clinical Subtype	Source	Rules
Invasive Breast Carcinoma	Triple Negative	Phenomic Data	ERBB2 Status = ERBB2 Negative Estrogen Receptor Status = Estrogen Receptor Negative Progesterone Receptor Status = Progesterone Receptor Negative
	ER Positive	Phenomic Data	Estrogen Receptor Status = Estrogen Receptor Positive
	ER Positive and HER2 Negative	Phenomic Data	Estrogen Receptor Status = Estrogen Receptor Positive ERBB2 Status = ERBB2 Negative
Gastric Adenocarcinoma	Hyper-Mutator	Oncomine NGS Data	Patient Mutation Count ≥ 400
Lung Adenocarcinoma	KRAS Mutation (No ALK Fusion and No EGFR Mutation)	Oncomine NGS Data	Oncomine Mutation Classification = Hotspot
	Triple Negative	Oncomine NGS Data	No EGFR Mutation (AND) No KRAS Mutation (AND) No ALK Fusion
Rectal Adenocarcinoma	KRAS Mutation	Oncomine NGS Data	Oncomine Mutation Classification = Hotspot
	KRAS Mutation, Stage 3 or 4	Oncomine NGS Data/ Phenomic Data	Oncomine Mutation Classification = Hotspot (AND) Stage = Stage III (OR) Stage IV
	KRAS Wildtype	Oncomine NGS Data	No KRAS Mutation

[00331] Pathways

[00332] Manually curated Compendia pathway definitions were used to summarize gene-level aberrations in the integrative analysis. The pathways represent clinically relevant pathway modules, and several modules may cover a major biological pathway, and a single gene may be present in one or more pathway module definitions.

[00333] Data Integration

[00334] The diagram in Figure 3 summarizes the data flow that integrates the various data types into a Genetic Event Database (GEDB). All further analyses are conducted using the GEDB. The process has 4 main steps.

[00335] Map the data to the internal IA gene and patient dimension

[00336] Define events and driver events in each data type

[00337] Roll-up individual events to the gene and pathway level

[00338] Combine the events into the Genetic Events Database.

[00339] Gene and Patient Dimensions

[00340] A single gene and patient dimension was constructed which encompassed all patients and genes measured across all disease and data types. The genes and patients were given internal identifiers, and all data in the IA was referenced against these identifiers for gene name and patient barcode consistency. The unique identifier for a gene is the gene Entrez ID. The unique identifier for a patient is the TCGA Patient Barcode (first 12 digits of the TCGA barcode).

[00341] Driver Event Definition

[00342] Mutation, fusion and copy number events are defined based on the following criteria for genomic events:

[00343] Fusions: Oncomine recurrent priority fusions

[00344] Mutations: Oncomine driver mutations from pan cancer driver genes

[00345] CNVs: CBI identified peaks, and gene amp/del within peaks

[00346] Genetic Event Definition and Roll-up

[00347] A genetic event is a genomic aberration, representing either an individual mutation, fusion, or copy number event, or a combination of events at the gene or pathway level. The events are 'rolled-up' according to the flowchart shown in Fig. 4. When multiple events are combined to construct rolled up events, the set of measured patients for the rolled up event becomes the intersection of the patients measured for all 3 data types. Patients positive are only included if fully measured.

[00348] Analyses

[00349] Once all the driver genetic events are constructed, a set of analyses is performed on each genetic event, calculating frequencies, associations and relationships within diseases (and pan-cancer where appropriate). The following are short descriptions of each analysis:

[00350] Frequency

[00351] Frequency is the occurrence of a driver event among the patients in which it was measured. Frequencies are calculated within disease and pan-cancer.

[00352] Clinical Association Analysis

[00353] Each driver event is tested for association against a set of available clinical subtypes. Each association is tested using a Fischer's exact test by comparing the occurrences of the genetic event in patients of one clinical subtype versus another. For example a Loss of Function mutation may be tested for over-representation in Smokers versus Non-Smokers, or in Stage I versus Stage II lung cancer. A total of 136 subtype pairs are tested against each event, the properties that define the subtypes are listed below (some properties may be disease-specific). At least 4 patients total, with at least 1 patient in each class are required to perform the test.

[00354] Clinical Outcome Analysis

[00355] Each driver event is tested for association with clinical outcome using log-rank test. Only the set of patients with available clinical data are used for the calculation, so the number of patients included in the test may be less than the number of patients measured for the driver event. At least 4 patients positive for driver event are required to perform the test. Survival time is presented in years, and individual alive/dead events are clearly marked on a Kaplan-Meier curve. P-values were corrected for multiple testing (q-values). Events with a q-value less than 0.1 were considered.

[00356] The results of the analysis are shown in Tables 15 and 39. In Tables 15 and 39, the columns provide the following information:

[00357] The "Subset" column provides the clinically relevant cancer type.

[00358] The p-value column is the p-value.

[00359] The q-value column is the corrected p-value. Events with $q < 0.1$ are included in the table.

[00360] The no. positive column is the number of patients positive for an event type.

[00361] The Total no. of patients column is the total number patients assessed.

[00362] The Cytoband column is the chromosomal location of the gene(s).

[00363] The Genes (Entrez ID) column is a List of gene(s) and corresponding Entrez id.

[00364] The Druggable genes column indicates if any gene(s) are targets for drugs in active trials, approved, or otherwise commercially available.

[00365] The KM Evidence column provides the Kaplan-Meier evidence. The KM evidence indicates if the event type supports good or poor prognosis in the particular cancer type.

[00366] Tables 15 and 39 contains more than 100 gain-of-function mutations, loss-of-function mutations, in-peak gene amplification/deletions, and fusion events for various cancer types with a $q < 0.1$. Gene(s) within each event and cancer type are included along with their chromosomal locations, druggability information and clinical outcome associations, as indicated in the column information above.

[00367] Example 5- Integrated data analysis

[00368] Oncomine NGS Integrated Analysis. The Oncomine NGS Integrative Analysis was designed to bring together the largest possible set of core integrated genomic and phenomic data to enable scientific workflows that interrogate relationships across data types and cancer types, summarizing the analyses at multiple biological levels of abstraction, such as genes and pathways.

Terminology:

[00369] *Aberration* - A genomic structural variation or alteration of DNA; Examples include: mRNA over/under-expression, copy number amplification/deletion, mutation, and gene fusion.

[00370] *Driver* – Aberration identified as a potential cancer driver by Oncomine methodology described in this document; examples include gain of function mutations, gene amplifications in a peak amplification region, or gene fusions

[00371] *Roll-up*- A summary of all mutation, fusion, or copy-number aberrations for the gene or pathway; Only patients measured for all three aberration types are included in the rolled-up.

[00372] *Hotspot Mutation* - A mutation that is recurrent ($n \geq 3$), and classified as either an in-frame insertion/deletion, nonstop or missense.

[00373] *Patient null set* - The set of patients measured for a genetic aberration

[00374] *Patient positive set* - The set of patients harboring the genetic aberration

[00375] *Gene null set* - The set of genes measured by the experimental platform used to assess the genetic aberration

[00376] *Mitelman* - Database of Chromosome Aberrations and Gene Fusions in Cancer manually curated from literature (hypertext transfer protocol:[//goo.gl/PnXMT](http://goo.gl/PnXMT))

[00377] *RPKM* – “Reads Per Kilobase per Million”; a method for RNASeq data quantification that normalizes for total read length and number of sequencing reads (Mortazavi et al. 2008)

[00378] *RSEM* – “RNA-Seq by Expectation Maximization” a method for RNASeq data quantification that estimates the best probable distribution of reads among the expected transcripts provides relative transcript abundances as a fraction of the total read pool. (Li and Dewey 2011)

[00379] **Data Sources.** An effort was made to collect the largest overlapping set of data available for each sample. The data in this release of the NGS Integrative Analysis Browser was obtained from The Cancer Genome Atlas (TCGA), the Cancer Cell Line Encyclopedia (CCLE), COSMIC Cell Lines Project, and a number of research publications, either directly or after being subjected to Oncomine processing and analysis methods. Due to the uneven coverage of all data types across the source datasets, some cancer types have a greater number of patients covered in multiple data types.

[00380] The Oncomine NGS Mutations release used in the Integrative Analysis contained a number of hand-curated datasets obtained from NGS mutation studies in peer-reviewed publications. For a full list of publications that contributed mutation data

to integrative analysis, please see the OncoPrint NGS Mutations methods documentation. The following datasets contained multi-dimensional NGS data, providing both, mutations and copy number data. Copy number data for these datasets was processed in the same way as the copy number data obtained from TCGA.

[00381] Cell line data includes mutation, fusion, and copy number datasets. Cell line data was processed in the same way as the clinical tumor data – with mutation and fusion cell line data obtained from the OncoPrint™ NGS Mutation and OncoPrint™ NGS Fusion Power Tools, respectively. Copy number data for cell lines was processed using the standard OncoPrint copy number pipeline. Although there were two disparate cell line datasets used – CCLE and COSMIC – our standardization of cell line disease types and names has enabled us to cross reference the two datasets and combine the CCLE copy number data, COSMIC mutation data and OncoPrint fusions calls (based on CCLE RNASeq data). Therefore, numerous cell lines in this release have had their exomes systematically characterized for all three types of aberrations. Cell line data was summarized using the OncoPrint cancer type definitions to be directly comparable to tumor data, although the summarization was performed separately for tumor and cell lines.

[00382] Phenomic Data

[00383] Clinical Patient Metadata Curation. Patient clinical data was obtained from primary sources and curated by Compendia. Curated data types include demographics, major clinical and histological disease subtypes, and clinical outcome data. All cancer type-independent properties (such as age or survival) were standardized for consistency across cancer types. Certain disease stages were merged to obtain higher patient counts within a stage. For example, Stage Ia and Ib may be combined as Revised Stage I.

[00384] Following is the list of most populated properties and corresponding values captured by the curation process. Not all properties were available for all patients.

Property Name	Property Value
Age	10 - 14 Years
	15 - 19 Years
	20 - 29 Years
	30 - 39 Years
	30 - 39 Years
	40 - 49 Years
	50 - 59 Years
	60 - 69 Years
	70 - 79 Years
	80 - 89 Years
	90+ Years

Property Name	Property Value
ERBB2 Status	ERBB2 Negative
	ERBB2 Positive
Estrogen Receptor Status	Estrogen Receptor Negative
	Estrogen Receptor Positive
FAB Subtype	FAB Subtype M0
	FAB Subtype M1
	FAB Subtype M2
	FAB Subtype M3
	FAB Subtype M4
	FAB Subtype M5
	FAB Subtype M6
Gleason Score	FAB Subtype M7
	Gleason Score 10
	Gleason Score 6
	Gleason Score 7
	Gleason Score 8
Grade	Gleason Score 9
	Grade 1
	Grade 2
	Grade 3
	Grade 3-4
Hepatitis Virus Infection Status	Grade 4
	Hepatitis B Virus Positive
Human Papillomavirus Infection Status	Hepatitis C Virus Positive
	HPV Negative
	HPV Positive
	HPV Type 16 and 52 Positive
	HPV Type 16 Positive
	HPV Type 45 Positive
Metastatic Event Status	HPV Type 58 Positive
	Metastatic Event
Microsatellite Status	Microsatellite Instable
	Microsatellite Stable
Overall Survival Status	Alive
	Dead
Overall Survival Status (Detailed)	Alive
	Alive With Disease
	Alive Without Disease
	Dead
	Dead With Disease
	Dead Without Disease
Patient Treatment Response	Unknown Therapy Complete Response
	Unknown Therapy Partial Response
	Unknown Therapy Progressive Disease
	Unknown Therapy Stable Disease
Progesterone Receptor Status	Progesterone Receptor Negative
	Progesterone Receptor Positive
Race/Ethnicity	American Indian or Alaska Native
	Asian
	Black or African American
	Hispanic or Latino
	Native Hawaiian or Other Pacific Islander
Recurrence Status	White
	Biochemical Recurrence
	No Biochemical Recurrence
Recurrence Status (Detailed)	Recurrence
	Local Recurrence
	Metastatic Recurrence
Revised M Stage	Recurrence
	M0
Revised N Stage	M1
	N0
	N1
	N2
	N3

Property Name	Property Value
Revised Smoking Status	Never Smoker
	Smoker
Revised Stage	FIGO Stage I
	FIGO Stage II
	FIGO Stage III
	FIGO Stage IV
	Stage I
	Stage II
	Stage III
	Stage IV
Revised T Stage	T
	T0
	T1
	T11
	T12
	T2
	T21
	T22
	T3
	T4
Sex	Female
	Male
*TCGA PAM50 Subtype	Basal-like
	HER2-enriched
	Luminal A
	Luminal B
	Normal-like
*TCGA RPPA Subtype	Basal
	Her2
	Luminal A
	Luminal A/B
	Reactive I Group
	Reactive II Group
*TCGA Subtype	Basal
	CIN
	Classical
	Invasive
	Mesenchymal
	MSI/CIMP
	Neural
	Primitive
	Proneural
	Secretory
Metastatic Event Follow-up Time	(Days)
Overall Survival Follow-up Time	(Days)
Recurrence Follow-up Time	(Days)

[00385] Properties prefixed by “TCGA” were obtained and curated from the TCGA publications that defined the molecular subtypes for invasive breast carcinoma, glioblastoma/squamous cell lung carcinoma and colorectal cancers.

[00386] Genomic Event Data: Fusions Data Filtering. Fusion data for the Integrative Analysis Browser was obtained from Oncomine NGS Fusion data released in November, 2013. Only fusions identified as Oncomine Priority Fusions were included in the Integrative Analysis Browser.

[00387] Oncomine Prioritized Fusion is a priority scheme developed at Compendia to capture attributes of known true positive fusion events and characterize a subset of observed gene fusions as high-confidence priority fusions. Criteria used to define priority fusions include: valid 5' to 3' orientation, non-adjacent fusion partners, uniquely mapping spanning reads, non-paralogous fusion partners, not observed in normal tissue, and non-overlapping with redundant regions in the genome.

[00388] The *patient null set* for the fusion data is the full set of patient tumor samples processed in the fusion analysis; data for only one tumor sample (preferably the primary, non-recurrent tumor) per patient was retained. The *gene null set* is the set of genes in RefGene as of May 2012. Fusions were included in the Integrative Analysis Browser if they were an Oncomine Priority Fusion, had exon expression evidence that was "supported" or "neutral", and met one of the following criteria:

[00389] Recurrent (occurred in 2 or more patients)

[00390] Annotated in the Mitelman database of known structural variations

[00391] Contained a gene partner that is an Oncomine Gain of Function gene that is involved in 3 or more Pan-Disease Priority Fusions

[00392] Contained a gene partner that is a Sanger Oncogene (hypertext transfer protocol:[//goo.gl/JQBw9](http://goo.gl/JQBw9)) that is involved in 3 or more Pan-Disease Priority Fusions.

[00393] Mutation Data Filtering. Mutation data for Integrative Analysis was obtained from Oncomine NGS Mutation data released in November, 2013. Only non-silent mutations in coding gene regions were included in the Integrative Analysis Browser.

[00394] The *patient null set* is the full set of patients processed in the mutation analysis; data for only one tumor sample (preferably the primary, non-recurrent tumor) per patient was retained. The *gene null set* is the set of genes in RefGene as of March 2012.

[00395] Mutations with the following variant classifications were not included in the Integrative Analysis Browser: Silent, 5' UTR, 3' UTR, RNA, Non-Coding Exon.

[00396] Calling Amplifications/Deletions. DNA copy number data for each sample was obtained from the 2013 Q4 Oncomine Standard Data Build, in which all copy number data available from TCGA and the hand-curated publications as of October 2013 was standardized.

[00397] The *patient null set* for this analysis was the set of patients measured for copy number data as of October 2013 and the set of patients measured in the hand-curated publications. Data for only one tumor sample (preferably the primary, non-recurrent tumor) per patient was retained. The *gene null set* for this data was the Oncomine DNA Copy Number platform, based on RefSeq coordinates (hg18) provided by UCSC RefGene build July 2009, and measures 18,796 genes. Measurements from multiple reporters for a single gene were averaged.

[00398] The $\log_2$ of the estimated copy value was used to make amplification/deletion (amp/del) calls, with cutoffs of > 1.0 and < -1.0 , respectively. No amp/del calls were made $\log_2$ (estimated copies) that were ≥ -1.0 or $\leq +1.0$.

[00399] Genomic Continuous Data: Copy Number Segmentation and Quantification. DNA copy number data for each sample was obtained from the 2013 Q4 Oncomine Standard Data Build, in which all copy number data available from TCGA as of September 2013 and all copy number data from the hand-curated publications was standardized.

[00400] The *patient null set* for this analysis was the set of TCGA patients measured for copy number data as of October 2013 and the set of patients measured in the hand-curated publications. Data for only one tumor sample (preferably the primary, non-recurrent tumor) per patient was retained. The *gene null set* for this data was the Oncomine DNA Copy Number platform, based on RefSeq coordinates (hg18) provided by UCSC RefGene build July 2009, and measures 18,796 genes. Measurements from multiple reporters for a single gene were averaged.

[00401] Copy number data was segmented and quantified using the standard Oncomine processing pipeline. Segmentation is a method used to identify contiguous regions of amplification or deletion. These regions or "segments" can include multiple genes or single genes. A copy number value is computed for each segment based on the mean value for the reporters contained in the segment. Genes are mapped to segments and assigned a value. This gene level data is then reported. Please see the Oncomine DNA Processing Pipeline White Paper for more information.

[00402] mRNA Expression Data. Expression data was obtained from the Broad GDAC's TCGA Standard Data build from September, 2013.

[00403] The *patient null set* for this data was the set of patients with available RNASeq data in the Broad GDAC 2013_08_09 stddata build; data for only one tumor

sample (preferably the primary, non-recurrent tumor) per patient was retained. The *gene null set* for this data was different per disease and corresponded to the TCGA Gene Annotation Files (GAFs) used for the RNASeq quantification.

[00404] The TCGA currently employs two methods of RNASeq quantification – V1 (RPKM) and V2 (RSEM) – which are not directly numerically comparable. To avoid a potentially inaccurate numerical conversion, we use data from a single quantification method on a per-disease basis, choosing the format based on maximal coverage. In line with efforts by the TCGA to process (and re-process) all available RNASeq data using RSEM (V2), RSEM (V2) data was available for most samples. An exception is Gastric Adenocarcinoma where RPKM (V1) data was used. Normalized, gene-level quantification values were obtained for both RSEM and RPKM and converted to $\log_2$ values (minimum non-zero RPKM or RSEM values were set at -12). A gene was considered to be expressed if it had a $\log_2$ value > -12 .

[00405] Oncomine Driver Reference Data: Minimum Common Region (MCR) Peak Generation and Gene Selection. In order to identify cancer driver genes subject to amplifications and deletions, a peak-clustering method was performed to select genes frequently aberrant across multiple cancer types. First copy number peaks were defined across the largest-available set of copy number data (i.e. data beyond what is included in the Integrative Analysis) within many cancer types. Next, the gene lists defined by the peaks were clustered in order to identify genes appearing in copy number peaks in multiple samples and multiple diseases. The parts of the method are described in more detail below.

[00406] An aberration may be classified as a “driver” aberration – or one that is considered potentially interesting according to one of the data type-specific Oncomine classification methods. Driver aberrations will be captured as events independently of other aberrations (non-driver aberrations are termed “any”). For example, a patient who has a “driver” mutation will be positive for two aberrations – a “driver” mutation, and an “any” mutation. Each of the measured data types has a set of rules for determining the driver events.

[00407] A set of continuous genomic regions subject to amplification or deletion were identified using the Oncomine MCR analysis by applying it to Oncomine’s 10,249 clinical samples grouped into 25 cancers.

[00408] The *patient null set* for the peak definition was 10,249 clinical samples from Oncomine (See Table below). The *gene null set* for this data was the Oncomine DNA Copy Number platform, based on RefSeq coordinates (hg18) provided by UCSC refGene build July 2009, and measures 18,796 genes.

[00409] Data for the minimal common region (MCR) analysis was sourced from Oncomine DNA copy number browser that contains >20,000 clinical specimens, xenografts and cell lines across diverse cancer types. MCR analysis identifies regions of recurrent copy number amplifications or deletions by analyzing the data at three levels— pan-cancer (across all cancer types), general cancer type (across cancer types), and intermediate cancer type or specific cancer sub-types. Briefly, the method first computes a common region (CR) defined as a contiguous genomic region that is amplified or deleted in 2 or more samples. The minimum thresholds for amplifications and deletions were set at $\log_2 \geq 0.9$ (3.7 copies or more) and $\log_2 \leq -0.9$ (1 copy or less) respectively. Then the peak regions within these common regions are defined as— (i) one or more genes that are aberrant in the highest number of samples (n) and also those that are aberrant in one less than the highest number (n-1) and (ii) genes that are aberrant in 90% of the highest aberrant sample count.

[00410] **Cluster Analysis to Identify Common Peaks Regions.** MCR analysis was performed. Peak regions identified by the MCR analysis were further filtered across the three analysis types (that is, pan-cancer, general cancer type, and specific cancer type analyses) using the criteria listed in table below. Note that only selected number (~40) of intermediate or specific cancer types (also listed further below) were included.

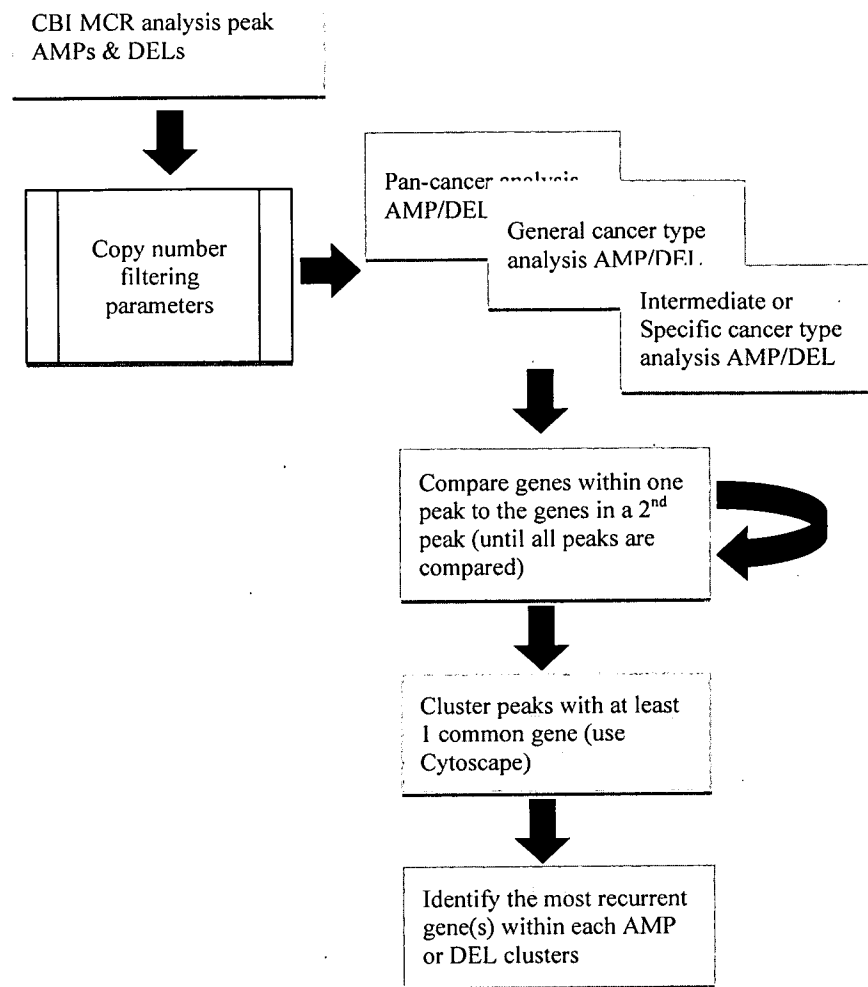
[00411] Filtering criteria to identify highly amplified/deleted regions from MCR analysis:

	Pan-cancer	General cancer type	Intermediate or specific cancer type
Aberrant sample count	≥ 4	≥ 4	≥ 4
Maximum $\log_2$ copy number- Amplifications	≥ 2 (8 or more copies)	≥ 2 (8 or more copies)	≥ 2 (8 or more copies)
Maximum $\log_2$ copy number- Deletions	N/A	≤ -1 (1 or less copies)	≤ -1 (1 or less copies)
Median frequency	$\geq 0.5\%$	$\geq 0.5\%$	$\geq 1.0\%$
Intermediate or specific cancer types	Include all	Include all	Selected ICTs (see Table 6)

[00412] Selected intermediate or specific cancer types included in the filtering criteria described above:

General Cancer Type	Intermediate or specific cancer types
Bladder	Bladder Urothelial Carcinoma
Brain and CNS	Glioblastoma; Medulloblastoma; Neuroblastoma
Breast	N/A
Cervical	Cervical Adenocarcinoma; Cervical Squamous cell carcinoma
Colorectal Cancer	Colorectal Adenocarcinoma
Esophageal	Esophageal Adenocarcinoma; Esophageal squamous cell carcinoma
Gastric	Gastric Adenocarcinoma
Head and Neck	Head-Neck Squamous Cell Carcinoma; Thyroid gland follicular carcinoma; Thyroid Gland Papillary Carcinoma
Kidney	Clear Cell Renal Cell Carcinoma; Papillary Renal Cell Carcinoma
Leukemia	Acute Lymphoblastic Leukemia; Acute Myeloid Leukemia; Chronic Lymphocytic Leukemia; Chronic Myelogenous Leukemia; Myelodysplastic Syndrome
Liver	Hepatocellular Carcinoma
Lung Cancer	Lung Adenocarcinoma; Small Cell Lung Carcinoma; Squamous Cell Lung Carcinoma
Lymphoma	Burkitt's Lymphoma; DLBCL; Follicular Lymphoma; Hodgkin's Lymphoma; Mantle Cell Lymphoma;
Melanoma	Cutaneous Melanoma; Multiple Myeloma
Other	Endometrial Endometrioid Adenocarcinoma
Ovarian	Ovarian Clear Cell Adenocarcinoma; Ovarian Serous Adenocarcinoma
Pancreas	Pancreatic Ductal Adenocarcinoma; Prostate Adenocarcinoma
Sarcoma	GIST

[00413] Next, to identify the most recurrent peak regions and genes across multiple cancer types we used Cytoscape 2.8.3 [Markiel et al. 2003; Smoot et al. 2001] to build network clusters. Briefly, the analysis compares every gene in a given peak region to genes in other peak regions and clusters peaks with at least one common gene. The most recurrent amplified or deleted gene(s) within each cluster was then considered as a potential candidate driver gene. The process is outlined in the schematic below:



[00414] Identification and Creation of Clinically Relevant Cancer Subtypes. In order to provide subsets of patients for more focused analysis, several clinically relevant cancer subtypes were identified and curated using a combination of clinical phenomic, and categorical genomic data. The phenomic data was sourced from the TCGA Web Portal or the Supplementary Methods of the hand-curated publications.

[00415] The following rules were applied to define the Clinical Subsets:

Cancer Type	Clinical Subtype	Data Source	Data Interpretation Rules for Inclusion in Subtype
Invasive Breast Carcinoma	Triple Negative	TCGA Web Portal	ERBB2 Status = "ERBB2 Negative" Estrogen Receptor Status = "Estrogen Receptor Negative" Progesterone Receptor Status = "Progesterone Receptor Negative"
	ER Positive	Phenomic	Estrogen Receptor Status = "Estrogen Receptor Positive"
	HER2 Positive	Phenomic	ERBB2 Status = "ERBB2 Positive"
	ER Positive and HER2 Positive	Phenomic	Estrogen Receptor Status = "Estrogen Receptor Positive" ERBB2 Status = "ERBB2 Positive"

Cancer Type	Clinical Subtype	Data Source	Data Interpretation Rules for Inclusion in Subtype
	ER Positive and HER2 Negative	Phenomic	Estrogen Receptor Status = "Estrogen Receptor Positive" ERBB2 Status = "ERBB2 Negative"
Gastric Adenocarcinoma	Diffuse	Phenomic	Cancer Type = "Diffuse Gastric Adenocarcinoma"
	Intestinal	Phenomic	Cancer Type = "Gastric Intestinal Type Adenocarcinoma"
	Hyper-Mutator	Oncomine NGS	Patient Mutation Count ≥ 400
Head and Neck Squamous Cell Carcinoma	HPV Positive	Phenomic	Human Papillomavirus Infection Status = "HPV Positive"
	HPV Negative	Phenomic	Human Papillomavirus Infection Status = "HPV Negative"
Cervical Squamous Cell Carcinoma	HPV Positive	Phenomic	Human Papillomavirus Infection Status = "HPV Positive"
Lung Adenocarcinoma	EGFR Mutation	Oncomine NGS Mutation	Oncomine Mutation Classification = "Hotspot"
	KRAS Mutation (No ALK Fusion and No EGFR Mutation)	Oncomine NGS Mutation and Fusion	Oncomine Mutation Classification = "Hotspot"
	ALK Fusion	Oncomine NGS Fusion	Have Oncomine Driver ALK fusions
	Triple Negative	Oncomine NGS Mutation and Fusion	No EGFR Mutation AND No KRAS Mutation AND No ALK Fusion
Colon and Rectal Adenocarcinoma	KRAS Mutation	Oncomine NGS Mutation	Oncomine Mutation Classification = "Hotspot"
	KRAS Mutation, Stage 3 or 4	Oncomine NGS Mutation and Phenomic	Oncomine Mutation Classification = "Hotspot" AND Stage = "Stage III" OR "Stage IV"
	KRAS Wildtype	Oncomine NGS Mutation and Fusion	Oncomine Mutation Classification = "Hotspot" AND No KRAS Fusion
	Microsatellite Stable	Phenomic	Microsatellite Status = "Microsatellite Stable"
	Microsatellite Unstable	Phenomic	Microsatellite Status = "Microsatellite Unstable"
Endometrial Endometrioid Carcinoma	Microsatellite Stable	Phenomic	Microsatellite Status = "Microsatellite Stable"
	Microsatellite Unstable	Phenomic	Microsatellite Status = "Microsatellite Unstable"

[00416] Pathways. Manually curated Compendia pathway definitions were used to summarize gene-level aberrations in the Integrative Analysis Browser. The pathways represent clinically relevant pathway modules, and several modules may cover a major biological pathway. A single gene may be present in one or more pathway definitions, but care was taken to eliminate largely redundant pathways, in which one module is a

complete subset of another. There are 67 total pathways, ranging in size from 42 genes (e.g. MAPK pathway) to 2 genes (e.g. IGF1/IGF1R and several others).

[00417] Data Integration. This section summarizes the data flow that integrates the primary data onto common patient and gene dimensions and constructs the Genetic Event Database (GEDB), which is comprised of all the aberrations which will be subject to Integrative Analyses. The process has 4 main steps: (1) Integrate primary data using universal gene and patient dimensions, (2) Call aberration events for each data type and define driver aberrations (3) Roll-up individual events to the gene and pathway level and integrate events, and (4) Construct the Genetic Event Database by defining patient status for each event.

[00418] Constructing and Mapping to the Gene and Patient Dimensions. The varied data types included in the Integrative Analysis may have been measured on different experimental platforms and on sets of patients that are not perfectly overlapping. Therefore, care was taken to include all patients and genes measured while avoiding duplicate or conflicting entries.

[00419] For each data type, a gene and patient “dimension” was constructed, enumerating the genes and patients measured in the data. The dimension for each data type may be different, as indicated by the patient dimension overlap diagram below (numbers for illustration only), in this case, for Invasive Breast Carcinoma.

[00420] Gene and patient dimensions were gathered from each cancer and data type, and a non-redundant superset of all the patient and gene identifiers in the Integrative Analysis was constructed. The bars in the figure below represent blocks of patient identifiers (if sorted by said identifier) for patients measured for the certain aberration types.

[00421] Redundancy for patients was determined based on the *unique patient identifier* – currently the first 12 characters of the TCGA Tumor Sample Barcode (e.g., TCGA-AB-1234).

[00422] Redundancy for genes was determined based on the *unique gene identifier* – currently the Entrez Gene ID. The genes were also compared against the Oncomine gene set, and when a gene symbol conflict was found – one Entrez ID assigned two or more gene symbols – the gene symbol from Oncomine was used. Several (12) of the measured genes contained Entrez ID that have been discontinued and thus may not

represent the most accurate gene model. The gene symbols for these genes were marked with the word “discontinued”.

[00423] Once constructed, the non-redundant gene and patient dimensions were indexed to provide a consistent internal identifier for each gene and patient in the dataset. All the unique patient and gene identifiers in the primary data were then mapped to the dimension patient and gene identifiers. Gene and patient metadata, such as gene symbols and patient clinical data, are thus always mapped through the respective dimensions, providing consistency in naming and annotation. The total number of unique genes and patients in the Integrative Analysis is as follows:

[00424]	Genes	23,340
	Patients	11,476

[00425] The patient dimension along with the dataset-specific mapping of the patients helps correctly identify fully wild-type patients – those who are measured for all aberration types but do not contain any aberrations.

[00426] A patient could thus be measured for any number of aberrations, and can only be aberrant for those events measured. The aberrations a patient is measured for determined the types of analyses that patient would be included in:

	Patient “X” Measured for:	Patient “X” Excluded from:	Patient “Y” Measured for:	Patient “Y” Excluded from:
Clinical*		Clinical Associations, Clinical Outcome Associations	✓	DNA vs. RNA Correlation, Differential Expression, Expressed Frequency
Mutations	✓		✓	
Fusions	✓		✓	
Copy Number	✓		✓	
Expression	✓			

*“Measured for : Clinical” indicates clinical metadata was present for patient.

[00427] Event Model. Each genomic aberration from the mutation, fusion, and copy number data sets was identified as an aberration event – a term used to define an event of interest that will be subject to the various pre-defined Integrative Analyses. Each aberration is part of three broad levels of events – data type-specific events, gene-

specific but data type independent events, and pathway-specific but gene or data type independent events. The latter two levels are considered “rolled-up” events.

[00428] The specific rules for aberration event definition as well as the “level” and “driver” schemes are described below.

[00429] Mutation Event Caller. Oncomine Pan-Cancer Mutation Classification: A mutation is classified as a “Hotspot” if it is: Recurrent (occurs in 3 or more samples at the same amino acid position) AND Annotated with one of the following variant classifications: In-Frame insertion/deletion, Nonstop, Missense. A mutation is classified as “Deleterious” if it is: Not recurrent AND Annotated with one of the following variant classifications: Frame-Shift insertion/deletion, Nonsense. Recurrence is measured across all cancer types analyzed as part of the Oncomine NGS Mutation Browser.

[00429] Oncomine Pan-Cancer Gene Classification. As part of the Oncomine NGS Mutation Browser pipeline, individual genes are classified into predicted functional classes, namely “Gain of Function” and “Loss of Function” to reflect their relative enrichment in potential activating or deleterious mutations. This classification is based on the combination of relative frequencies and the significance of the mutations observed in the gene assessed by a p-value. A “Gain of Function” gene will have a relatively high frequency of hotspot (recurrent in 3 or more samples) missense mutations and a low frequency of deleterious mutations, while a “Loss of Function” gene contains a large fraction of deleterious mutations. Pan-cancer gene classifications are based on the mutations observed across all cancer types.

[00430] Mutation Aberration Events. For each patient gene mutation (as defined by the Mutation Data Filtering section), either one or two mutation events will be created, depending on whether the mutation is classified as a *driver* aberration. A *driver* mutation aberration is defined as a “Hotspot” mutation present in a “Gain of Function” gene, or a “Hotspot” or “Deleterious” mutation present in a “Loss of Function” gene. For *driver* mutations, both a *driver* event and an *any* event is created. For non-driver mutations, only an *any* event is created. Pan-Cancer mutation and gene classification was used for all analysis subsets; so, gene classifications may sometimes differ between Integrative Analysis and Oncomine NGS Mutation Browser.

[00431] The table below gives the description and examples of mutation events that could be created for each gene:

Event Type	Description of Event	Example Aberration Name	Driver Status
Gain of Function Mutation	A "Hotspot" mutation and a "Gain of Function" gene classification	EGFR Gain of Function Mutation	driver
Loss of Function Mutation	A "Hotspot" or "Deleterious" mutation in a "Loss of Function" gene	APC Loss of Function Mutation	driver
Any Gene Mutation	Any mutation in a gene	TTN <Any Gene Mutation>	any

[00432] Fusion Event Caller. Only Oncomine Priority fusions are included in the Integrative Analysis. Of the Priority Fusions, the driver fusions were defined as those labeled known oncogenes by the Mitelman database OR fusions that either did not have sufficient exon expression data and are recurrent, OR fusions that have exon expression data and a significant p-value for exon expression imbalance of the two gene partners (See Oncomine NGS Fusions Methods Documentation for details of exon imbalance classification). For each gene, an event will be created for each unique observed 5'-3' combination of the gene partners. For example, for PML-RARA balanced translocation both isoforms are observed and hence two fusion events will be called - for PML-RARA and RARA-PML respectively.

Event Type	Description of Event	Example Aberration Name	Driver Status
Fusion	Driver fusion involving gene	PML-RARA Fusion	driver
Any Fusion	Any fusion involving gene	FRS2-LYZ Fusion	any

[00433] Copy Number Event Caller. Each Amp/Del (see Calling Amp/Dels) that was called was defined as an *any* event for the aberrant gene. If the amp/del occurred in a gene that was part of a peak definition (see MCR Peak Generation) in a certain cancer type, a *driver* ampdel event was also created for that gene. The driver definition for copy number events is thus cancer type specific.

[00434] The following are the copy number aberration events that maybe be called for a gene amplification or deletion:

Event Type	Description of Event	Example Aberration Name	Driver Status
In-Peak Gene Amplification	An amplification in a gene observed in an Amplification Peak within the same cancer type	EGFR In-Peak Gene Amplification	driver
In-Peak Gene Deletion	A deletion in a gene observed in a Deletion Peak within the same cancer type	CDKN1A In-Peak Gene Deletion	driver
Any Gene Amplification	An amplification in a gene	ERBB2 <Any Gene Amplification>	any
Any Gene	A deletion in a gene	FGFR <Any Gene Deletion>	any

Event Type	Description of Event	Example Aberration Name	Driver Status
Deletion			

[00435] Genetic Event Roll-Up. Both *driver* and *any* events are “rolled-up” to gene-level and pathway-level events to capture a data type-independent aberration statistics and associations. For example, it may be interesting to see the association of any aberrations in a tumor suppressor gene with clinical outcome; not just the association of the deleterious mutations.

[00436] A gene-level aberration event is created for each gene that has at least one aberration of any data type. A pathway-level aberration event is created for each pathway in which at least one of the component genes has an aberration of any data type. *Driver* and *any* aberrations are rolled-up independently into gene-level or pathway-level *driver* or *any* events. The diagram below shows the hierarchical relationships between the various aberration event types.

[00437] Patient Event Status. A patient can be measured for any number of aberrations but can only be aberrant for those events measured. Patient status for each event-level aberration is thus recorded as aberrant, wild type, or not measured.

[00438] The patient dimension along with the data set-specific mapping of the patients helps correctly identify fully wild-type patients – those who are measured for all aberration types but don't contain any aberrations.

[00439] When gene-level and pathway-level events are defined, only the patients measured for all 3 genetic data types – mutations, fusion, and copy number – are marked as “aberrant” or “wild type” for the event. This assumption has the effect of potentially reducing the number of patients summarized for a gene or pathway-level aberrations as compared to the data type-specific event-level aberrations. A patient is considered aberrant for a gene-level event if the patient is aberrant for at least one of the event-level aberration types (Fusion, Mutation, Amplification, or Deletion). A patient is considered aberrant for a pathway-level event if the patient has an aberration in at least one gene that is part of the pathway definition. In each case, the patient must have been measured for all the event types.

[00440] In the case of the Gain of Function and Fusion, the aberration frequency is ~50%. For the Driver Gene Aberration event the aberration frequency is also ~50% but only half as many patients are included in the numerator and denominator of the frequency.

[00441] Analysis. Once all the driver genetic events are constructed, a set of analyses is performed on each genetic event, calculating frequencies, associations, and relationships within cancer types, clinically relevant subtypes, and among cancer types (pan-cancer). The following are short descriptions of each analysis, including which data is used, and what constraints, if any, are put on the reported results: frequency, expressed frequency, co-occurrence and mutual exclusivity, clinical association analysis, etc.

[00442] Frequency. Frequency is the occurrence of an aberration among the patients in which it was measured. Frequencies are calculated within cancer types, clinically relevant cancer subtypes, and pan-cancer. All events with at least one aberrant patient are reported.

[00443] Expressed Frequency. Expressed frequency is the frequency at which the gene(s) is expressed among the event-positive patients. For each event, expression level of the gene(s) is used to ascertain the expressed frequencies. Frequencies are calculated within cancer types and clinically relevant cancer subtypes, but not pan-cancer.

[00444] Co-Occurrence and Mutual Exclusivity. Co-occurrence and mutual exclusivity is calculated for each pair of events using a Fischer's Exact test. At least 2 patients positive for each event and 5 patients measured for the events in total are required for the calculation. Co-occurrence or exclusivity of two individual copy number deletion or amplification events is not calculated. Also, co-occurrence and mutual exclusivity is not calculated between pairs of events with "any" driver status (i.e. only drivers vs. drivers and drivers vs. any are compared). Associations are calculated within cancer types and clinically relevant cancer subtypes, but not pan-cancer.

[00445] Clinical Association Analysis. Each driver event is tested for association against a set of available clinical subtypes. Each association is tested using a Fischer's exact test by comparing the occurrences of the genetic event in patients of one clinical subtype versus another. For example, a Loss of Function mutation may be tested for over-representation in Smokers versus Non-Smokers, or in Stage I versus Stage II lung cancer. A total of 136 subtype pairs are tested against each event, and the properties that define the subtypes are listed below (some properties may be disease-specific). At least 4 patients total, with at least 1 patient in each class are required to perform the test. Associations are calculated within cancer types, clinically relevant cancer subtypes, and pan-cancer.

[00446] Clinical Subtype Property Names:

Race/Ethnicity
Revised Smoking Status
ERBB2 Status
Estrogen Receptor Status
Progesterone Receptor Status
TCGA PAM50 Subtype
BRAF Mutation Status
Revised T Stage
Revised N Stage
Revised M Stage
Revised Stage
KRAS Mutation Status
EGFR Amplification Status
TCGA Subtype
Microsatellite Status
Human Papillomavirus Infection
Status

[00447] Clinical Outcome Analysis. Each event is tested for association with clinical outcome using the Logrank test. Only the set of patients with available clinical data are used for the calculation, so the number of patients included in the test may be less than the number of patients measured for the driver event. At least 4 patients aberrant for an event are required to perform the test. Survival time is presented in years, and individual alive/dead events are clearly marked on a Kaplan-Meier curve. Associations are calculated within cancer types and clinically relevant cancer subtypes, but not pan-cancer.

[00448] DNA-RNA Correlation Analysis. For each gene, the RNA expression and DNA copy number values are tested for correlation among all patients within a disease who were measured for these data types using Pearson's correlation. Correlations are calculated within cancer types and clinically relevant cancer subtypes, but not pan-cancer.

[00449] Differential Expression Analysis. For each event, each gene associated with the event was tested for differential expression in event-positive patients vs. event-negative patients using Student's T-Test. For events involving several genes – such as fusions – each gene was tested. Differential expression is calculated within cancer types and clinically relevant cancer subtypes, but not pan-cancer.

Table 15: Events associate with cancer prognosis

Table 15	Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
	Hepatocellular Carcinoma	In-Peak Gene Amplification	3.31E-02	9.93E-02	4	65	1q21.2	ADAMTSL4 (54507), MCL1 (4170)	MCL1	Poor prognosis
	Hepatocellular Carcinoma	In-Peak Gene Deletion	2.47E-02	9.89E-02	4	65	13q14.2	LPAR6 (10161)	N	Poor prognosis
	Squamous Cell Lung Carcinoma	Loss of Function Mutation	1.60E-02	9.59E-02	7	175	4q31.3	FBXW7 (55294)	N	Poor prognosis
	Squamous Cell Lung Carcinoma	Loss of Function Mutation	3.14E-02	9.42E-02	7	175	9q34.3	NOTCH1 (4851)	NOTCH1	Poor prognosis
	Squamous Cell Lung Carcinoma	Loss of Function Mutation	7.73E-03	9.28E-02	5	175	1p35.3	ARID1A (8289)	N	Poor prognosis
	Clear Cell Renal Cell Carcinoma	In-Peak Gene Deletion	7.12E-03	9.25E-02	8	493	9p21	CDKN2B (1030)	No	Poor prognosis
	Invasive Breast Carcinoma:ER Positive	In-Peak Gene Amplification	2.17E-03	9.13E-02	15	635	17q11.2	TIAF1 (9220), MYO18A (399687), CRYBA1 (1411)	N	Poor prognosis
	Ovarian Serous Cystadenocarcinoma	In-Peak Gene Amplification	1.00E-03	8.99E-02	10	557	19q13.1-q13.2	PSG2 (5670), PSG5 (5673), CEACAM1 (634), CEACAM8 (1088), CXCL17 (284340), RABAC1 (10567), ATP1A3 (478)	CEACAM1 (Preclinical)	Poor prognosis

Table 15									
Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
Clear Cell Renal Cell Carcinoma	Loss of Function Mutation	2.44E-02	8.55E-02	14	293	3p21	BAP1 (8314)	No	Poor prognosis
Ovarian Serous Cystadenocarcinoma	In-Peak Gene Amplification	5.45E-04	8.39E-02	89	557	19q12	C19orf2 (8725)	N	Poor prognosis
Lung Adenocarcinoma	In-Peak Gene Amplification	6.80E-03	8.16E-02	4	320	1q12	CHD1L (1105)	N	Poor prognosis
Lung Adenocarcinoma	In-Peak Gene Amplification	6.80E-03	8.16E-02	4	320	1q21.1	FMO5 (2330), PRKAB2 (5565)	N	Poor prognosis
Lung Adenocarcinoma	In-Peak Gene Amplification	8.57E-03	7.71E-02	9	320	12p12.1	KRAS (3845), CASC1 (55259), LYRM5 (144363), LRMP (4033)	KRAS (Preclinical)	Poor prognosis
Invasive Breast Carcinoma: Triple Negative	In-Peak Gene Amplification	8.10E-03	7.29E-02	5	88	8p12	BRF2 (55290), ERLIN2 (11160), GPR124 (25960), PROSC (11212), RAB11FIP1 (80223), ZNF703 (80139)	N	Poor prognosis
Head and Neck Squamous Cell Carcinoma	In-Peak Gene Amplification	1.02E-02	6.93E-02	8	316	5q35	THOC3 (84321)	No	Poor prognosis
Rectal Adenocarcinoma	In-Peak Gene Deletion	2.08E-03	6.86E-02	4	145	16p13.3	A2BP1 (54715)	N	Poor prognosis
Lung Adenocarcinoma	In-Peak Gene Deletion	9.37E-03	6.09E-02	5	320	3q25.1	AADAC (13)	N	Poor prognosis

Table 15	Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
Hepatocellular Carcinoma		In-Peak Gene Deletion	3.03E-02	6.06E-02	4	65	8p21.2	GNRH1 (2796)	GNRH1	Poor prognosis
			5.58E-04	6.02E-02	22	557	20q11	ID1 (3397), BCL2L1 (598), COX4I2 (84701)	ID1 (Preclinical), BCL2L1	Poor prognosis
Rectal Adenocarcinoma:KRAS Wildtype		Loss of Function Mutation	2.69E-02	5.38E-02	11	28	5q21-q22	APC (324)	N	Poor prognosis
Papillary Renal Cell Carcinoma		In-Peak Gene Amplification	2.68E-02	5.35E-02	6	100	17q21.1	CCL3L3 (414062), CCL3L1 (6349)	N	Poor prognosis
Acute Myeloid Leukemia		PML+RARA Fusion	1.26E-02	5.03E-02	15	169	17q and 15q	RARA (5914), PML (5371), ACOT8 (10005), ADA (100), C20orf111 (51526), C20orf123 (128506), C20orf165 (128497), CD40 (958), CDH22 (64405), CTSA (5476), DBNDD2 (55861), DNTTIP1 (116092), ELMO2 (63916), FITM2 (128486), GDAP1L1 (78997), GTSF1L (149699), HNF4A (3172), IFT52 (51098), JPH2 (57158), KCNK15 (60598), KCNS1 (3787), L3MBTL (26013), MATN4 (8785), MMP9 (4318), MYBL2 (4605), NCOA5 (57727), NEURL2 (140825), PABPC1L	Y	Good prognosis
Rectal Adenocarcinoma:KRAS Wildtype		In-Peak Gene Amplification	4.55E-02	4.96E-02	10	27	20q	ADA, CD40 (958), MMP9, PI3		Poor prognosis

Table 15	Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
								(80336), PCIF1 (63935), PI3 (5266), PIGT (51604), PKIG (11142), PLAGL2 (5326), PLTP (5360), POFUT1 (23509), R3HDM1 (140902), RBPJL (11317), RIMS4 (140730), SDC4 (6385), SEMG1 (6406), SEMG2 (6407), SERINC3 (10955), SFRS6 (6431), SGK2 (10110), SLC12A5 (57468), SLC13A3 (64849), SLC35C2 (51006), SLPI (6590), SNAIL (6615), SNX21 (90203), SPINLW1 (57119), SPINT3 (10816), SPINT4 (391253), STK4 (6789), SYS1 (90196), TM9SF4 (9777), TNNC2 (7125), TOMM34 (10953), TOX2 (84969), TP53RK (112858), TP53TG5 (27296), TTPAL (79183), UBE2C (11065), WFDC10A, (140832), WFDC10B, (280664), WFDC11, (259239), WFDC12, (128488), WFDC13, (164237), WFDC2 (10406), WFDC3 (140686), WFDC5 (149708), WFDC6 (140870), WFDC8 (90199), WFDC9 (259240), WISP2 (8839), YWHAB (7529), ZNF334 (55713), ZNF335 (63925), ZSWIM1 (90204), ZSWIM3 (140831)		

Table 15									
Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
Gastric Adenocarcinoma	Loss of Function Mutation	4.09E-03	4.50E-02	4	131	6p21	HLA-B (3106)	Yes	Poor prognosis
Endometrial Endometrioid Adenocarcinoma	In-Peak Gene Deletion	1.36E-02	4.43E-02	6	446	16Q24	SLC7A5 (8140), CTU2 (66965), FAM38A (9780), CDT1 (81620), APRT (353), GALNS (2588)	SLC7A5 (preclinical)	Poor prognosis
Lung Adenocarcinoma	In-Peak Gene Deletion	3.15E-03	4.09E-02	8	320	19q13.4	KIR2DS4 (3809)	N	Poor prognosis
Head and Neck Squamous Cell Carcinoma	In-Peak Gene Amplification	1.45E-03	3.82E-02	6	316	20p12	C20orf94 (128710), JAG1 (182), MKKS (8195), SNAP25 (6616)	JAG1 (Preclinical)	Poor prognosis
Lung Adenocarcinoma: Triple Negative	In-Peak Gene Amplification	8.80E-04	3.78E-02	6	174	7q31	MET (4233), CAPZA2 (830)	MET	Poor prognosis
Endometrial Endometrioid Adenocarcinoma	In-Peak Gene Deletion	1.09E-02	3.77E-02	4	446	3Q26	APOD (347)	No	Poor prognosis
Cutaneous Melanoma	Loss of Function Mutation	3.74E-03	3.74E-02	16	148	17q11	NF1 (4763)	No	Poor prognosis
Acute Myeloid Leukemia	CBFB+MYH11 Fusion	1.83E-02	3.67E-02	11	169	16Q22 and 16P13.11	CBFB (865), MYH11 (4629)	N	Good prognosis
Head and Neck Squamous Cell Carcinoma	In-Peak Gene Amplification	6.01E-04	2.86E-02	5	316	7p12	ABCA13 (154664), C7orf57 (136288), C7orf65 (401335), C7orf69 (80099), C7orf72 (100130988), DDC (1644), FIGNL1 (63979), GRB10 (2887)	No	Poor prognosis

Table 15									
Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
							HUS1 (3364), IKZF1 (10320), PKDIL1 (168507), SUN3 (256979), TNS3 (64759), UPP1 (7378), VWC2 (375567), ZPBP1 (11055)		
Lung Adenocarcinoma	In-Peak Gene Amplification	1.28E-03	2.76E-02	7	320	7q31	MET (4233), CAPZA2 (830)	MET	Poor prognosis
Head and Neck Squamous Cell Carcinoma	In-Peak Gene Amplification	3.00E-03	2.31E-02	80	316	11q13	FADD (8772), PPPIA1 (8500), ANO1 (55107), CTTN (2017)	No	Poor prognosis
Gastric Adenocarcinoma	In-Peak Gene Amplification	1.89E-04	2.14E-02	4	172	18q11	GATA6 (2627)	No	Poor prognosis
Invasive Breast Carcinoma	In-Peak Gene Deletion	2.27E-03	1.82E-02	11	863	10q23.31,q23.2	ATAD1 (84896), KILLIN (100144748)	N	Poor prognosis
Head and Neck Squamous Cell Carcinoma	In-Peak Gene Amplification	1.89E-03	1.55E-02	6	316	2q32	GLS (2744), MYO1B (4430), NAB1(4664),STAT1 (6772), STAT4 (6775), TMEM194B (100131211)	No	Poor prognosis
Colon Adenocarcinoma	In-Peak Gene Deletion	2.27E-04	1.48E-02	4	412	3q26	APOD (347)	No	Poor prognosis
Gastric Adenocarcinoma: Hyper-Mutator	Loss of Function Mutation	5.32E-04	1.22E-02	4	32	2q31	HOXD8 (3234)	No	Poor prognosis

Table 15									
Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
Glioblastoma	Loss of Function Mutation	1.23E-03	1.11E-02	6	276	Xq25	STAG2 (10735)	No	Poor prognosis
Head and Neck Squamous Cell Carcinoma	Gain of Function Mutation	2.61E-03	1.04E-02	13	304	2q31	NFE2L2 (4780)	NO	Poor prognosis
Endometrial Endometrioid Adenocarcinoma	In-Peak Gene Amplification	9.20E-04	1.03E-02	7	446	1q21	SSR2 (6746), ARHGEF2 (9181), UBQLN4 (56893)	No	Poor prognosis
Endometrial Endometrioid Adenocarcinoma	In-Peak Gene Deletion	2.47E-03	9.17E-03	7	446	16p13	LOC339047 (339047)	No	Poor prognosis
Hepatocellular Carcinoma	In-Peak Gene Amplification	2.57E-03	8.89E-03	4	65	1q21.3	DCST1 (149095), ADAM15 (8751), EFNA4 (1945), EFNA3 (1944), EFNA1 (1942), RAG1API (55974), DPM3 (54344), KRTCAP2 (200185), TRIM46 (80128), MUC1 (4582), THBS3 (7059), MTX1 (4580), GBA (2629)	ADAM15, MUC1	Poor prognosis
Clear Cell Renal Cell Carcinoma	In-Peak Gene Deletion	3.16E-04	8.23E-03	8	493	9p21	CDKN2A (1029)	Yes	Poor prognosis
Glioblastoma	Gain of Function Mutation	2.72E-03	8.15E-03	14	276	2q33	IDH1 (3417)	preclinical	Good prognosis
Lung Adenocarcinoma: KRAS Mutation (No ALK Fusion and No EGFR Mutation)	In-Peak Gene Amplification	2.56E-03	5.98E-03	4	78	12p12.1	LYRM5 (144363), KRAS (3845), CASC1 (55259)	KRAS (Preclinical)	Poor prognosis

Table 15									
Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
Endometrial Endometrioid Adenocarcinoma	In-Peak Gene Amplification	4.40E-04	5.55E-03	6	446	1q22	ROBLD3 (28956), RAB25 (57111), MEX3A (92312)	No	Poor prognosis
Colon Adenocarcinoma:KRAS Mutation	Gain of Function Mutation	4.97E-03	4.97E-03	17	53	3q26	PIK3CA (5290)	Yes	Poor prognosis
Head and Neck Squamous Cell Carcinoma	Loss of Function Mutation	1.79E-04	3.95E-03	161	304	17p13	TP53 (7157)	TP53	Poor prognosis
Head and Neck Squamous Cell Carcinoma	In-Peak Gene Amplification	6.41E-05	3.81E-03	4	316	22q11	CRKL (1399), PI4KA (5297), SERPIND1 (3053), SNAP29 (9342)	No	Poor prognosis
Gastric Adenocarcinoma	Loss of Function Mutation	2.14E-04	3.53E-03	5	131	17q22	RNF43 (54894)	No	Poor prognosis
Lower Grade Glioma	Loss of Function Mutation	3.00E-04	2.70E-03	5	166	17q11.2	NF1 (4763)	N	Poor prognosis
Lung Adenocarcinoma:Triple Negative	Gain of Function Mutation	5.06E-04	2.53E-03	11	175	3q26.3	PIK3CA (5290)	Y	Poor prognosis
Lung Adenocarcinoma	Loss of Function Mutation	5.24E-05	9.96E-04	4	283	5q21-q22	APC (324)	N	Poor prognosis
Bladder Urothelial Carcinoma	In-Peak Gene Amplification	8.34E-05	9.31E-04	5	125	5p15.33	PLEKHG4B (153478), LRRCL4B (389257), CCDC12 (151903), SDHA (6389), PDCD6 (10016), AHRH (57491), C5orf55 (116349), EXOC3 (11336),	AHRH, TERT	Poor prognosis

Table 15	Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
								SLC9A3 (6550), CEP72 (55722), TPPP (11076), BRD9 (65980), TRIP13 (9319), NKD2 (85409), SLC12A7 (10723), SLC6A19 (340024), SLC6A18 (348932), TERT (7015), CLPTMIL (81037), SLC6A3 (6531), LPCAT1 (79888), MRPL36 (64979), NDUFS6 (4726)		
	Endometrial Endometrioid Adenocarcinoma	In-Peak Gene Deletion	1.12E-04	8.32E-04	9	446	10q23	PTEN (5728), ANKRD22 (118932), STAMBPL1 (57559), ACTA2 (59), FAS (355), ATAD1 (84896), KILLIN (100144748), RNLS (55328)	Yes	Poor prognosis
	Lower Grade Glioma	In-Peak Gene Amplification	5.69E-04	6.74E-04	5	206	1q32.1	Clorf157 (284573), ETNK2 (55224), GOLTA1 (127845), KISS1 (3814), LAX1 (54900), LRRN2 (10446), MDM4 (4194), PIK3C2B (5287), PLEKHA6 (22874), PPP1R15B (84919), REN (5972), SNRPE (6635), SOX13 (9580), ZC3H11A (9877)	MDM4 (Preclinical)	Poor prognosis
	Ovarian Serous Cystadenocarcinoma	In-Peak Gene Deletion	1.05E-06	6.28E-04	7	557	9q22	FAM75C1 (441452)	N	Poor prognosis
	Endometrial Endometrioid Adenocarcinoma	In-Peak Gene Amplification	8.93E-06	3.01E-04	25	446	8q24	MYC (4609), TAF2 (6873), DSCC1 (79075), DEPC6 (64798)	No	Poor prognosis
	Acute Myeloid Leukemia	Loss of Function Mutation	2.35E-05	9.42E-05	12	184	17p	TP53 (7157)	Y	Poor prognosis

Table 15	Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
	Colon Adenocarcinoma	In-Peak Gene Amplification	3.93E-06	6.24E-05	7	412	12p13	CCND2 (894), TULP3 (7289), TEAD4 (7004), TSPAN9 (10867), PRMT (563418), EFCAB4B (84766), PARP11 (57097), C12orf5 (57103), FGF23 (8074), FGF6 (2251), FKBP4 (2288), ITFG2 (55846), NRIP2 (83714), FOXM1 (2305)	No	Poor prognosis
	Gastric Adenocarcinoma	Loss of Function Mutation	8.74E-07	2.88E-05	4	131	2q31	HOXD8 (3234)	No	Poor prognosis
	Lower Grade Glioma	Gain of Function Mutation	9.38E-08	2.81E-07	130	166	2q33.3	IDH1 (3417)	IDH1 (Preclinical)	Good prognosis
	Lower Grade Glioma	In-Peak Gene Amplification	1.31E-08	3.48E-08	14	206	7p11.2	EGFR (1956), SEC61G (23480)	EGFR	poor prognosis
	Lower Grade Glioma	In-Peak Gene Amplification	1.48E-10	1.18E-09	5	206	12q14.1	CDK4 (1019), CYP27B1 (1594), MARCH9 (92979), TSPAN31 (6302), AGAP2 (116986), AVIL (10677), CTDSP2 (10106), FAM119B (25895), METTL1 4234, OS9 (10956), TSEF (10102)	CDK4	Poor prognosis
	Lower Grade Glioma	Gain of Function Mutation	1.09E-10	6.56E-10	6	166	7p12	EGFR (1956)	EGFR	Poor prognosis
	Lung Adenocarcinoma	In-Peak Gene Amplification	1.30E-12	4.66E-11	4	320	12p11	LOC100133893 (100133893), MRPS3 (604885), REPI5 (387849)	N	Poor prognosis

Table 15	Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
Lower Grade Glioma		In-Peak Gene Deletion	4.57E-12	6.85E-12	21	206	9p21	CDKN2A (1029), CDKN2B (1030), MTAP (4507)	CDKN2A (1029)	Poor prognosis
		In-Peak Gene Amplification	2.00E-15	1.01E-13	4	446	17q21	CCL3L3 (414062), CCL3L1 (6349)	No	Poor prognosis
Astrocytoma		Loss of Function Mutation		3.88E-03	34	59	17p13.1	TP53 (7157)	TP53	favorable outcome
		Loss of Function Mutation		8.15E-03	22	59	Xq21.1	ATRX (546)	no	favorable outcome
Breast Carcinoma		In-Peak Gene Deletion		8.14E-03	4	36	8p23.2	CSMD1 (64478)	no	poor outcome
		In-Peak Gene Amplification		5.71E-02	12	407	8q24.3	PARP10 (84875), MAPK15 (225689), PTK2 (5747), KHDRBS3 (10656)	PTK2	poor outcome
Colorectal Adenocarcinoma		In-Peak Gene Amplification		9.18E-02	17	407	13q34	FAM70B (348013)	no	poor outcome
		Gain of Function Mutation		8.10E-03	8	32	3q26.3	PIK3CA (5290)	PIK3CA	poor outcome
Cutaneous Melanoma		In-Peak Gene Amplification		2.60E-06	6	231	8q22.3	ODF1 (4956)	no	poor outcome

Table 15	Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
	Cutaneous Melanoma	In-Peak Gene Amplification		1.54E-04	7	231	8q24.3	PARP10 (84875), MAPK15 (225689), PTK2 (5747), KHDRBS3 (10656)	PTK2	poor outcome
	Cutaneous Melanoma	In-Peak Gene Amplification		7.21E-03	8	231	8q21	HEY1 (23462)	no	poor outcome
	Cutaneous Melanoma	In-Peak Gene Amplification		7.59E-03	6	231	11q13.3	FADD (8772), CCND1 (595), ORAOV1 (220064), FGF19 (9965)	CCND1	poor outcome
	Cutaneous Melanoma	In-Peak Gene Amplification		1.82E-02	4	231	1q44	OR2T27 (403239)	no	poor outcome
	Cutaneous Melanoma	In-Peak Gene Amplification		9.36E-02	6	231	1q21.3	LCE1E (353135)	no	poor outcome
	Ductal Breast Carcinoma	In-Peak Gene Amplification		2.77E-03	4	665	3q29	OSTalpha (200931)	no	poor outcome
	Ductal Breast Carcinoma	In-Peak Gene Amplification		2.28E-02	7	665	6q23.3	AH11 (54806)	no	poor outcome
	Ductal Breast Carcinoma	In-Peak Gene Amplification		2.64E-02	8	665	3q26.3	PIK3CA (5290), SOX2 (6657), ATP11B (23200)	PIK3CA	poor outcome
	Ductal Breast Carcinoma:ER Positive and HER2 Negative	In-Peak Gene Amplification		7.92E-06	6	263	1q21.3	ADAMTSL4 (54507), MCL1 (4170), ENSA (2029)	MCL1	poor outcome

Table 15									
Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
Ductal Breast Carcinoma:ER Positive and HER2 Negative	In-Peak Gene Amplification		4.02E-02	7	263	1q32	MDM4 (4194)	MDM4 (pre-clinical)	poor outcome
Ductal Breast Carcinoma:ER Positive and HER2 Negative	In-Peak Gene Deletion		4.35E-02	4	263	8p11.2	FKSG2 (59347)	no	poor outcome
Ductal Breast Carcinoma:ER Positive and HER2 Positive	In-Peak Gene Deletion		7.48E-02	4	84	9q22	FAM75C1 (441452)	no	poor outcome
Ductal Breast Carcinoma:HER2 Positive	In-Peak Gene Deletion		4.47E-02	4	116	15q13.1	CHRFAM7A (89832)	no	poor outcome
Ductal Breast Carcinoma:HER2 Positive	In-Peak Gene Deletion		5.17E-02	4	116	9p21	CDKN2B (1030)	CDKN2B (pre-clinical)	poor outcome
Ductal Breast Carcinoma:Triple Negative	In-Peak Gene Amplification		2.58E-02	5	75	1q23.3	APOA2 (336), SDHC (6391), FCGR2B (2213)	no	poor outcome
Ductal Breast Carcinoma:Triple Negative	In-Peak Gene Amplification		7.21E-02	8	75	1q21	ACP6 (51205), ECM1 (1893), ADAMTSL4 (54507), MCL1 (4170), ENSA (2029)	MCL1	poor outcome
Endometrial Endometrioid Adenocarcinoma:Microsatellite Stable	Loss of Function Mutation		5.55E-02	19	113	5q13.1	PIK3R1 (5295)	no	poor outcome
Endometrial Serous Adenocarcinoma	In-Peak Gene Amplification		6.37E-04	4	52	19p13.2	DNMT1 (1786)	DNMT1	poor outcome

Table 15									
Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
Gastric Adenocarcinoma:Hyper-Mutator	In-Peak Gene Deletion		9.05E-02	8	106	9p21	CDKN2A (1029), CDKN2B (1030)	CDKN2A, CDKN2B (pre-clinical)	poor outcome
Glioblastoma	In-Peak Gene Deletion		2.58E-02	300	565	9p21	CDKN2A (1029), CDKN2B (1030)	CDKN2A, CDKN2B (pre-clinical)	poor outcome
Glioblastoma	In-Peak Gene Amplification		8.80E-02	189	565	7p11.2	SEC61G (23480)	no	poor outcome
Lung Adenocarcinoma	Fusion		5.79E-02	7	343	17q23.1	RPS6KB1 (6198), VMP1 (81671)	RPS6KB1	poor outcome
Lung Adenocarcinoma:Triple Negative	Loss of Function Mutation		1.31E-03	4	99	7q36.1	MLL3 (58508)	no	poor outcome
Oligoastrocytoma	Loss of Function Mutation		1.97E-02	38	53	17p13.1	TP53 (7157)	TP53	favorable outcome
Oligodendroglioma	Loss of Function Mutation		5.90E-02	6	89	9q34.3	NOTCH1 (4851)	NOTCH1	poor outcome
Oligodendroglioma	Loss of Function Mutation		6.62E-02	15	89	1p31.1	FUBP1 (8880)	no	poor outcome
Ovarian Serous Adenocarcinoma	In-Peak Gene Amplification		1.15E-02	17	562	19q13.1	FCGBP (8857), PAK4 (10298)	PAK4 (pre-clinical)	poor outcome

Table 15	Subset	event type	p-value	q-value	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggable genes	KM Evidence
	Ovarian Serous Adenocarcinoma	In-Peak Gene Amplification		6.59E-02	17	562	20q11.2-13.2	ZNF217 (7764), MYLK2 (85366), KIF3B (9371)	no	poor outcome
	Ovarian Serous Adenocarcinoma	In-Peak Gene Deletion		7.86E-02	7	562	17p13.1	ATP1B2 (482)	no	poor outcome
	Ovarian Serous Adenocarcinoma	In-Peak Gene Amplification		8.43E-02	53	562	19q12	CCNE1 (898)	CCNE1	poor outcome
	Squamous Cell Lung Carcinoma	In-Peak Gene Amplification		7.93E-02	63	320	3q26.2	MECOM (2122)	no	favorable outcome

Example 5: Additional Fusion Methods

[00450] Clinical Data Sources. All RNASeq data for gene fusion analysis was obtained from the Cancer Genomics Hub (CGHub), the current repository for TCGA genomic data - <https://cghub.ucsc.edu/>.

[00451] Cell Line Data Sources. All CCLE RNASeq data for gene fusion analysis was obtained from the Cancer Genomics Hub (CGHub), the current repository for CCLE NGS data - <https://cghub.ucsc.edu/>.

[00452] BAM to FASTQ conversion. The input to the fusion callers consists of RNASeq reads in FASTQ format, which required conversion of the BAM file provided by TCGA to one or two FASTQ files for single or paired end data (respectively).

[00453] BAM files varied in provenance and processing, and many required special handling. For example, older BAM files provided by UNC were aligned using BWA (Burrows-Wheeler Aligner), while newer BAMs contained reads aligned by MapSplice. TCGA recently updated the RNASeq pipeline to support alternative gene expression reporting. (The former pipeline relied on the RPKM measurements for gene expression, while the latter uses RSEM.) These different RNASeq analysis pipelines are referred to by UNC as V1 and V2 respectively (<https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2>). We used the following BAM prioritization pipeline to select a single "primary BAM" when both formats are available for the same TCGA sample: 1) V2 BAMs were chosen over V1 BAMs and 2) BAMs with newer upload dates were selected when multiple files for the same case were present.

[00454] The custom SamToFastq converter described above was used to generate FASTQ files from a TCGA BAM file.

[00455] There were 2 cancer types (COADREAD and UCEC) only available as single-end RNASeq data. For single-end BAM file conversion, the program BamTools (hypertext transfer protocol secure:<https://github.com/pezmaster31/bamtools>) was used to generate FASTQ files.

[00456] With the goal of supporting both single and paired-end data, we processed all single-end data using TopHat and all paired-end data using deFuse.

[00457] Broadly, our analysis pipeline consists of 5 main steps:

[00458] Pre-process the raw data to obtain FASTQ files

[00459] Run fusion callers

[00460] Filter breakpoints to gene regions of interest

[00461] Annotate the breakpoints with the Oncomine transcript set

[00462] Summarize and prioritize potentially interesting novel fusions

[00463] Steps 1 and 2 were executed in parallel for all samples on a high-performance cloud computing cluster. The filtering and annotation was conducted on the aggregated data as a post-processing step, to enable exploratory analyses of effects of various filters and annotation schemes. After finalizing filtering criteria to minimize false positive fusions (Step 5), the list of Oncomine Prioritized Fusions is validated with RNASeq Exon Expression data.

[00464] **TopHat.** TopHat-Fusion was obtained from the authors hypertext transfer protocol://tophat.cbcb.umd.edu. Software and reference data dependencies were configured as specified by the TopHat documentation:

[00465] **Software:**

[00466] TopHat: 2.0.4, includes TopHat-Fusion Post ([release 04/09/2012](#))

[00467] bowtie: 0.12.8 ([release 05/06/2012](#))

[00468] samtools: 0.1.18 ([release 09/02/2011](#))

[00469] blast (2.2.26) ([release 03/03/2012](#))

[00470] blast+ (2.2.26) ([release 10/21/11](#))

[00471] **Reference and Annotation:**

[00472] Reference Genome: UCSC [hg19](#) (downloaded May 2012)

[00473] Gene Models: [refGene](#), [ensGene](#) (downloaded May 2012)

[00474] BLAST DB: [nt](#), [human](#), [other](#) (downloaded May 2012)

[00475] **Parameters:**

[00476] We ran TopHat with largely default parameters on single and paired-end TCGA Illumina data as specified in the TopHat documentation. The following is a list of parameters used.

[00477] **Table 25**

TopHat Parameter	Value Used
------------------	------------

TopHat Parameter	Value Used
--fusion-search	<i>Flag</i>
--keep-fasta-order	<i>Flag</i>
--no-coverage-search	<i>Flag</i>
--mate-inner-dist	0
--mate-std-dev	80
--min-anchor-length	8
--splice-mismatches	0
--min-intron-length	70
--max-intron-length	500,000
--max-insertion-length	3
--max-deletion-length	3
--num-threads	4
--max-multihits	20
--transcriptome-mismatches	2
--genome-read-mismatches	2
--read-mismatches	2
--segment-mismatches	2
--segment-length	25
--fusion-min-dist	100,000
--fusion-anchor-length	13
--fusion-read-mismatches	2
--fusion-multireads	2
--fusion-multipairs	2

TopHat Parameter	Value Used
--fusion-ignore-chromosomes	chrM

[00478] The --mate-inner-dist and --mate-std-dev parameters have no default values. The first parameter specifies an expected insert size for the RNASeq paired-end reads, while the second parameters specifies the expected standard deviation of that value. The values of 0 and 80 are recommended by TopHat authors for most data sets.

[00479] Table 26

TopHat-Fusion Parameter	Post	Value Used	Explanation of Values
--num-fusion-reads		3	<i>Recommended value</i>
--num-fusion-pairs		0	<i>Set to 0 to not penalize low-evidence, but potentially important fusions</i>
--num-fusion-both		0	

[00480] TopHat-Fusion was executed on one sample at a time, immediately followed by TopHat-Fusion Post. We retained both, unfiltered TopHat-Fusion output and filtered TopHat-Fusion Post output, to enable deeper analyses.

[00481] deFuse. deFuse was obtained from the authors: [hypertext transfer protocol://defuse.sf.net](http://hypertext-transfer-protocol://defuse.sf.net). Software and reference data dependencies were configured as specified by the deFuse documentation:

[00482] Software:

[00483] deFuse: 0.5.0 ([released](#) 04/07/2012)

[00484] bowtie: 0.12.8 ([release](#) 05/06/2012)

[00485] R 2.15.0 ([release](#) 03/30/2012)

[00486] [blat](#), [faToTwoBit](#) (obtained on 05/01/2012)

[00487] Reference and Annotation:

[00488] Reference Genome: Ensembl [GRCh37.62 fa](#) (downloaded May 2012)

[00489] Gene Models: Ensembl [gtf](#) (downloaded May 2012)

[00490] Genomic Data:

[00491] UCSC EST fasta, EST alignments, and repeats (downloaded May 2012)

[00492] NCBI UniGene (downloaded May 2012)

[00493] **Parameters:**

[00494] We ran deFuse with default parameters, as specified in the deFuse program documentation.

[00495] **Table 27**

deFuse Parameter	Value Used
-bowtie_qual	phred33-quals
-max_insert_size	500
-discord_read_trim	50
-clustering_precision	0.95
-span_count_threshold	5
-split_count_threshold	3
-percent_identity_threshold	0.90
-max_dist_pos	600
-num_dist_genes	500
-split_min_anchor	4
-max_concordant_ratio	0.1
-splice_bias	10
-denovo_assembly	No
-probability_threshold	0.5
-covariance_sampling_density	0.01
-reads_per_job	1,000,000
-regions_per_job	20
-p	4

[00496] deFuse was executed on one sample at a time. We kept both the filtered and unfiltered results of deFuse output to enable deeper analysis.

[00497] Integration. We integrated the “Level I” data--the output from TopHat-Fusion Post’s *potential_fusion.txt* file and the output from deFuse’s *results.classify.tsv* file. deFuse reports many more potential calls at this level than TopHat, and thus may also report more false-positive predictions. The Level I data was chosen to strike a balance between utilizing the caller’s built-in filtering and allowing through enough results to identify potentially real fusions with somewhat weaker evidence.

[00498] As each caller provided a different level of annotation and supporting evidence for the fusion calls, the breakpoints of the predicted fusions from both callers were extracted and integrated into a common format for filtering and annotation. The integration steps consisted of converting the reported breakpoints to ones-based genomic coordinate system, and consolidation into a common file format.

[00499] Breakpoint Filtering. The predicted fusions from the “Level I” output of the callers were filtered to only retain those calls where each breakpoint was either in the 5’UTR or CDS region of a RefSeq transcript (refGene circa July18, 2012, obtained from UCSC). This was done to enrich the predicted fusions for those containing functional gene regions, filtering out, for example, fusions calls where the 3’UTR of one gene is predicted to be fused to a 3’UTR of another gene. Although at the genomic DNA level breakpoints may occur in introns, in RNASeq data such breakpoints would be observed at the nearest exon-intron boundary. Therefore, breakpoints predicted to occur in intronic sequences were also excluded.

[00500] Breakpoint Annotation. After excluding fusions outside of the 5’UTR or CDS region of a RefSeq transcript, the annotation from the RefSeq transcripts was transferred to the remaining breakpoints with some predictions annotated against multiple Entrez IDs.

[00501] For each pair of breakpoints, only one transcript per Entrez ID was retained. In case of multiple transcripts, the transcript with the shortest transcript accession was chosen; further ties were broken by sorting the accessions alphanumerically and retaining the first accession. This scheme ensured consistency in annotating breakpoints at the same location. However, predicted breakpoints at different locations for the same gene partners may still result in multiple transcripts representing a pair of genes – possible evidence of alternative transcripts.

[00502] Basic annotation coming from the callers themselves was discarded, as it was based on the default annotation source of each respective caller. However, certain output fields from both TopHat and deFuse were retained to help prioritize the predicted fusions. Additionally, certain annotation properties that weren't explicitly reported by the callers were inferred from other caller properties.

[00503] Inferred Properties. Supporting and Spanning read counts were obtained from each caller and summarized in two columns – **Reads Span** and **Reads Span Support**. The latter column is a sum of reads spanning the fusion and those supporting the fusion (not to be confused with TopHat's count of "spanning mate pairs where one end spans a fusion," which is sometimes referred to as 'spanning and supporting reads').

[00504] The breakpoint sequence reported by the callers was trimmed to include 50 bases on each side of the fusion and consolidated into one column – **Breakpoint Sequence**. The fusion breakpoint is delineated by a "|". Note that this is the breakpoint sequence as inferred by the caller, and is not simply obtained from the reference genome. Because the inferred sequence may reflect actual sequence observed by the spanning reads, this sequence may represent the complement of the reference genome sequence.

[00505] Since neither of the callers provides a definitive '5-prime' or '3-prime' flag, we infer the relative **5'-3' orientation** of the fusion partners by combining a caller parameter with the gene strand annotation. For deFuse, the orientation was inferred for each partner based on the following combination of the gene strand and the deFuse output property '*genomic_strand*.'

[00506] Table 28

Gene Strand	<i>deFuse_genomic_strand</i>	
	+	-
+	5'	3'
-	3'	5'

[00507] TopHat reports a different metric – the relative orientation of reads mapped to the gene partners, so a different rule set is required for inferring 5'-3' order for a pair of genes:

[00508] Table 29

Gene Strand	A/B	<i>tophat_orientation</i>			
		ff	fr	rr	Rf
+/+		5'-3'		3'-5'	
+/-			5'-3'		3'-5'
-/-		3'-5'		5'-3'	
-/+			3'-5'		5'-3'

[00509] A **Valid Orientation** field was labeled as "Y" if there was an inferred 5' and 3' partner for a given gene fusion call.

[00510] **RepeatMasker Annotation.** Each predicted breakpoint location was also annotated with RepeatMasker features in the neighborhood of the breakpoint. This was done to identify breakpoints in highly repetitive genomic regions, where alignment errors were likely to affect the prediction of the chimeric transcript.

[00511] Specifically, a 25 bp sequence upstream or downstream of the 5' and the 3' partner breakpoint respectively was selected as a 'breakpoint flank'. These flanks were intersected against the RepeatMasker elements set (<http://www.repeatmasker.org/>) downloaded from UCSC Table Browser on August 24, 2012. We reported the element name, element length, and amount of overlap with the 26 base breakpoint flank region for each breakpoint. Currently, the RepeatMasker elements are not filtered for specific element types (LINES, SINES, simple repeats, etc.).

[00512] For each fusion prediction, we set a RepeatMasker Overlap field to equal the number of bases the breakpoint flank sequences overlaps with a RepeatMasker element, and considered overlaps of 12 or more bases to be significant. The frequency of significantly overlapping fusion calls is used in the OncoPrint Prioritization described

below such that gene fusions with a lower frequency of overlap are considered higher quality.

[00513] Fusion Exon Expression Imbalance. Fusions were visualized using RNASeq exon expression data to provide secondary evidence of true positive fusion events by searching for exon expression imbalance before and after the breakpoint call. Specifically, if the 3' partner's expression is impacted by the 5' partner's promoter region, then exon expression should increase post-predicted breakpoint. This effect is especially visible when viewing fused versus non-fused patient samples.

[00514] TCGA Exon Expression Data. TCGA exon expression data was downloaded from the Broad's GDAC Firehose site. The RPKM RNASeq values are listed for each patient as Gene Annotation Format (GAF) features corresponding to a composite of UCSC exons from several different gene definitions including RefSeq. After downloading data for 21 diseases, we found that 4 different sets of GAF features were used to annotate RPKM expression. Finally, availability of patient expression data varied per disease in V1 and V2 RNASeq analysis pipelines described above.

[00515] To address these challenges we first mapped UCSC RefSeq exons to available GAF features and calculated the percentage overlap between each RefSeq exon and GAF feature. This step is critical since all CBI processed fusion breakpoints are mapped to UCSC Refgene definitions downloaded on 7/18/2012 and these breakpoints must in turn be mapped to GAF features. 80.8% of the 396,298 RefSeq exons map perfectly to GAF features in the plot shown below. We selected and reported on the RefSeq exon and GAF feature pair that resulted in the largest overlap.

[00516] A value called *rg_pct* provides a metric of the mapping quality of a given RefSeq exon with a GAF feature based on the following formula:

$$\text{[00517]} \quad rg_pct = \text{overlap} / \text{length}_{refseq} * \text{overlap} / \text{length}_{GAF \text{ feature}}$$

[00518] Mappings with an *rg_pct* value of 1 overlap perfectly, while values less than 1 indicate the RefSeq exon or GAF feature did not map to the exact same genomic regions and the RPKM value may be suspect.

[00519] We selected RNASeq V2 data for all diseases except STAD due to non-availability of V2 data.

[00520] Cell Line Exon Expression Data. Exon expression data for cell line samples was generated from the CCLE BAM files obtained from CGHub. The method

employed was similar to Step 18 as described in the "TCGA mRNA-seq Pipeline for UNC data" method available here: [hypertext transfer protocol secure://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/UNC_mRNAseq_summary.pdf](https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/UNC_mRNAseq_summary.pdf)

[00521] A difference between the UNC method and our method is the use of RefSeq Exons BED in our method instead of a composite exons BED used by the TCGA.

[00522] Exon Expression Imbalance Calculation. Each sample was systematically analyzed for evidence of potential 5' promoter-induced imbalance in 3' partner expression. Expression levels for each gene were first converted to a log scale, and then z-score normalized across each disease's sample cohort. This normalization was performed at the exon level to account for population-wide trends such as 3' bias or poor RefSeq exon / GAF feature match (see below).

[00523] Raw RPKM expression values (top) vs. z-score normalized values for PLXNB21 and COL7A1 in Ovarian Serous Carcinoma patients (See Figure 8 A-D). The population-wide dips in PLXNB1 expression at exons 12, 17 and 23 are smoothed out in the normalized data. A sample predicted to harbor a fusion between these genes is highlighted in red; wild-type patients are shown in blue. The red diamond indicates the caller-predicted breakpoint exon.

[00524] Prior to normalization, samples that were considered wild-type for the fusion under consideration but that were predicted to harbor other fusions involving one of the gene partners were removed from the wild-type population, so as not to contaminate z-score calculations.

[00525] After normalization, each sample was assigned a p-value calculated via one-sided Student's t-test on the hypothesis that the sample's post-breakpoint normalized expression values (Population A) have a higher mean than the pre-breakpoint values ($H_0: \mu_A \leq \mu_B$). The caller-predicted breakpoint was used to separate the expression populations for samples identified by either fusion caller.

[00526] P-values were also calculated for each wild-type sample to facilitate analysis of p-values for fusion-positive samples in the context of the overall population. This allows us to discard fusions involving genes that exhibit population-wide exon imbalance trends that are not fusion-induced. Any sample whose p-value did not rank within the top fraction of wild-type sample p-values was discarded. The breakpoint that

maximized the difference between pre- and post-breakpoint expression levels was used for wild-type sample p-value calculation.

[00527] Fusion Summarization. Fusions were summarized within a disease based on the occurrence of unique gene pairs, and based on the occurrence of individual genes, possibly with multiple partners.

[00528] For a unique fusion pair (unique by Entrez ID pair), the number of samples within a disease with at least one prediction of that fusion by either caller is the **Fused Sample Count**. Since multiple breakpoints for the same pair of genes may be reported in one sample and across the samples, the number of unique fusion pairs within each disease is much less than the total number of fusion calls. In order to filter and prioritize fusions at the gene pair level rather than the fusion call level, several of the fusion caller properties were summarized. The following table shows the properties that were summarized for a given fusion partner pair across the individual predictions:

[00529] Table 30

Property	Summary Method
DEFUSE_EVERSION	% of total fusion calls = 'Y'
DEFUSE_VALID_ORIENTATION	% of total fusion calls = 'Y'
DEFUSE_NUM_MULTI_MAP	% of total fusion calls > 0
TOPHAT_VALID_ORIENTATION	% of total fusion calls = 'Y'
3P/5P_REPEATMASKER_OVERLAP	% of total fusion calls ≥ 12

[00530] The **Adjacent** flag is set for a fusion if the genes are <1Mb apart on the genome and the *defuse_eversion* flag is set in ≤ 75% of the individual fusion prediction for these fusion partners.

[00531] Gene-Level Summary. Fused sample counts were also summarized at the gene level (unique by Entrez gene ID) within each disease type and across diseases (pan-cancer). This summarization approach was irrespective of inferred orientation within the fusion. In addition, fused sample counts were tallied for only the Oncomine Priority fusions (described below).

[00532] Individual unique fusion pairs were cross-referenced to the **Mitelman** database of genomic aberrations (hypertext transfer

protocol://cgap.nci.nih.gov/Chromosomes/Mitelman). The match was done based on gene names and not disease type. Therefore, gene fusions reported in Mitelman in a certain disease may have occurred in a different disease type in the TCGA datasets.

[00533] Gene fusions summarized at the gene level were cross-referenced to the Mitelman database based on gene name. Thus, there is more potential for the gene as reported in Mitelman to be of different histology or altogether different aberration type (for example a large chromosome-level deletion instead of a fusion) than the predicted unique fusion pairs.

[00534] Normal Sample Fusion Blacklist. With the assumption that all fusions called in TCGA normal samples are false positives, we asked the following questions: 1) Are fusion calls in tumor samples identified in normal samples? 2) Are Oncomine Prioritized Fusions identified in tumor samples also identified in normal samples? Answering the first question provides a baseline sense of the technical false positive rate in tumor gene fusion calls. The second question is a sanity check on how well the Oncomine Priority Fusion filter is overcoming this problem. 344 paired-end normal samples across 10 diseases were downloaded and processed using the same deFuse pipeline described above. A total of 56,579 total fusion calls consisting of 6,024 unique fusions were observed. These normal sample fusion calls were used to generate a blacklist and remove these false positives from Oncomine Priority gene fusions.

[00535] Paralogous Fusion Partner Blacklist. A blacklist of fusions between paralogous gene family members was assembled using two strategies: 1) manually inspecting high frequency fusion partner gene names and 2) comparing the first 3 characters of all Priority Fusion partner gene names. In the latter strategy, fusion partners were verified to be “paralogous” using HomoloGene, Ensembl, SIMAP, and GeneDecks V3 before inclusion in the final blacklist. The table below shows the top 10 most commonly observed gene fusion calls between paralogous fusion partners. The entire table consists of more than 400 unique paralogous gene fusions and is used to remove these false positives from our Oncomine Priority gene fusions.

[00536] Table 31

GeneA Symbol	GeneB Symbol	Observed in Normal	TCGA Cancer Types
HLA-B	HLA-C	YES	BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LAM L,LGG,LIHC, LUAD,LUSC,OV, PAAD,PRAD,SKCM,STAD,THCA,UCEC

HLA-A	HLA-B	YES	BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,LAML,LGG,LIHC,LUAD, LUSC,OV,PAAD,PRAD,SKCM,STAD,THCA
HLA-A	HLA-C	YES	BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,LAML,LGG,LIHC,LUAD, LUSC,OV,PAAD, PRAD,SKCM,STAD,THCA
TTLL12	TTLL12	YES	BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,LAML,LGG,LIHC,LUAD, LUSC,OV,PAAD,PRAD,SKCM,STAD,THCA
TRPV1	TRPV1	YES	BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,LAML,LGG,LIHC,LUAD, LUSC,OV,PAAD,PRAD,SKCM,STAD,THCA
B9D1	B9D1	YES	BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,LGG,LIHC,LUAD,LUSC, OV,PAAD,PRAD,SKCM,THCA
TGIF2-C20ORF24	TGIF2-C20ORF24	YES	BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,LAML,LGG,LIHC,LUAD, LUSC,OV,PRAD,SKCM,STAD,THCA
HLA-B	HLA-E	YES	BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LAML,LGG,LIHC,LUAD,LUSC,OV,PAAD,PRAD,READ,SKCM,STAD,THCA,UCEC
SEC16A	SEC16A	YES	BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,LAML,LGG,LUAD,LUSC,OV, PRAD,SKCM,THCA
LOC390940	LOC390940	YES	BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,LGG,LUAD,LUSC,OV, SKCM,STAD,THCA

[00537] Fusion Prioritization - Oncomine Priority scheme. The Oncomine Priority scheme outlined below was designed by iterative exploration of the top results in the Level I fusion predictions and systematic elimination of suspect false-positive fusions, while retaining previously discovered 'true-positive' (Mitelman) fusions. This scheme was meant to highlight fusions that conformed to certain features expected of a 'true-positive' fusion, and conversely, lack features observed in many 'false-positive' fusions.

[00538] A fusion is an **Oncomine Priority** fusion if:

[00539] Table 32

Fusion Summary Property	Value	Explanation
DEFUSE_VALID_ORIENTATIONTOPHAT_VALID_ORIENTATION	> 0.75	Most predictions in correct orientation
ADJACENT	'N'	
REPEATMASKER_FREQUENCY	< 0.25	Minority or none of predicted breakpoints are in repetitive regions
DEFUSE_NUM_MULTIMAP	> 0	Most spanning reads map uniquely to fusion breakpoint
PARALOGOUS_PARTNERS	Not on Paralogous Blacklist	Manually curated blacklist of predicted fusions between paralogous genes
OBSERVED_IN_NORMAL	Not on Normal Blacklist	List derived from processing 344 Normal samples using deFuse.

Example 6: Oncomine NGS Mutation Methods

[00540] Mutation Integration. The goal of the data integration was to create the most complete set of NGS mutation data currently available. We considered the following sources:

[00541] Primary Data Sources

[00542] COSMIC Cell Lines Project

[00543] TCGA Data from Broad GDAC Mutation_Packager (stddata build)

[00544] TCGA Data from DCC level 2

[00545] Compendia mutation calls based on TCGA Data

[00546] Publications containing NGS mutation data

[00547] COSMIC Cell Lines Project

[00548] The Cancer Genome Project has characterized the exomes of over 1000 cancer cell lines for mutations. The database provides the mutation data, filtered for quality, in a flat-file format. The cell line data was subjected to the same OncoPrint curation and annotation processes used for clinical mutation data. Cell line names were vetted against the OncoPrint ontology, and cancer types were standardized to be comparable with clinical mutation data.

[00549] The dataset was obtained from the Wellcome Trust Sanger Institute Cell Lines Project website: http://cancer.sanger.ac.uk/cancergenome/projects/cell_lines/ as it appeared in November 2013.

[00550] Broad GDAC Mutation_Packager. Broad has been working since Q3 2011 on gathering and integrating mutation data from multiple sources.

[00551] <https://docs.google.com/document/d/18XIVWv-a9xLBOflNikOa9rCXOyiravMM8-PVJxAQPPo/edit>

[00552] The above document details the provenance of the MAF files the Broad integrates into Mutation_Packager standard data runs. The Broad has integrated many MAF files that are maintained outside of the central TCGA DCC system, often by members of the Analysis Working Groups themselves. We have performed extensive comparisons between all MAF files available to us. It is our belief that the Broad has the most complete mutation data available.

[00553] For this release, we integrated data from the 2013_08_09 stddata build.

[00554] TCGA DCC Level 2. This is the controlled access mutation data available from the DCC. TCGA has a page on their wiki that provides additional details about the MAF files available:

[00555] <https://wiki.nci.nih.gov/display/TCGA/TCGA+MAF+Files>

[00556] For this release, we considered all MAF files available as of September 15, 2013.

[00557] NGS DNaseq Mutation Calls. PRAD mutation calls available from TCGA were of low quality and resulted in false-positive 'Gain of Function' predictions. Therefore, all calls for this disease were sourced from Compendia's own mutation calling pipeline. The Compendia mutation calls were made to conform to the MAF file format for integration. Please see the Appendix: Compendia NGS DNaseq Mutation Calling for more details. Included in this release are 170 Prostate Adenocarcinoma patients.

[00558] Hand-Curation of All NGS Data. TCGA and Non-TCGA NGS datasets were sourced by the Oncomine curation team directly from their primary sources- mainly peer-reviewed cancer publications and the above publically accessible databases. Mutation data, usually available in the Supplementary Materials, was brought to the standard required for mutation re-annotation and classification as part of the overall NGS Mutation processing pipeline. Cancer types were curated using the Oncomine cancer type ontology, assigning the appropriate Oncomine Cancer Type based on the best-available clinical metadata present in the publication. Since all the published experiments claimed whole-genome ('NGS') coverage, the null gene set for each dataset was assumed to be inclusive of all human RefSeq genes. The non-TCGA data was processed in the same exact way as the TCGA MAF-file data for the rest of the mutation analysis pipeline.

[00559] Remove Duplicate Mutations. We performed some simple clean-up operations to remove duplicate mutation records present in the source data. We also performed several file-column name re-mappings, as many of the sources do not adhere to the MAF file standard. Duplicate mutations from various tumor/normal aliquot pairs of the same patient sample were removed.

[00560] Mutation Annotation. Data obtained from the TCGA and non-TCGA sources contains mutation results from datasets processed and annotated by different genome sequencing centers or authors over the course of several years. This leads to the

mutation calls annotated using different gene models and using different conventions for variant classification. Since Compendia's approach to defining mutations relies on accurate variant annotation, we re-annotated the mutations against a single set of transcripts and consistent variant classification rules. A standard annotation pipeline ensured that mutations across disease types are evaluated consistently and are subject to common interpretation during the nomination of potential oncogenes or tumor suppressor genes. It also provided important annotation not consistently available from the primary sources, such as the HGVS-style mutation nomenclature (e.g., V600E).

[00561] Mutations obtained from primary sources are processed by Compendia according to the following general steps (details provided below).

[00562] We first re-annotated each mutation using Compendia's OncoPrint transcript set. Successfully annotated mutations received Compendia-derived annotation, while the rest retain annotation obtained from the primary source. Annotation includes:

[00563] Variant classification

[00564] Variant position

[00565] Variant change

[00566] Several filtering steps are implemented to remove redundant annotation in multiple transcripts, and mutations located outside of gene regions of interest.

[00567] Excluding "Ultra-mutator" Samples. In certain diseases, such as Endometrial Carcinoma, several highly-mutated samples may dominate the overall mutation counts. We also observed such "ultra-mutator" samples in Lung Adenocarcinoma, Gastric cancer, Melanoma, and Colorectal cancer. Based on a cut-off determined by analyzing ultra-mutator outliers in several cancer types, we decided on <5,000 non-silent exon mutations as the threshold for inclusion of a sample in our recurrence analysis. We therefore excluded a number of ultra-mutator samples in this dataset from our downstream analysis pipelines.

[00568] In the Mutation Annotation step, we attempted to re-annotate the mutations obtained from the primary sources against a standard transcript set compiled by Compendia. This transcript set included RefGene transcripts from hg18 and hg19 genome builds, obtained from UCSC on February 19, 2012.

[00569] Each mutation is individually mapped against a contig in the OncoPrint Transcript Set within the specified genome build. SNP mutations were mapped directly

to their start location, while for small insertion (INS) and deletion (DEL) mutations a position of interest is selected for mapping. For insertions, the position of interest is the base at which the insertion occurred. Depending on the direction of the transcript, this can either be the start or the end coordinate of the mutation, depending on whether the gene is on the positive or negative strand respectively. For deletions, the position of interest is the deleted base if the transcript is on the positive strand or the last base deleted if the transcript is on the negative strand. This adjustment ensures that the mutation position is defined as the first base affected by the insertion/deletion with respect to the direction of the transcript translation, i.e. 5' → 3'.

[00570] For a mutation successfully mapped to a transcript, the Compendia mutation annotation was inferred with respect to that transcript. For mutations that failed to map, the annotation from the primary data source was retained, and a variant position for Hotspot calculations was constructed based on the genomic coordinate (more details below). Since only the standard set of 23 chromosomes was included in our transcript set, mutations located on mitochondrial or other non-standard contigs were not mapped.

[00571] Below is a description of the criteria used in annotating the mutations that map to the Oncomine Transcript Set.

[00572] Variant Classification. For each mutation successfully mapped to a transcript, the variant classification was inferred using a combination of mutation and annotation properties. Our approach identified six main mutation variant classifications, all located within transcript. Variant classifications for mutations outside a gene region (e.g. intergenic) are currently not considered (see filtering section below). The following are the criteria used for inferring the variant classification:

[00573] Table 33

Variant Classification	Criteria	Transcript Region
Splice_Site	Mutation is within 2 bp of a splice site	exon or intron
3'UTR, 5'UTR	Mutation is in UTR region and not within 2 bp of splice site	UTR exon
Intron	Mutation is in an intron and is	intron

Variant Classification	Criteria	Transcript Region
	between 3 to 10 bp from a splice site	
Missense, Nonsense, Nonstop, Silent	Mutation is a SNP	coding exon
Frame_Shift_Ins/Del	Mutation is an INS/DEL not divisible by 3	coding exon
In_Frame_Ins/Del	Mutation is an INS/DEL divisible by 3	coding exon
Non_Coding_Exon	Mutation is in a non-coding transcript	non-coding exon

[00574] This list of variant classifications is a subset of the allowed variant classification specified by the TCGA for the MAF file format.

[00575] <https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+%28MAF%29+Specification>

[00576] This subset covers the mutation classes of interest for recurrence analysis and identification of potential Gain or Loss of Function genes, and is thus sufficient for the vast majority of the mutations that are mapped to the Oncomine Transcript Set. The following table describes the likely variant classification that would be assigned versus an original author classification (assuming mutation maps to the same transcript as that used in defining classification), and the relative abundance of that type of mutation in the source dataset:

[00577] Table 34

Example TCGA Variant Classification	Equivalent Compendia Variant Classification	Potential Mutation Classification (H)otspot, (D)eleterious or (O)ther

Example TCGA Variant Classification	Equivalent Compendia Variant Classification	Potential Mutation Classification (H)otspot, (D)eleterious or (O)ther
Missense_Mutation	Missense_Mutation	H,O
Nonsense_Mutation	Nonsense_Mutation	D
Nonstop_Mutation	Nonstop_Mutation	H,O
Silent	Silent	O
Frame_Shift_Del	Frame_Shift_Del	D
Frame_Shift_Ins	Frame_Shift_Ins	D
Translation_Start_Site	Missense_Mutation	O
In_Frame_Del	In_Frame_Del	H,O
In_Frame_Ins	In_Frame_Ins	H,O
3'UTR	3'UTR	O
5'UTR	5'UTR	O
Non_coding_exon (or "RNA")	Non_coding_exon	H,O
Splice_Site	Splice_Site	O
Intron	Intron	--
5'Flank	--not supported by Oncomine transcript set--	--
IGR	--not supported by Oncomine transcript set--	--
Other (classification present in mutation list but not supported by TCGA)		--

[00578] Variant Position. One of the primary goals of the current analysis is to identify genes with Hotspot mutations, which are mutations of a certain classification that are observed at the same location in multiple tumor samples. To effectively identify recurrence and define a hotspot for each mutation, we must construct a mutation spot identifier that encompasses the mutation position, the identity of the amino acid or base affected, and the variant classification. We aggregated mutations that occur at the same location irrespective of the specific base change they generate. Therefore, we only used the reference base or amino acid to define the variant position. This ensures that mutations affecting the same codon or genomic position will be counted towards a possible hotspot, even if the alternate alleles they generate are different. For example, for a given gene, missense mutations V600E, V600F and V600G would all have a variant position of V600 and would thus be aggregated together when identifying hotspot mutations. Our variant position is thus defined as follows:

[00579] Variant Position = mutation spot {base|codon} + reference {base|AA} + [variant classification]

[00580] If the mutation is in a coding region, then the codon number and the respective amino acid at the base of interest is used to identify the mutation spot—p.L116_in_frame_del—for example. If the mutation is in a non-coding region, such as the UTR, then the position and identity of the reference nucleotide at the base of interest is used to identify the mutation spot—c.*110C—for example.

[00581] For Splice_Site mutations outside of the coding region, the variant position is specified relative to the splice boundary. The relative position is identified using a +{1|2} or a −{1|2} (splice site mutations are those within 2 bases of a splice junction). As with insertions and deletions, a suffix of “_Splice_Site” is added for a Splice_Site mutation. For INS and DEL mutations, a suffix indicating an in frame (“_in_frame_ins” or “_in_frame_del”) or frame shift (“_frame_shift_ins” or “_frame_shift_del”) is added to the variant position.

[00582] In summary, the following are examples of the different possible variant position formats:

[00583] Table 35

Variant Type	Near Splice Site?	In Coding Region?	Variant Position
SNP	YES	YES	p.A42_Splice_Site
		NO	c.42+1_Splice_Site
SNP	NO	YES	p.A42 (Missense, Nonstop, Silent) p.Stop42 (Nonsense)
		NO	c.*42T (3'UTR) c.-42C (5'UTR) c.42 (Non_coding_exon)
INS	YES	YES	p.A42_Splice_Site
		NO	c.42+1_Splice_Site
	NO	YES	p.A42_{in_frame_ins frame_shift_ins}
		NO	c.*42G_{in_frame_ins frame_shift_ins} (3'UTR) c.-42G_{in_frame_ins frame_shift_ins} (5'UTR) c.42 (Non_coding_exon)
DEL	YES	YES	p.A42_Splice_Site
		NO	c.42+1_Splice_Site
	NO	YES	p.A42_{in_frame_del frame_shift_del}
		NO	c.*42T_{in_frame_del frame_shift_del} (3'UTR) c.-42C_{in_frame_del frame_shift_del} (5'UTR) c.42 (Non_coding_exon)

[00584] For mutations that **do not** map to the OncoPrint Transcript Set, and hence do not have a transcript-based location, the genomic location (start position) and the

reference nucleotide (reference allele) is used as the variant position irrespective of the coding region or splice site proximity. The variant classification supplied by the primary data is then added as a suffix. For example, a SNP missense mutation would have a variant position such as "chr19_c.C22952756_Missense_Mutation", and a splice site SNP would have a variant position "chr1_c.A155025094_Splice_Site". The variant change (see below) for these mutations is not defined.

[00585] Although the suffix of the variant position often implicitly incorporated the variant classification, when calculating hotspots, both the variant position and the variant classification are explicitly used for aggregating mutations. Therefore, mutations that may produce identical variant positions but have different variant classifications (such as a missense and a nonsense SNP) were tallied separately.

[00586] Variant Change. The variant change provides HGVS-like information about the alternate allele change of the mutation. For SNP mutations in the coding region, the variant change is a full HGVS protein-level sequence variant description, indicating the alternate amino acid. For SNPs outside of the coding region, the alternate allele nucleotide base is provided.

[00587] For INS and DEL variant types, the variant position (see above) was used as the variant change. In these cases, the consequence of the change at the amino acid level is not inferred. As such, variant change for INS/DEL does not strictly follow HGVS specification.

[00588] The following are illustrative examples of variant changes for Compendia-derived mutation annotation:

[00589] Table 36

Mutation	Variant Classification	Variant Position	Variant Change
SNP in CDS, E>K, residue 137	Missense_Mutation	p.E137	p.E137K

Mutation	Variant Classification	Variant Position	Variant Change
SNP in Intron C>2, two bp from splice site	Splice_Site	c.4913-1_splice_site	c.4913-1
INS in CDS at residue Gly 264	Frame_Shift_ins	p.G264_frame_shift_ins	p.G264_frame_shift_ins
DEL of one base in a UTR	3'UTR	c.*1007A_frame_shift_del	c.*1007A_frame_shift_del

[00590] For mutations that **do not** map to the OncoPrint Transcript Set, the variant classification from the primary data source was retained.

[00591] **Transcript Filter.** To avoid retrieving multiple transcripts, and hence multiple annotations for a single mutation within a gene, we kept only one transcript per mutation per gene (unique Entrez ID). If a mutation mapped to several transcripts of a gene, only one was chosen. However, if a mutation mapped to several genes, then only one transcript per gene was selected. It is thus possible for a mutation to receive two different annotations, but only if they stemmed from transcripts with different Entrez IDs.

[00592] We chose the representative transcript for a mutation based on the following priority scheme:

[00593] Transcript with the most impactful variant classification:

[00594] High impact in coding: Missense, Nonsense, Nonstop, Frame-shift

[00595] Low impact in coding: In-frame, silent

[00596] Outside of coding region: Splice Site, 3' or 5' UTR, Non-coding exon

[00597] Outside of exon: Intron

[00598] If there is a tie based on priority, the transcript with the shortest (by length) RefSeq transcript accession is chosen, followed by the alphanumerically smallest

transcript accession in event of further ties. For example – of the transcripts NM_003319, NM_133378, and NM_00125685 for the TTN gene, we would choose NM_003319 as the representative transcript.

[00599] These steps allowed us to repeatedly choose a consistent transcript for the same type of mutation at one location. One consequence of choosing the most impactful transcript is that multiple transcripts may be utilized for mutations at multiple locations in a single gene. However, the benefit of this scheme is that any mutations of the same variant classification at the same location are always assigned to the same transcript, and hence will be in the same frame of reference when computing recurrence for hotspot identification.

[00600] Filter by Mutation Class and Type. All mutations were further filtered by variant type and class. To avoid including mutations of minor interest to gene function analysis, we filtered out mutations that were not resolved to a gene region, either because they fell significantly far outside of a transcript, or because they were in a location not associated with a RefSeq gene. These mutations were evident either by **their lack of gene identifier (Entrez ID = 0 or blank)**, or membership in the following variant classes: *Intron*, *5'Flank*, *IGR*, and *miRNA*.

[00601] We also filtered out mutations with variant type *DNP*, *TNP*, *ONP*, *Complex_substitution*, and *Indel*, as their annotation was not supported by our pipeline

[00602] Since certain data sources included extensive amounts of intronic and intergenic mutations, this filtering step significantly reduces the size of the dataset as many NGS datasets don't apply these filters pre-publication.

[00603] Classifying Mutations as Hotspot, Deleterious, or Other. The next step in our analysis pipeline identified recurring mutations in multiple samples based on their variant position, and categorized them into Hotspot, Deleterious or Other variant categories. For this step, and the subsequent frequency calculations, mutations for each disease type were processed independently. Only mutations of the same variant classification were tallied together, so, for example, a missense mutation and a silent mutation at the same position are counted separately.

[00604] To identify driver events, each mutation for a given Entrez Gene ID was categorized as "Deleterious" or "Hotspot" depending on the following criteria:

[00605] A mutation was deemed '**recurrent**' if it was observed in the same variant position in 3 or more tumor samples.

[00606] A mutation belongs to the "Hotspot" variant category if it is:

[00607] Recurrent **AND**

[00608] Annotated with one of the following variant classifications:

In-frame insertion/deletion

Nonstop

Missense

Non_Coding_Exon

[00609] A mutation belongs to the "Deleterious" category if it is:

[00610] Non-recurrent **AND**

[00611] Annotated with one of the following variant classifications:

Frame shift insertion/deletion

Nonsense

[00612] A mutation is considered in the "Other" variant category if it did not fit the above criteria.

[00613] The Oncomine Mutation Classification and the Variant Classification can be used to summarize the relative frequencies of various mutations at the gene level.

[00614] **Nominating "Gain of Function" and "Loss of Function" Genes.** Individual genes were classified into predicted functional classes, namely "Gain of Function", "Recurrent Other", and "Loss of Function", to reflect their relative enrichment in potential activating or deleterious mutations. Details of the scheme used to make the classification are provided below.

[00615] **Mutated Sample Frequency Calculation.** Mutation frequencies for each gene were calculated with respect to a given variant classification and variant category across all samples within a disease type. Overall mutation frequency for a gene within a disease was calculated by combining mutations of all variant classifications.

[00616] **Overall Mutation Frequency.** Overall mutation frequency for a gene was obtained by dividing the total number of samples with at least one mutation of any

variant classification in that gene (Mutated Sample Count) by the total number of samples in the given cancer type (Sample Count).

[00617] Hotspot Frequency. Hotspot frequency for a gene was obtained by dividing the total number of samples with at least one mutation belonging to the "Hotspot" Oncomine Mutation Classification by the Mutated Sample Count – the total number of samples with at least one mutation for the given gene. If a sample had both Hotspot Missense and a Hotspot In-Frame Deletion, for example, it would only be counted once.

[00618] Hotspot Missense Frequency. To obtain a Hotspot Missense Frequency for a gene, the number of samples containing at least one Missense mutation with an Oncomine Mutation Classification of "Hotspot" was divided by the Mutated Sample Count – the number samples with at least one mutation of any type in this gene. Samples with more than one mutation of such type were only counted once.

[00619] Deleterious Frequency. To obtain the Deleterious frequency for a gene, the number of samples containing at least one mutation with an Oncomine Mutation Classification of "Deleterious" was divided by the Mutated Sample Count – the number of samples with at least one mutation for the given gene. Samples with more than one mutation of that type were only counted once.

[00620] Other Frequency. To obtain the Other frequency for a gene, the total number of samples with at least one mutation with an Oncomine Mutation Classification "Other" was divided by the Mutated Sample Count – the total number of samples with at least one mutation for the given gene. If a sample contained both splice site and UTR mutations, for example, it would only be counted once.

[00621] Hotspot, Other, and Deleterious Frequency Consideration. Hotspot, Other, and Deleterious frequencies should not be expected to add up to 100%, since a sample may have been counted in more than one of these categories.

[00622] Assessing Significance of Hotspot and Deleterious Mutations. The Hotspot and Deleterious p-values for each gene within a disease are calculated by two independent methods.

[00623] Significance of Deleterious Mutation Enrichment. To assess whether a gene was significantly enriched for deleterious mutations compared with other genes, given the background mutation rate, we performed Fisher's exact test using the following contingency table:

[00624] Table 37

		Deleterious	Other
Gene of Interest		A	B
All Other Genes		C	D

[00625] where A, B, C, and D are counts of mutations across a disease. Nonsense mutations, frame shift insertions and frame shift deletions are classified as deleterious mutations, while mutations of any other type (UTR, silent, missense, etc., but non-intergenic) count as others.

[00626] Q-values are calculated within each disease, by counting the number of genes with deleterious mutations (N), and calculating the rank of each association. The q-value for a given p-value is then $Q=p*N/\text{rank}$.

[00627] Significance of Recurrent Hotspot Mutations. In order to calculate gene-specific p-values, the significance of the most recurrent hotspot on that gene is assessed. Given the assumption that each sequence position was equally likely to mutate, each gene can be tested whether the most recurrent is significantly greater than that expected using a multinomial test. This is an exact test of the sampling algorithm that has been implemented in previous versions. One of the advantages of this test is that the p-value precision is increased to $1E-16$, so no flooring occurs. To obtain hotspot mutations, we filtered the mutations to remove any that did not affect the coding sequence (i.e. by removing silent, UTR, stop codon, and splice site mutations), and then removed mutation data for genes that we could not annotate with RefSeq transcript identifier. We then counted the mutations observed for each transcript in each disease. We calculated the amino acid sequence length by dividing the CDS length by three and subtracting 1.

[00628] The exact calculation of the p-value is framed as the following. Given an amino acid sequence of length x , an observed number of hotspot mutations n , what is the likelihood of observing r or more mutations at the most recurrent spot by chance For each gene, the p-value is calculated by the following formula:

$$\begin{aligned}
p &= \Pr(y_{(x)} \geq r) \\
&= 1 - \Pr(y_1 < r, y_2 < r, \dots, y_x < r) \\
&= 1 - \sum_{i=0}^{r-1} \frac{n!}{y_1! y_2! \dots y_x!} (1/x)^n
\end{aligned}$$

[00629] where $y_{(x)}$ is the mutation count at the most recurrent hotspot, and $y_1, \dots, y_x$ stands for the mutation count at each spot 1, ..., x.

[00630] When n and x are large, the above formula can be very slow, an approximation with Bonferroni-Mallows (BM) bounds were used:

$$1 - \frac{n!}{n^n e^{-n}} \left\{ \prod_{i=1}^x P(y_i \leq r-1) \right\} P(W = n)$$

[00631] where y_i is a Poisson random variable with mean n/x , and $W = \sum_{i=1}^x Y_i$ where Y is a truncated Poisson. $P(W = n)$ is estimated by Edgeworth Expansion. The lower and upper Bonferroni-Mallows bounds for the p-value are:

$$\mathbf{[00632]} \quad 1 - \text{Binomial}\left(r-1, n, \frac{1}{x}\right)^x \leq p \leq x * (1 - \text{Binomial}\left(r-1, n, \frac{1}{x}\right)).$$

[00633] If the approximation falls outside of the BM bounds, either the lower bound or upper bound was used. It rarely occurred in our data, and it mostly occurred for small p-values ($p < 1e-16$) or large p-values ($p \sim 1$).

[00634] Q-values are calculated using the Benjamini-Hochberg method, which is $Q = p * N / \text{rank}$, where N is the number of transcripts and rank is the rank of each p-value.

[00635] Silent Hotspot Mutations. Recurrent silent mutations – silent hotspots – seem to be an indication of sequencing errors, occurring in regions of low sequence quality and serving as a ‘canary in the coal mine’ for false-positive missense mutation peaks in the neighboring nucleotides. Based on reviewing genes with silent hotspots, and the evaluation of neighboring silent peaks, we believe that these genes are subject to systematic sequencing errors, and hotspot mutations in these genes should not contribute to the gene classification.

[00636] Oncomine Gene Classification Rules. Once the mutations have been classified, individual genes are nominated to one of three classes – “Gain of Function,” “Loss of Function,” and “Recurrent Other.” The classification is based on the combination of relative frequencies and the significance of the mutations observed in the gene. The significance of the mutations per gene is assessed by a p-value.

[00637] Recurrent silent mutations. A “Gain of Function” gene will have a relatively high frequency of Hotspot Missense mutations and a low frequency of Deleterious mutations, while a “Loss of Function” gene contains a large fraction of Deleterious mutations. “Recurrent Other” genes tend to contain recurrent insertion/deletion mutations, some of which - for example recurrent frame shift indels of 1 base – exhibit signs of potential false-positive calls that may arise from local alignment errors. In general, we are more confident about the functional importance of genes classified as Gain/Loss of Function.

[00638] Pan-Cancer Analysis. To summarize mutations across diseases we performed identical calculations as we did for within-disease analyses, but without stratifying the mutation records by disease. All mutation records were aggregated, and frequencies, variant categories and gene classes were calculated in this pan-cancer context. For the pan-cancer summary, genes (unique by Entrez ID) are summarized across all diseases with one row per disease. However, a summary of the genes within disease is also provided, but in a pan-cancer context. This means, for example, that samples with Hotspot mutations are totaled within a disease, but only for the mutations considered Hotspots in a pan-cancer context. Cancer types with < 20 samples were included in Pan-Cancer analysis, even though they were not eligible for within-disease analysis due to low sample count.

[00639] Cell Line Annotations. Cell line mutation data was subjected to the same Oncomine curation and annotation processes described above *except* for mutation and gene classification. Instead, mutations from cell lines were annotated with Oncomine mutation classification and gene classifications whenever a mutation in a cell line was also observed in a clinical sample. This annotation was performed only for mutations having a Hotspot or Deleterious or Other Oncomine mutation classification. If a mutation was not observed in tumors, it would receive “Unobserved in Tumor” mutation classification.

[00640] Mutations from a cell line and a tumor sample are considered equivalent if they belong to the same gene, and have the same variant position and variant classification.

[00641] Cell lines names were vetted against internal Oncomine ontology, and cell line cancer types were standardized to be comparable with clinical mutation data. Several cell lines whose identity or cancer type could not be independently verified through databases or publications were removed from our analysis. The mutation annotation from clinical data was performed in a pan-cancer and within-disease contexts.

[00642] NGS DNaseq Mutation Calling

[00643] BAM File Selection. We queried TCGA's CGHub to identify patients having a single tumor-normal BAM pair. We did so to remove the possibility of mutation call differences due to different tumor-normal pairs.

[00644] Reference Genome Builds. We identified the reference genome builds used to align the reads in the BAM files by parsing the SAM headers. We located, downloaded, and indexed all the reference genome builds which are needed as inputs to the mutation caller packages.

[00645] Mutation Calling. We employed the following somatic mutation calling packages for this analysis:

[00646] MuTect (1.0.27783), Broad Institute, Cancer Genome Analysis Group (CGA) (Cibulskis, 2013)

[00647] SomaticIndelDetector (1.6-13-g91f02df), Broad Institute, Genome Analysis Toolkit (GATK)

[00648] MuTect. MuTect performs initial preprocessing to remove "reads with too many mismatches or very low quality scores" (MuTect documentation). Next, for a candidate mutation two log odds (LOD) scores are calculated that describe the likelihood of a mutation being present in the tumor sample (LOD_T) and not mutated in the normal sample (LOD_N):

$$\mathbf{[00649]} \quad LOD_T = \log_{10} \left(\frac{P(\text{observed data in tumor} | \text{site is mutated})}{P(\text{observed data in tumor} | \text{site is reference})} \right)$$

$$\mathbf{[00650]} \quad LOD_N = \log_{10} \left(\frac{P(\text{observed data in normal} | \text{site is reference})}{P(\text{observed data in normal} | \text{site is mutated})} \right)$$

[00651] MuTect expects somatic mutations to occur at a rate of ~1 in a Mb and requires $LOD_T \geq 6.3$. MuTect requires that a mutation not be in dbSNP and have a $LOD_N \geq 2.3$ since non-dbSNPs are expected to occur at a rate of 100 per Mb. Both cutoffs are chosen to guarantee a false positive rate less than half of the expected somatic mutation rate. Finally, additional post-processing steps are performed, including testing that the alternate allele is observed in both read directions. MuTect requires at least 14 tumor reads and 8 normal reads for a mutation to be considered.

[00652] SomaticIndelDetector (SID). For a given mutation site, SID considers candidate indels using counts-based thresholding and an indel consensus voting scheme. The indel with the largest number of supporting reads, or votes, is chosen as the putative indel call. This call is reported if there is:

[00653] Enough coverage (default: normal ≥ 4 reads, tumor ≥ 6 reads),

[00654] A large fraction of reads at that site support the putative call (default: $\geq 30\%$)

[00655] This fraction is sufficiently large compared to those supporting any site of the indel (default: $\geq 70\%$)

[00656] Indel calls in a tumor sample are annotated as "Germline" if there is even weak evidence for the same indel in the normal sample; otherwise, they are labeled "Somatic." Calls only observed in the normal samples are ignored. SID takes BAM files as input and outputs VCF and BED formatted putative calls.

[00657] Mutation Filtering. The callers output all candidate mutation calls, including germline mutations and other calls with low statistical confidence. We filtered the mutation caller output to only somatic mutations, mutations designated "KEEP" by MuTect and mutations occurring within the CDS of RefSeq Genes. The tables below detail the specific filters applied to MuTect and SomaticIndelDetector output:

[00658] Table 38

MuTect Filter	Description
tumor_f > 0.1	At least 10% of the tumor reads must be variant
t_alt_sum / t_alt_count > 28	Average quality of the variant base calls > 28
t_alt_count map_Q0_reads	– Conservatively require at least 3 reads where variant not – in Q0 or in improperly paired reads.

improper_pairs >= 3	
t_alt_count > 10 * n_alt_count	When MuTect allows one variant normal read, require at least 10 variant tumor reads.
dbsnp_site NE 'DBSNP'	Ignore variants present in dbSNP v132

SomaticIndelDetector Filter	Description
T_STRAND_COUNTS_C[12] / (..._C1 + ..._C2) > 0.1	At least 10% of the tumor variant reads must be on each strand
T_AV_MAPQ_C > 28	Average quality of the variant calls > 28

[00659] References

[00660] Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnology (2013).doi:10.1038/nbt.2514

[00661] MuTect: [hypertext](http://www.broadinstitute.org/cancer/cga/mutect) transfer
protocol://www.broadinstitute.org/cancer/cga/mutect

[00662] SID: [hypertext](http://gatkforums.broadinstitute.org/discussion/35/somatic-indel-detection) transfer
protocol://gatkforums.broadinstitute.org/discussion/35/somatic-indel-detection

Table 16: Druggability status for Table 2 genes/fusions

Table 16						
Gene	Approved	Pre-registrati on (pre-approval)	Phase III	Phase II	Phase I	preclinical
TOP1	belotecan hydrochloride; irinotecan hydrochloride; topotecan	N	cositecan; irinotecan, HyACT; irinotecan, PharmaEngi ne; etirinotecan pegol	gimatecan; camptotheci n, Calando; irinotecan HCl + floxuridine, Celator; firtecan pegol; TLC-388 hydrochlori de; hRS7-SN-38; irinotecan bead, Biocompati bles	irinotecan, liposomal, Yakult; HM-30181A; namitecan; camptothecin prodrug, Mersana; labetuzumab-SN-38; Genz-644282; simmitecan hydrochloride prodrug	camptothecin (Aphios); irinotecan (BioAlliance); cisplatin + irinotecan (Celator); APH-0804; irinotecan (Champions); SER-203; SN-38; topotecan + vincristine (LipoCure); topotecan (EnduRx Pharmaceutical s)
SRDSA 1	dutasteride	N	idronoxil	N	N	N
VIM	N	N	N	pritumumab	N	N

Table 16						
Gene	Approved	Pre-registrati on (pre-approval)	Phase III	Phase II	Phase I	preclinical
IGFBP 2	N	N	N	N	N	OGX-225
SPP1	N	N	N	N	N	N
MDK	N	N	N	N	N	CAMI-103; CMS-101
MUC1 6	N	N	oregovomab	N	DMUC-5754A	N
RET	sorafenib; vandetanib; sunitinib malate; cabozantinib; regorafenib	apatinib	motesanib diphosphate ; SAR-302503	N	JNJ-26483327	MG-516; NMS-173; RET kinase inhibitor (Bionomic)
MAP2 K2	trametinib	N	ARRY-438162	selumetinib; refametinib; pimasertib; WX-554	PD-0325901; ARRY-704; TAK-733; GDC-0623; BI-847325; AS-703988	N
MAPK I	N	N	N	N	N	AEZS-129; AEZS-136; AEZS-134; SCH-722984; SCH-772984
BRAF	pazopanib; vemurafenib; dabrafenib	N	N	RAF-265; XL-281; LGX-818	ARQ-761; ARQ-736	AB-024; b-raf inhibitors (Sareum); BRAF kinase inhibitor (Selexagen Therapeutics); BeiGene-283; DP-4978; TL-241
MUC1 6	N	N	oregovomab	N	DMUC-5754A	N
MET	cabozantinib; crizotinib	N	oregovomab; rilotumumab; onartuzumab;	MGCD-265; foretinib; ficlatuzumab; BMS-777607; golvatinib; INCB-028060; LY-2875358	AMG-208; TAS-115; volitinib; SAR-125844; S-49076	X-379; metatinib; PRS-110; ASP-08001; ARGX-111; DCC-2701; DCC-2721; MG-516; AL-2846; CG-206481; T-1840383; cMet-EGFR dual inhibitors (CrystalGenomics); bispecific antibodies (Hoffmann-La Roche)
PTK2	N	N	N	PF-04554878	GSK-2256098; BI-853520; VS-4718	CFAK-C4; FAK inhibitor (Verastem); CTX-0294945;

Table 16						
Gene	Approved	Pre-registrati on (pre-approval)	Phase III	Phase II	Phase I	preclinical
						CTx-0294886; FAK inhibitors (Takeda)
ACE*	alacepril; benazepril; delapril+manidipine (Chiesi); captopril; captopril + HCTZ; captopril slow release (Sankyo); cilazapril; delapril; delapril + indapamide (Chiesi); diltiazem, Alza; enalapril maleate; enalapril maleate + HCTZ; enalapril + nitrendipine; enalapril (KRKA); enalaprilat; felodipine + enalapril; fosinopril; imidapril; lisinopril; lisinopril + HCTZ; moexipril; perindopril; quinapril hydrochloride; quinaprilat; ramipril; felodipine + ramipril; perindopril + indapamide, Serv; saralasin acetate; spirapril; temocapril; trandolapril; zofenopril; trandolapril + verapamil, Aven; lercanidipine + enalapril (Recordati); zofenopril + HCTZ; pirtanide + ramipril; benazepril + HCTZ; amlodipine + benazepril; moexipril + HCTZ; amlodipine + perindopril, Servier; ASA + atorvastatin + ramipril + metoprolol ER (Zydus Cadila); ramipril + hydrochlorothiazide; (S)-amlodipine + ramipril (Emcure); quinapril/hydrochlorothiazide	N	perindopril + indapamide + amlodipine (Servier)	N	amlodipine + enalapril maleate (GlaxoSmithKline)	N
ADAM 9	N	N	N	N	N	N
CDK6	N	N	palbociclib	alvocidib; LY-2835219	LEE-011	N
IKBKB	N	N	N	N	N	EC-70124

Table 16						
Gene	Approved	Pre-registrati on (pre-approval)	Phase III	Phase II	Phase I	preclinical
RARA	tamibarotene	N	N	IRX-5183	N	N
LYN	dasatinib	N	nintedanib	bafetinib	JNJ-26483327	Bcr-Abl/Lyn inhibitor (AB Science)
NTRK3	N	N	N	N	N	PLX-7486
ERBB2	trastuzumab; trastuzumab emtansine; pertuzumab; lapatinib ditosylate; catumaxomab; afatinib	trastuzumab, Enhance	neratinib; XL-647; dacomitinib; nelipepimut-S; trastuzumab (Celltrion, Biocad, Biocon, Synthon, Harvest Moon, Aryogen)	lapuleucel-T; AVX-901; AE-37; BMS-690514; MVA-BN-HER2; varlitinib; MM-111; AC-480; ovarian cancer vaccine (Generex); margetuximab; poztotinib; PR-610	Her-VAXX; VM-206; ARRY-380; JNJ-26483327; S-222611; doxorubicin (Merrimack); cipatinib; TrasGEX; trastuzumab (Hanwha Chemical); trastuzumab (Pfizer); IDN-6439	Lovaxin B; TH-1 (Algeta); trastuzumab-antibody conjugates (Synthon); CUDC-101; Her-2/neu Stradobody (Gliknik); ARX-788; Etbx-021; SN-34003; IBI-302; NT-004; ICT-140; ONS-1050; Sym-013; anti-HER2 X anti-CD3 (Emergent Biosolutions); Z-650; breast cancer vaccine (Cel-Sci); JNJ-28871063; trastuzumab (PlantForm, BioXpress, biOasis Technologies, Stada, Natco, Curaxys, Oncobiologics, Alteogen, Mabion)
RHOA	N	N	N	N	N	N
RBI	N	N	N	N	SGT-RB94	N
THRA	N	N	N	N	N	N
CBL	N	N	N	N	N	N

Table 16						
Gene	Approved	Pre-registrati on (pre-approval)	Phase III	Phase II	Phase I	preclinical
ALK	crizotinib	N	N	AP-26113; RG-7853; LDK-378; TSR-011; NMS-E628	X-396; ASP-3026	NMS-E628; aurora kinase + ALK inhibitor (Sareum, AstraZeneca) ; ALK inhibitors (AstraZeneca , Cephalon, Aurigene); ARN-5032; DLX-521

Table 17: Druggability status for Table 3 genes/fusions

Gene	approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	preclinical
ESR1	estramustine phosphate sodium; ethinyl estradiol sulfonate; fulvestrant; raloxifene hydrochloride; tamoxifen; toremifene citrate; trilostane;	N	acolbifene	TAS-108; estetrol; GTx-758; endoxifen; afimoxifene	icaritin; ARN-810	SR-16388; VAL-201; SERM + toxin (SEEK); estradiol (BHR Pharma); NDC-1407; anticancer MAb (Shenogen)
RPS6KB1	N	N	N	N	AZD-5363; AT-13148; LY-S6KAKT1	p70S6 kinase inhibitors (Sentinel)

Table 19: Gene Fusions

Table 19	5' gene symbol	3' gene symbol	Druggable gene
Cancer Type			
Prostate Adenocarcinoma	ABCD3	DPYD	DPYD
Sarcoma	ACTG2	ALK	ALK
Lung Adenocarcinoma	ADAMTS16	TERT	TERT
Brain Lower Grade Glioma	ATRX	BCL2	BCL2

Table 19 Cancer Type	5' gene symbol	3' gene symbol	Druggable gene
Gastric Adenocarcinoma	B4GALT1	RAF1	RAF1
Gastric Adenocarcinoma	BRD3	LCN2	BRD3
Gastric Adenocarcinoma	CASZ1	MTOR	MTOR
Acute Myeloid Leukemia	CHD1	MTOR	MTOR
Uterine Corpus Endometrioid Carcinoma	CPA6	PTK2	PTK2
Breast invasive carcinoma	DAB1	IL12RB2	IL12RB2
Lung Adenocarcinoma	DDI2	MTOR	MTOR
Sarcoma	FRS2	MDM2	MDM2
Sarcoma	GLIS3	TERT	TERT
Lung Adenocarcinoma	HIF1A	PRKCH	HIF1A
Breast invasive carcinoma	HPRT1	CTPS2	HPRT1
Breast invasive carcinoma	IL12RB2	DAB1	IL12RB2
Breast invasive carcinoma	IL6R	C1orf112	IL6R
Breast invasive carcinoma	KCMF1	PRKDC	PRKDC
Lung Adenocarcinoma	KIF5B	MET	MET
Breast invasive carcinoma	MAPK14	EFHA1	MAPK14
Sarcoma	MDM2	SPATS2	MDM2
Thyroid carcinoma	MTMR12	TERT	TERT
Bladder Urothelial Carcinoma	NOTCH2	EIF2B3	NOTCH2
Sarcoma	NTRK1	DYNC2H1	NTRK1
Kidney renal clear cell carcinoma	PDCD6	TERT	TERT
Lung Adenocarcinoma	PHKB	PDE3A	PDE3A
Uterine Carcinosarcoma	RARA	SLC9A3R1	RARA
Liver hepatocellular carcinoma	SLC12A7	TERT	TERT
Sarcoma	SMARCA4	EEF2	EEF2
Breast invasive carcinoma	STARD13	TNFRSF8	TNFRSF8
Lung Adenocarcinoma	TICAM1	IL12RB1	IL12RB1
Sarcoma	TRIO	TERT	TERT
Prostate Adenocarcinoma	TRPM8	UGT1A9	TRPM8
Sarcoma	TSPAN3	MDM2	MDM2
Breast invasive carcinoma	TTLL7	TERT	TERT
Brain Lower Grade Glioma	USP46	PDGFRA	PDGFRA
Gastric Adenocarcinoma	WNK2	BRD3	BRD3
Cervical squamous cell carcinoma and endocervical adenocarcinoma	ZNF226	AKT2	AKT2

Table 20. Breakpoints for Gene Fusions from Table 19

Table 20	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromo some	5' Break point	3' Gene Symbol	3' Accession	3' Chromo some	3' Break point	Breakpoint Sequence	SEQ ID NO
PDCD6 TERT		Clear Cell Renal Cell Carcinoma	TCGA-BP- 4991-01A-01R- 1334-07	PDCD6	10016	chr5	2728 52	TERT	7015	chr5	1282 548	TTCCTGTGGAACGTT TTCAGAGGGTCGA TAAAGACAGGAGTG GAGTGATATCAGA CAGCACTTGAAGAG GGTGAGCTGCGGG AGCTGTCGGAAGCA GA	200
TSPAN 3 MD M2		Sarcoma	TCGA-DX- A23R-01A- 11R-A26T-07	TSPAN3	10099	chr15	7734 4775	MDM2	4193	chr12	6920 2269	ACCTCTATGCTGAGG GGTGTGAGGCTCTA GTAGTGAAGAAAGCT ACAAGAA CAGGCA AATGTGCAATACCA ACATGCTGTACCTA CTGATGGTGCTGTA A	201
SLC12A 7 TERT		Hepatocel lular Carcinoma	TCGA-BC- A3KG-01A- 11R-A213-07	SLC12A7	10723	chr5	1111 983	TERT	7015	chr5	1282 739	CGGAGGCTCCGGGC ACCCCGAGGGCCC CGAGCCCGAGCGCC CCAGCCCG GGGGT TGGCTGTGTTCCGGC CGCAGAGCACCGTC TGCCTGAGGAGATC CT	202

Table 20	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromo some	5' Break point	3' Gene Symbol	3' Accession	3' Chromo some	3' Break point	Breakpoint Sequence	SEQ ID NO
FRS2 MDM2	Sarcoma		TCGA-DX- A3M1-01A- 11R-A22K-07	FRS2	10818	chr12	6986 4310	MDM2	4193	chr12	6920 2988	GTGGTTACAGCACC ATCAGTAGGTACAG ACATGTTGGTATTGC ACATTTG CCGTCCG CCCAGGTGCTGAGA GGGAGCAGGGCGC GGGTCGGCGGGCGC GA	203
CHD1 MTOR	Acute Myeloid Leukemia		TCGA-AB- 2939-03A-01T- 0740-13	CHD1	1105	chr5	9819 9112	MTOR	2475	chr1	1127 3623	GAATGCTCTAAAAGA GTATACAAATCCTGA ACAAATTAAGCAAT GGAGAAA GAATTC TGGGTCATGAACAC CTCAATTCAGAGCAC GATCATTCTTCTCAT	204
CHD1 MTOR	Acute Myeloid Leukemia		TCGA-AB- 2939-03A-01T- 0740-13	CHD1	1105	chr5	9820 4199	MTOR	2475	chr1	1127 3623	TCCCAATTTCTGAAG AATCTGAAGAGCTG GATCAGAAGACATT CAGCATT GAATTC GGGTCATGAACACC TCAATTCAGAGCAC GATCATTCTTCTCAT	205
CHD1 MTOR	Acute Myeloid Leukemia		TCGA-AB- 2939-03A-01T- 0740-13	CHD1	1105	chr5	9819 9112	MTOR	2475	chr1	1127 3623	AATGAGAAGAATGA TCGTGCTCTGAATTG AGGTGTCATGACCC AGAATT CTTCTCC ATTGCTTAATTTGTT CAGGATTTGTATACT CTTTAGACATT	206

Table	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
20 Fusion Name												
MAPK14 EFH A1	Invasive Breast Carcinoma	TCGA-AO-A129-01A-21R-A10J-07	MAPK14	1432	chr6	3604 4379	EFHA1	221154	chr13	2211 3824	GGGATGCATAATGG CCGAGCTGTTGACT GGAAGAACATTGTT TCCTGGTA AAACTT CAGTCAAGAAAGCTG ACAAAAAGGACAT CGAGGATACACTGT CA	207
TICAM1 IL12 RB1	Lung Adenocarcinoma	TCGA-05-4426-01A-01R-1206-07	TICAM1	148022	chr19	4831 636	IL12RB1	3594	chr19	1818 0463	GTCCTGGCCCCACAG GCTGCCATTCAATGC AATACGTCATGCTCT GAGCCC GGGCTGC CGGCTGCGCCACTG GGTCCTGGGGTCTT GGGGGCTGGGGCTT C	208
TICAM1 IL12 RB1	Lung Adenocarcinoma	TCGA-05-4426-01A-01R-1206-07	TICAM1	148022	chr19	4831 630	IL12RB1	3594	chr19	1818 2962	CCACTGGTCTGTGT GGGTGTCGGCAGGA ATGTGCCACGTCTG GTTCAGG GATCCG GGGTGCCGCTGC GCCACTGGGTCTTG GGGTCTGGGGGCT GG	209
DAB1 IL12RB2	Invasive Breast Carcinoma	TCGA-AN-A0AM-01A-11R-A034-07	DAB1	1600	chr1	5761 1102	IL12RB2	3595	chr1	6784 5789	CCCTTCACCTTTAAA CCTCTTTATCAAAGT GGCTTCACTGCGATC CTGAC GGGAATTTT GTCTGCAAGGTGAG AGGCAGTGTAAAGG ATGATGAGTCCAC	210

Table	Cancer	TCGA Tumor	5' Gene	5'	5'	5'	5'	3'	3'	3'	3'	Breakpoint	SEQ
20	Type	Sample	Symbol	Accession	Chromo	Break	Gene	Accession	Chromo	Break	Sequence	ID NO	
Fusion		Barcode			some	point	Symbol		some	point			
Name													
IL12RB	Invasive	TCGA-AN-	IL12RB2	3595	chr1	6784	DAB1	1600	chr1	5761	CTGCTGGTGAAGT	211	
2 DAB	Breast	A0AM-01A-				5806				1102	TCCACGGAATGA		
1	Carcinoma	11R-A034-07									GAGGGAATTTGTCT		
											GCAAGGT CAGGAT		
											CGCAGTGAAGCCAC		
											TTTGATAAAGAGGTT		
											TAAAGGTGAAGGGG		
											T		
IL12RB	Invasive	TCGA-AN-	IL12RB2	3595	chr1	6784	DAB1	1600	chr1	5761	TCTCCAAAATTAC	212	
2 DAB	Breast	A0AM-01A-				5733				1052	ATCCAATAAACAGCC		
1	Carcinoma	11R-A034-07									TGCAGCCCCGAGTG		
											ACATAT GTCCGGTA		
											CAAAGCCAAAATTGA		
											TCGGGATTGATGAA		
											GTTCCGCAGCTCG		
GLIS3	Sarcoma	TCGA-DX-	GLIS3	169792	chr9	4117	TERT	7015	chr5	1282	CTGCTGATCCACATG	213	
TERT		A3LS-01A-11R-				768				739	AGAGTCCACTCTGG		
		A21T-07									GGAGAAGCCCCAACA		
											AGTGATAC GGGGTT		
											GGCTGTGTTCCGGC		
											CGCAGAGCACCGTC		
											TGCGTGAGGAGATC		
											CT		
ADAM	Lung	TCGA-44-	ADAMTS	170690	chr5	5191	TERT	7015	chr5	1282	GATACAGGCTCTGG	214	
TS16 T	Adenocar	2662-01A-01R-	16			903				739	ACTGGCCTTCACCAT		
ERT	cinoma	0946-07									TGCCCATGAGTCTG		
											GACACAA GGGTTG		
											GCTGTGTTCCGGCC		
											GCAGAGCACCGTCT		
											GCGTGAGGAGATCC		
											TG		
ABCD3	Prostate	TCGA-CH-	ABCD3	5825	chr1	9495	DPVD	1806	chr1	9798	CTTTAGCAACGCCAA	215	
DPVD	Adenocar	5764-01A-21R-				6803				1497	ATGGAGATGTTTG		

Table 20 Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromo some	5' Break point	3' Gene Symbol	3' Accession	3' Chromo some	3' Break point	Breakpoint Sequence	SEQ ID NO
	cinoma	1580-07									ATCCGAGACCTTAAT TTTGAA TCACAATA TGGAGCTTCGGTTTC TGCCAAGCCTGAACCT ACCCCTCTTTTA	
SMARC A4 EEF 2	Sarcoma	TCGA-K1- A3PO-01A- 11R-A21T-07	SMARCA 4	6597	chr19	1115 1982	EEF2	1938	chr19	3983 208	TCTGCCGGACCTCCT CTTCGATCTCCTCCA GCGTGCCCTCCTCGA TGGCC CAACTCAT TGACTCCCCGGGC ATGTCGACTTCTCCT CGGAGGTGACTG	216
ZNF22 6 AKT2	Cervical Squamous Cell Carcinoma	TCGA-IR-A3LH- 01A-21R- A213-07	ZNF226	7769	chr19	4466 9953	AKT2	208	chr19	4074 8529	ATTGAGCCCTGACTT CTCAAAAAGCACTG CACAGAGGAGGAG GCAGCAGA ACCCC ATGGACTACAAGTG TGGCTCCCCCAGTGA CTCCTCCACGACTGA G	217
ZNF22 6 AKT2	Cervical Squamous Cell Carcinoma	TCGA-IR-A3LH- 01A-21R- A213-07	ZNF226	7769	chr19	4466 9953	AKT2	208	chr19	4074 8529	AATTCTCCTGACTT CTCAAAAAGCACTG CACAGAGGAGGAG GCAGCAGA ACCCC ATGGACTACAAGTG TGGCTCCCCCAGTGA CTCCTCCACGACTGA G	218

Table 20	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
ACTG2 ALK	Sarcoma		TCGA-IW- A3M6-01A- 11R-A21T-07	ACTG2	72	chr2	7412 8558	ALK	238	chr2	2944 6380	GAGATGATGCCCC CGGGCTGTCTCCCC TCCATTGTGGCCGC CCTCGC CACCAGGA GCTGCAAGCCATGC AGATGGAGCTGCAG AGCCTGAGTACAA	219
ACTG2 ALK	Sarcoma		TCGA-IW- A3M6-01A- 11R-A21T-07	ACTG2	72	chr2	7412 8564	ALK	238	chr2	2944 9940	ATGCCCCCGGGCT GTCTTCCCTCCATT GTGGCCCGCCCTCG CCACCAG TGATGG AAGCCACGGGGAA GTGAATATTAAGCAT TATCTAAACTGCAGT	220
ACTG2 ALK	Sarcoma		TCGA-IW- A3M5-01A- 22R-A21T-07	ACTG2	72	chr2	7412 8564	ALK	238	chr2	2944 9940	TGATGCCCGGGG CTGTCTTCCCTCCA TTGTGGCCCGCCCTC GCCACC AGTGATG GAAGCCACGGGGA AGTGAATATTAAGC ATTATCTAAACTGCA	221
CASZ1 MTOR	Gastric Adenocar cinoma		TCGA-BR- 8590-01A-11R- 2402-13	CASZ1	54897	chr1	1076 5549	MTOR	2475	chr1	1128 8975	ATGAAGTGACACCC CCAGCTACATCCGA GGAGTTCTAGGAC CTGCTACG AGCTGA CTATAGCACTAGTGA AATGCTGGTCAACAT GGGAAACTTGCCTC	222

Table 20 Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
DDI2 MTOR	Lung Adenocar cinoma	TCGA-MP- A4SW-01A- 21R-A24X-07	DDI2	84301	chr1	1594 4303	MTOR	2475	chr1	1122 7574	ATTCTAACACTCCGG CCGCTGCCTCCGGCT GCTGTAGCTTATTAT TAATG CTGGCTCTC GGCTGCGGGGATGC CAGACTCGAGCTCG CACAGCGCGCGGA	223
B4GAL T1 RAF 1	Gastric Adenocar cinoma	TCGA-HU- A4GH-01A- 11R-A24K-31	B4GALT1	2683	chr9	3316 6756	RAF1	5894	chr3	1264 1914	CTGGACAGGGCTGA AGGTGAGGCTGATT CGCTGTGACTTCGAA TTGCATC CAAGCAG CGGGACTCCTCAG GGCAGGCGGGCAGC GACAGTGGGTGGT G	224
HIF1A PRKCH	Lung Adenocar cinoma	TCGA-44- 2668-01A-01R- 0946-07	HIF1A	3091	chr14	6220 7906	PRKCH	5583	chr14	6199 5793	AAAAATCTCATCCAA GAAGCCCTAACGTG TTATCTGTGCTTTG AGTCAA AGAGATCT GAAACTGGACAATG TCCTGTTGGACCACG AGGTCACCTGTAA	225
HIF1A PRKCH	Lung Adenocar cinoma	TCGA-44- 2668-01A-01R- 0946-07	HIF1A	3091	chr14	6220 7766	PRKCH	5583	chr14	6199 5805	CGAAGTCTGCCAGTT TACAGTGACCTCGT GGTCCAAACAGGACA TTGTCC AGTTTCTTT ATGTATGTGGGTAG GAGATGGAGATGCA ATCAATATTTAA	226

Table 20	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromo some	5' Break point	3' Gene Symbol	3' Accession	3' Chromo some	3' Break point	Breakpoint Sequence	SEQ ID NO
HPRT1 CTPS2	Invasive Breast Carcinom a	TCGA-AR- A24W-01A- 11R-A169-07	HPRT1	3251	chrX	1336 2754 2	CTPS2	56474	chrX	1665 7355	GATGATCTCTCAACT TTAACTGGAAAGTCT AGGTTGTTGGCAGA AGATAT GCCCGAG CACAAACCCTGGCAAT TTGGGAGGGAACAAT GAGACTGGGAATAA	227	
HPRT1 CTPS2	Invasive Breast Carcinom a	TCGA-AR- A24W-01A- 11R-A169-07	HPRT1	3251	chrX	1336 0934 0	CTPS2	56474	chrX	1668 5822	ATAAATCTTTGCTG ACCTGCTGGATTACA TCAAAGCACTGAAT AGAAAT AGTGATA GAGTTTGCAAGAAA CTGCCCTTAACCTTGAA AGATGCTGATTCCA	228	
HPRT1 CTPS2	Invasive Breast Carcinom a	TCGA-AR- A24W-01A- 11R-A169-07	HPRT1	3251	chrX	1336 0937 5	CTPS2	56474	chrX	1663 8444	GCACTGAATAGAAA TAGTGATAGATCCAT TCCTATGACTGTAGA TTTTAT GGTGATGT TCCTTTTATAGAAGA AAGACACAGACATC GGTTCGAGGTAAA	229	
HPRT1 CTPS2	Invasive Breast Carcinom a	TCGA-AR- A24W-01A- 11R-A169-07	HPRT1	3251	chrX	1336 2754 2	CTPS2	56474	chrX	1665 7355	GATGATCTCTCAACT TTAACTGGAAAGAA TGTCTTGATTGTGGA AGATAT GCCCGAG CACAAACCCTGGCAAT TTGGGAGGGAACAAT GAGACTGGGAATAA	230	

Table 20	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromo some	5' Break point	3' Gene Symbol	3' Accession	3' Chromo some	3' Break point	Breakpoint Sequence	SEQ ID NO
HPRT1 CTPS2	Invasive Breast Carcinom a	TCGA-AR- A24W-01A- 11R-A169-07	HPRT1	3251	chrX	1336 0936 3	CTPS2	56474	chrX	1668 5820	GATTACATCAAAAGC ACTGAATAGAAATA GTGATAGATCCATTCC CTATGAC TGATAGA GTTTGCAAGAAACT GCCTTAACTTGAAG ATGCTGATTCCACA	231	
IL6R C 1orf11 2	Invasive Breast Carcinom a	TCGA-E9- A1RF-01A- 11R-A157-07	IL6R	3570	chr1	1544 2064 7	C1orf11 2	55732	chr1	1697 9082 0	GGACAGAATCCAGG AGTCCTCCAGCTGA GAACGAGGTGTCCA CCCCCATG CAGGAT AATGCTGACTACAG ATTATTCAGAAAAAC ACTCAAATTGTGTCG	232	
KIF5B MET	Lung Adenocar cinoma	TCGA-93- A4JN-01A-11R- A24X-07	KIF5B	3799	chr10	3230 4500	MET	4233	chr7	1164 1161 7	CCAACTCACCCAAGT GCAATTCGTGGAGG AGGTGCATTGTGTTCA GAACAG AGGATTG ATTGCTGGTGTGTC TCAATATCAACAGCA CTGTTATTACTAC	233	
KIF5B MET	Lung Adenocar cinoma	TCGA-93- A4JN-01A-11R- A24X-07	KIF5B	3799	chr10	3230 6145	MET	4233	chr7	1164 1193 2	GCACTGAAAGAAAGC TAAAGAAAAATGCAT CTCGTGATCGCAAAAC GCTATCA GCAAGA GTACACACTCCTCAT TTGGATAGGCTTGTA AGTGCCCGAAGTGT	234	

Table 20	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
BRD3 LCN2		Gastric Adenocar cinoma	TCGA-HU- A4H2-01A- 11R-A251-31	BRD3	8019	chr9	1369 1742 8	LCN2	3934	chr9	1309 1251 7	GTATGCAGGACTTC AACACCATGTTTACA AATTGTTACATTAT AACCAAG TCCAGG GGAAGTGGTATGTG GTAGGCCTGGCAGG GAATGCAATTCTCAG	235
MDM2 SPATS 2		Sarcoma	TCGA-DX- A1KZ-01A- 11R-A24X-07	MDM2	4193	chr12	6923 3549	SPATS2	65244	chr12	4988 3267	CATTGTCCATGGCAA AACAGGACATCTTAT GGCTGCTTTACATG TGCAA TAGTTCCTA ATAAGAGCAACAAT GAAATTATCCTGGTT TTGCAGCACTTT	236
NOTCH 2 EIF2 B3		Bladder Urothelial Carcinoma	TCGA-FD- A5BS-01A- 21R-A26T-07	NOTCH2	4853	chr1	1204 5896 3	EIF2B3	8891	chr1	4539 2411	CATGCCTACTAGCCT CCCTAACCTTGCCAA GGAGGCAAAAGGATG CCAAGG TGGAGCA GCGTGACTTTCATTGG AGTGGACAGCACAG GAAAGAGGCTGCTC	237
NTRK1 DYNC 2H1		Sarcoma	TCGA-DX- A3LY-01B-11R- A27Q-07	NTRK1	4914	chr1	1568 5140 1	DYNC2H1	79659	chr11	1033 0670 8	AACGCCACAGCATC AAGGATGTGCACGC CCGGCTGCAAGCCC TGGCCAG AAGAT CCCTACAATACCTG AGAGGTCTGTGCCC CGTGCCCTTGAATA	238

Table ID	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
20	PHKB PDE3A	Lung Adenocarcinoma	TCGA-MN-A4N5-01A-11R-A24X-07	PHKB	5257	chr16	4772	PDE3A	5139	chr12	2079	ACTTCAGATCCGTGG	239
							3028				9464		
	USP46 PDGFR	Lower Grade Glioma	TCGA-CS-6665-01A-11R-1896-07	USP46	64854	chr4	5352	PDGFRA	5156	chr4	5514	GTGGAAGCAACCAC TAATATAAACACCTC CCATGTATAGGAAG GCTGGAG CGTTTG GGAAGGTGGTTGAA GGAACAGCCTATGG ATTAAGCCGGTCCCA A	240
							2650				3576		
	USP46 PDGFR	Lower Grade Glioma	TCGA-CS-6665-01A-11R-1896-07	USP46	64854	chr4	5349 4288	PDGFRA	5156	chr4	5514 0771	GGTCAATTTTGGAA ACACATGCTACTGTA ACTCCGTGCTTCAGG CATTGT CCTGGTTG TCATTTGGAACACAG AAACCGAGGTATGA AATTCGCTGGAGGG	241
MTMR12 TERT		Thyroid Gland Carcinoma	TCGA-BJ-A409-01A-11R-A250-07	MTMR12	54545	chr5	3226	TERT	7015	chr5	1282 739	ACATGAAGTACAAA GCAGTGAGTGTCAA CGAAGGCTATAAAG TCTGTGAG AGGGG TTGGCTGTGTTCCGG CCGCAGAGCACCGT CTGCGTGAGGAGAT CC	242
							3219						

Table 20	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromo some	5' Break point	3' Gene Symbol	3' Accession	3' Chromo some	3' Break point	Breakpoint Sequence	SEQ ID NO
ATRX BCL2	Lower Grade Glioma	TCGA-DB- A4XF-01A- 11R-A27Q-07	ATRX	546	chrX	7704 1468	BCL2	596	chr18	6079 5992	AATCAACAGAGGC CGCATGCTGGGGCC GTACAGTTCACAAA GGCATCC TCATGGG CTCAGCGGTCATGTT TTCGCTTGAACGCCT TGTCGGCTTCTGT	243	
												244	
TRPM8 UGT1 A9	Prostate Adenocar cinoma	TCGA-CH- 5766-01A-11R- 1580-07	TRPM8	79054	chr2	2348 9450 9	UGT1A9	54600	chr2	2346 7568 0	CATGTTATCCACAA CATCCTGCTGGTCAA CCTGCTGGTCGCCAT GTTTG GGAAATTGA AGCTACATTAAATGC TTCTGGAGAACATG GAATTGTGGTTT	245	
KCMF1 PRKD C	Invasive Breast Carcinom a	TCGA-EW- A1P4-01A- 21R-A144-07	KCMF1	56888	chr2	8526 2227	PRKDC	5591	chr8	4877 2278	CACAGTCTTTTACTT GTCCTATTGTGGAA AAATGGGCTATACG GAGACA GTACCCT GAGTGAGGAAATGA GTCAATTTGATTCT CAACCGGAGTTCAG	246	
CPA6 PTK2	Endometri al Endometri oid Adenocar cinoma	TCGA-A5- A0G5-01A- 11R-A040-07	CPA6	57094	chr8	6853 6411	PTK2	5747	chr8	1417 7438 9	AAACAGAAGAGGAA GCATATGCACCTGAA GAAAATATCCTATCA ACTTAAG AAACAG ATGATTATGCTGAG ATTATAGATGAAGA AGATACTTACACCAT G	247	
RARA SLC9A3 R1	Carcinosar coma	TCGA-N8- A4PQ-01A- 11R-A28V-07	RARA	5914	chr17	3850 8759	SLC9A3 R1	9368	chr17	7275 8151	ACCATCGCGGACCA GATCACCTCTCTCAA GGCTGCCTGCTCTGG ACATCCT GCGCGAG		

Table 20	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
												CTTCGGCCTCGGCTC TGTACCATGAAGAA GGGCCCCAGTGGCT	
WINK2 BRD3	Gastric Adenocar cinoma	TCGA-HU- A4H2-01A- 11R-A251-31	WINK2		65268	chr9	9594 7892	BRD3	8019	chr9	1369 1054 3	ACAAGGGGCTGGAC ACGAGACCTGGGT GGAGGTGGCCTGGT GTGAGCTG CAGAG GAAGATGGATGGCC GAGAGTACCCAGAC GCACAGGGCTTTGC TGC	248
TRIO T ERT	Sarcoma	TCGA-DX- A1L3-01A-11R- A24X-07	TRIO		7204	chr5	1442 0130	TERT	7015	chr5	1282 739	ATGCCCCACTCCAGA AGTAGCATGGAAT GGAGGGCATCTTCA ACCACAA AGGGGT TGGCTGTGTTCCGGC CGCAGAGCACCGTC TGCCTGAGGAGATC C	249
TTLL7 TERT	Invasive Breast Carcinoma	TCGA-C8- A131-01A- 11R-A115-07	TTLL7		79739	chr1	8446 4614	TERT	7015	chr5	1282 739	CCGCTTGACAGCGGG GACGCGAGGACCCG GGCTGGGCTTTCCTC ACCCGGG GGTTGG CTGTGTTCCGGCCGC AGAGCACCGTCTGC GTGAGGAGATCCTG G	250

Table 20 Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
STARD13 TNFRSF8	Invasive Breast Carcinoma	TCGA-BH-A0C7-01B-11R-A115-07	STARD13	90627	chr13	3385 9649	TNFRSF8	943	chr1	1216 4568	CTCACAGACCGTGTTCTTCTGCGCCGTGCC TGGGAACCTTGACAA TCATCC GGCTCATC CTGTAAGGAGAGCG TCTTGTAAGTCTGATC AAATCGCAAGTAC	251

Table 21. Druggability Status for Table 19 Genes/Fusions

Table 21 Gene	approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	Preclinical
AKT2	N	N	N	N	ARQ-092; BAY-1125976	RX-1792; NT-113; TAS-117
ALK	crizotinib	N	N	AP-26113; RG-7853; LDK-378; TSR-011; NMS-E628	X-396; ASP-3026;	NMS-E628; aurora kinase + ALK inhibitor (Sareum, AstraZeneca); ALK inhibitors (AstraZeneca, Cephalon, Aurigene); ARN-5032; DLX-521
BCL2	N	N	N	PBI-1402; PNT-2258; R-(-)-gossypol; navitoclax; RG-7601	N	VAL-101; BP-100-1.02; sabutoclax

Table 21 Gene	approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	Preclinical
BRD3	N	N	N	N	Y-803	N
DPYD	N	N	N	eniluracil	TAS-114	N
EEF2	denileukin difitox	N	moxetumomab pasudotox	cintredekin besudotox	N	Glioblast-13
FGFR3	ponatinib	asitinib	lenvatinib	dovitinib lactate; ENMD-2076; AZD- 4547	JNJ-42756493; BGJ- 398; LY-2874455; S- 49076	N
HIF1A				camptothecin, Calando	2-methoxyestradiol; SPC-2968	RX-0047; ATSP- 9172; ATSP-9172; P- 3971
HPRT1	Butocin	N	N	N	N	N
IL12RB1	N	N	N	INXN-2001/1001; IL- 12	AS-1409 ; NHS-IL12	N
IL12RB2	N	N	N	IL-12	NHS-IL-12; AS-1409	N
IL6R	tocilizumab	N	ARRY-438162	givinostat; ALX-0061		L-6 inhibitors, Interprotein; IL-6 antagonists, Protagonist Therapeutics; APX- 007
MAPK14	pirfenidone	N	N	ralimetinib	ARRY-614; thioureidobutyronitrile	N

Table 21 Gene	approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	Preclinical
MDM2	N	N	N	N	SAR-405838; RG-7388; RO-5503781; CGM-097; DS-3032	p53-mdm2/mdm4 dual inhibitors. Adamed; PXN-527; ATSP-7041; MDM2 inhibitors, Amgen
MET	cabozantinib; crizotinib	N	tivantinib; rilotumumab; onartuzumab;	MGCD-265; foretinib; ficlatuzumab; BMS- 777607; golitinib; INCB-028060; LY- 2875358; apitolisib	AMG-208; TAS-115; volitinib; SAR-125844; S-49076	X-379; metatinib; PRS-110; ASP- 08001; ARGX-111; DCC-2701; DCC- 2721; MG-516; AL- 2846; CG-206481; T- 1840383; cMet- EGFR dual inhibitors(CrystalGenomics); bispecific antibodies (Hoffmann-La Roche)
MTOR	everolimus; temsirolimus	ridaforolimus	N	quinacrine; XL-765; dactolisib; PKI-587; PF-04691502; CC- 223	P-7170; CBL-137, INK- 128, AZD-2014; CC-115; PWT-33957; DS-7423; GDC-0084; DS-3078; LY-3023414; PI3 kinase/mTOR inhibitor, Lilly	nPT-MTOR; SB2343; STP-503; X-480; ABTL-0812; X-414; CC214; HMPL-518; PQR-309; PQR-401; mTOR inhibitor/PI3 kinase inhibitor, Lilly- 1; PIM/PI3k/mTOR inhibitors, Inflection Biosciences
NOTCH2	N	N	N	OMP-59R5	N	N
NTRK1	N	N	N	milciclib maleate	N	tyrosine kinase inhibitors (Bristol- Myers Squibb); PLX- 7486

Table 21 Gene	approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	Preclinical
PDE3A	amrinone; anagrelide hydrochloride; enoximone; cilostazol; lopinone hydrochloride; lopinone hydrochloride; lopinone hydrochloride	N	N	parogrelli hydrochloride; K-134; RPL-554; cilostazol, Genovate	CR-3465	CLC-2001
PDGFRA	imatinib mesilate; pazopanib; sunitinib; dasatinib; nilotinib; regorafenib	nintedanib	orantinib; motesanib; linifanib	ENMD-2076; olaratumab; X-82; crenolanib;	N	DCC-2618; CG- 206481
PRKDC	N	N	vosaroxin	N	SF-1126, Dbait; CC-115	N
PTK2	N	N	N	defactinib	GSK-2256098; CEP- 37440; BI-853520; VS- 4718	CFAK-C4; FAK inhibitor; Verastem; CTX-0294945; x- 0294886
RAF1	sorafenib	N	N	iCo-007; XL-281	RO-5126766; MLN-2480	BIB-024; STP503; DP-4978; HM-95573; TAK-632
RARA	tamibarotene	N	N	IRX-5183	N	N

Table 21 Gene	approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	Preclinical
TERT	N	N	GV-1001	VX-001; GX-301-	TeloB-Vax	telomerase vaccine, Geron; hTERT DNA vaccine, Inovio
TNFRSF8	brentuximab vedotin	N	N	AFM-13; XmAb-2513	N	N
TRPM8	N	N	N	N	D-3263	N

Table 22. Cancer Types Newly Associated with Gene Fusions

Cancer Type	Gene A	Gene B	Orientation (5'/3')	Druggable gene	Cancer type precedent
Papillary renal cell carcinoma	FGFR3	TACC3	FGFR3/TACC3	FGFR3	Bladder cancer; Squamous cell lung cancer; Glioblastoma; Head & Neck squamous cell carcinoma; Cervical squamous cell carcinoma; Low grade glioma
Squamous cell Lung Carcinoma	SEC16A	NOTCH1	SEC16A/NOTCH1	NOTCH1	Breast Cancer; Thyroid Gland Carcinoma

Table 23. Breakpoints of Gene Fusions from Table 22

Table 23 Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromo some	5' Break point	3' Gene Symbol	3' Accession	3' Chromo some	3' Break point	Breakpoint Sequence	SEQ ID NO
-------------------------------	----------------	---------------------------------	-------------------	--------------	-------------------	-------------------	-------------------	--------------	-------------------	-------------------	------------------------	-----------------

Table ID	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
23	Fusion Name											
FGFR3 TACC3	Papillary Renal Cell Carcinoma	TCGA-A4-7287-01A-11R-2139-07	FGFR3	2261	chr4	1808661	TACC3	10460	chr4	1741429	TCCTCACA CCTGCTCC TCAGCTCC CGGTTCTC CTCCTGTG TCGCCCTT AC GTCGG TGGACGTC ACGGTAAG GACACGGT CCAGGTCC TCCACCAG CTGCT	252
FGFR3 TACC3	Papillary Renal Cell Carcinoma	TCGA-A4-7287-01A-11R-2139-07	FGFR3	2261	chr4	1808633	TACC3	10460	chr4	1741500	GCCGCGCC CTCCCAGA GGCCCACC TTC AAGCA GCTGGTGG AGGACCTG GA ACTGG GGAAGATC ATGGACAG GTTCGAAG AGGTTGTG TACCAGGC CATGG	253

Table ID	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
23	FGFR3 TACC3	Papillary Renal Cell Carcinoma	TCGA-A4-7287-01A-11R-2139-07	FGFR3	2261	chr4	1808661	TACC3	10460	chr4	1741429	AGCAGCTG GTGGAGG ACCTGGAC CGTGTCTT TACCGTGA CGTCCACC GAC GTAA AGGCGACA CAGGAGG AGAACCGG GAGCTGAG GAGCAGGT GTGAGGA	254
3	FGFR3 TACC3	Papillary Renal Cell Carcinoma	TCGA-A4-7287-01A-11R-2139-07	FGFR3	2261	chr4	1808637	TACC3	10460	chr4	1742650	CGCCCTCC CAGAGGCC CACCTTCA AGCAGCTG GTGGAGG ACCTGGAC CGT GTCC TTCTCCGA CCTCTTCA AGCGTTTT GAGAAACA GAAAGAG GTGATCG	255

Table ID	Fusion Name	Cancer Type	TCGA Tumor Sample Barcode	5' Gene Symbol	5' Accession	5' Chromosome	5' Break point	3' Gene Symbol	3' Accession	3' Chromosome	3' Break point	Breakpoint Sequence	SEQ ID NO
23	FGFR3 TACC3	Papillary Renal Cell Carcinoma	TCGA-A4-7287-01A-11R-2139-07	FGFR3	2261	chr4	1808561	TACC3	10460	chr4	1741689	GAGGGCC ACCGCATG GACAAGCC CGCCAAC GCACACAC GACCTGTA CAT GATC ATGGACAG GTTCGAAG AGGTTGTG TACCAGGC CATGGAGG AAGTTC	256
SEC16A NOTCH1	Squamous Cell Lung Carcinoma	TCGA-NC-A5HK-01A-11R-A26W-07	SEC16A	9919	chr9	139352036	NOTCH1	4851	chr9	139418396	GTACGCC AGTCCCTG GGTGCCGA GACCTGCC CCCTGCCT AGTTTCCA GG ACCCC AACCCGTG CCTCAGCA CCCCCTGC AAGAACGC CGGGACAT GCCAC	257	

Table 24. Druggability Status of Genes/Fusions of Table 22

Gene	approved	Pre-registration (pre-approval)	Phase III	Phase II	Phase I	preclinical
FGFR3	ponatinib	masitinib	lenvatinib	dovitinib lactate; ENMD- 2076; AZD- 4547	JNJ-42756493; BGJ-398; LY- 2874455; S- 49076	N
NOTCH1	N	N	N	N	OMP-52M51	Debio-0826; TR-4; Notch antibody (AVEO); Notch1 inhibitors (Interprotein); BMS871; NTR-4

Table 39

Cancer	Event type	Q	No. positive	Total no. of patients	Cytoband	Genes (Entrez ID)	Druggabl e genes	KM evidence
Endometrial Endometrioid Adenocarcinoma	Fusion	2.18E-03	5	258	11p15.5, 4p13	RPLP2 (6181), ATP8A1 (609542)		poor outcome
Cervical Squamous Cell Carcinoma	Fusion	3.56E-03	5	54	17q21.2	KRT15 (3866), KRT19 (3880)		poor outcome
Colorectal Adenocarcinoma:KRAS Mutation	Loss of Function Mutation	9.69E-03	4	105	11q22- q23	ATM (472)		poor outcome
Ductal Breast Carcinoma:ER Positive and HER2 Negative	Fusion	1.46E-02	7	265	17p11.2, 17p13	USP22 (23326), MYH10 (160776)		poor outcome
Endometrial Endometrioid Adenocarcinoma:Microsatel lite Stable	In-Peak Gene Amplificat	3.40E-02	8	171	3q26.2	MECOM (2122)		poor outcome

	ion											
Endometrial Endometrioid Adenocarcinoma	Loss of Function Mutation	5.04E-02	4	188	16p13.3	CREBBP (1387)						poor outcome
Cutaneous Melanoma	Gain of Function Mutation	6.69E-02	5	214	7q34	PRSS37 (136242)						poor outcome
Endometrial Serous Adenocarcinoma	In-Peak Gene Deletion	7.52E-02	4	94	8p11.2	FKSG2 (59347)						poor outcome
Cutaneous Melanoma	Gain of Function Mutation	7.94E-02	5	214	6p21.3	STK19 (8859)						poor outcome
Endometrial Serous Adenocarcinoma: Microsatellite Stable	Loss of Function Mutation	8.05E-02	30	38	17p13.1	TP53 (7157)			TP53			favorable outcome
Colorectal Adenocarcinoma: KRAS Mutation, Stage 3 or 4	In-Peak Gene Amplification	8.58E-02	4	45	13q12.3	CDX2 (1045)						poor outcome
Colorectal Adenocarcinoma: KRAS Mutation	Loss of Function Mutation	8.77E-02	4	105	18q21.1	SMAD4 (4089)						poor outcome
Colorectal Adenocarcinoma: Microsatellite Stable	Gain of Function Mutation	9.10E-02	10	21	12p12.1	KRAS (3845)			KRAS (pre-clinical)			poor outcome

Example 7 Identification of Status of TP53

[00663] Advances in both molecular diagnostics and the understanding of cancer biology are raising the bar for clinical trial paradigms with the expectation that more effective patient stratification will improve outcome and expedite approval of effective cancer drugs.

[00664] Mutational status of TP53 has been identified as a predictive biomarker of treatment response and prognosis. For example, TP53 wild-type (WT) patients have been shown to exhibit significantly increased progression-free survival following therapies including adjuvant 5-fluorouracil and cetuximab combination treatments compared to patients harboring TP53 mutations.

[00665] TP53 mutation annotations were obtained from ONCOMINETTM NGS Mutation Browser (Compendia Biosciences, MI). In total 776 patients were assessed for TP53 mutation status; 259 patients contained at least one mutation in TP53 and were annotated as TP53 mutant while 519 patients lacked a detected TP53 mutation and were annotated as TP53 wild type. TP53 wild type and TP53 mutant annotations were then mapped at the patient level to corresponding microarray samples from the TCGA breast dataset. When mutation annotations were mapped to patients with corresponding microarray data, 327 patients were annotated as TP53 wild type and 188 were annotated as TP53 mutant. TP53 wild type and TP53 mutation signatures were generated from a differential expression analysis of the TCGA breast datasets. Gene lists were ranked by p-value according to Student's two class t-test. Genes differentially upregulated in TP53 wild type patients contributed to the TP53 wild type signature whereas genes that were upregulated in TP53 mutant patients contributed to the TP53 mutant signature. Each signature contained the top 1% of ranked genes (n=204). All genes in the TP53 wild type and TP53 mutation signature were highly significant after correcting for false discovery ($Q < 0.0001$). The Q-value was calculated as $(p\text{-value}/p\text{-value rank}) \times \text{number of genes measured}$.

[00666] Five ONCOMINETTM cancer types contained sufficient TP53 mutation status data to complete an analysis. Of these, significantly increased signature expression was found in TP53 WT compared to TP53 mutated clinical samples from breast ($p < 0.001$; n=189 WT, 37 mutant), lung ($p = 0.0003$; n=23 WT, 18 mutated), liver ($p = 0.0069$; n= 74 WT, 11 mutated) and ovarian ($p = 0.05$; n=22 WT, 15 mutated) cancer patients and a trend was found within lymphoma patients ($p = 0.068$; n=65 WT, 16

mutated) (see Figs. 5-7 and 9-10). Table 40 contains the TP53 WT TCGA breast cancer signature.

[00667] The clinically-derived expression signature effectively distinguishes TP53 WT from mutant tumor samples.

[00668] TABLE 40

TP53 WT Signature Genes				
SUSD3	BAG1	ZNF214	USP30	CEP120
DMXL1	ERBB4	SLC24A1	MKL2	CA12
P4HTM	PCP2	AGBL2	SYTL4	SLC7A2
KIF12	C1orf64	NME5	HEXIM2	ANKHD1-EIF4EBP3
ACBD4	TMEM161B	RERG	BRD8	EIF4EBP3
FSIP1	SLC16A6	VEZF1	LOC644189	TMEM128
CAMLG	MLPH	ZNF484	PJA2	HVCN1
FAM47E	LRBA	FBXO38	TCEAL5	TCTN1
C14orf25	EXOC6	LOC100129623	CHIC1	TOX4
USP47	FAM174A	WFS1	RNF135	SEPSECS
POLK	C14orf19	TRIM4	LOC646976	KIAA1370
SPG11	TCEAL3	SLC7A8	XPC	RG9MTD2
TLE3	CCNH	ZC3H6	MED13L	CELSR1
GLIPR1L2	ANXA9	SFRS12	CXXC5	TBC1D9B
PCBD2	TTC8	LOC100131801	C9orf68	
TCEAL4	TCEAL6	GAMT	CACNA1D	
KCTD3	MAN2B2	ABCC8	ANKRD42	
OBFC1	CST5	CRY2	LOC440459	
MRFAP1L1	SCAMP1	LRRC48	PCM1	
GMPR2	PTGER3	ZNF24	C7orf63	
DDB2	CST3	TMEM101	RHBDD1	
TIGD6	PTPRT	NDFIP1	WDFY3	
KIAA0232	RAI2	CHCHD5	REEP5	
TMEM26	GREB1	KCNE4	FUT8	
PCDH19	CCDC103	PGR	ZFYVE1	

EXAMPLE 8

[00669] In accordance with methods of the Example 9, 8,690 clinical samples were analyzed for mutation status and 7,202 clinical samples were analyzed for fusion status.

[00670] Novel mutations and fusion events were identified. Table 44 shows novel mutations identified in CDKN2A and ALK. Table 41-43 identifies novel fusion events identified in the analysis.

[00671] The druggability of each target and available clinical trials are also provided in the Tables.

[00672] The novel mutations and fusion events provide an opportunity not only for further development for target treatments but also potential options for future targets of chemotherapeutics.

EXAMPLE 9

[00673] In accordance with methods of previous Examples 1-8, 8,690 clinical samples were analyzed for mutation status and 7,202 clinical samples were analyzed for mutation status.

[00674] Novel mutation and cancer association events were identified. Tables 45 and 46 show the novel mutations/cancer associations.

[00675] The druggability of each target and likely outcome are also provided in Table 45.

[00676] The novel mutation/cancer association events provide an opportunity not only for further development for target treatments but also potential options for future targets of chemotherapeutics.

Example 10

[00677] In accordance with methods of previous Examples 1-8, 8,690 clinical samples were analyzed for mutation status and 7,202 clinical samples were analyzed for mutation status

[00678] Novel mutation/cancer association events are shown in Table 47.

[00679] Novel fusion events and associated cancers are shown in Table 48. breakpoints and druggability are shown in Table 49. The status of additional drugs relevant to fusion partners are shown in Table 50.

[00680] In certain embodiments, the disclosure provides:

[00681] A kit comprising a set of probes that specifically recognize at least one fusion of two genes chosen from: Table 42 or 49.

[00682] A composition comprising a set of probes that specifically recognize at least one gene fusion of two genes chosen from: Table 42 or 49.

[00683] A set of probes that specifically recognize a nucleic acid comprising at least one of SEQ ID NOs: 1-289.

[00684] An isolated nucleic acid comprising at least one sequence selected from SEQ ID NOs: 1-289.

[00685] A method of detecting bladder urothelial carcinoma, breast carcinoma, endometrial endometrioid adenocarcinoma, colon adenocarcinoma, glioblastoma multiforme, clear cell renal cell carcinoma, papillary renal cell carcinoma, acute myeloid leukemia, brain lower grade glioma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, rectal cutaneous melanoma, and thyroid gland carcinoma in a sample, the method comprising: amplifying a nucleic acid comprising a sequence selected from SEQ ID NOs: 1-289; and detecting the presence of the nucleic acid comprising a sequence selected from SEQ ID NOs: 1-289; wherein detecting the nucleic acid comprising a sequence selected from SEQ ID NOs: 1-289, indicates that one or more of the above cancers is present in the sample.

[00686] A kit comprising a set of probes that specifically hybridize to at least one nucleic acid comprising a break point from Table 42 or 49.

[00687] A set of probes that specifically hybridize to at least one nucleic acid comprising a break point from Tables 42 or 49.

[00688] In the kit, the nucleic acid can be a gene fusion chosen from Table 41.

[00689] A method of detecting bladder urothelial carcinoma, breast carcinoma, endometrial endometrioid adenocarcinoma, colon adenocarcinoma, glioblastoma multiforme, clear cell renal cell carcinoma, papillary renal cell carcinoma, acute myeloid leukemia, brain lower grade glioma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, rectal cutaneous melanoma, and thyroid gland carcinoma in a sample, the method comprising: amplifying a gene fusion chosen from Table 41 or 48 or a fusion gene product; and detecting the presence of the gene fusion chosen from Table 41 or 48 or fusion gene product in the sample; wherein

detecting the presence of the gene fusion chosen from Table 41, indicates that one or more of the above cancers is present in the sample.

[00690] An isolated gene fusion nucleic acid of between 100 and 10,000 nucleotides in length and comprising at least 25 nucleotides on either side of one of the break points in Table 42 or 49.

[00691] The isolated gene fusion nucleic acid can have a gene fusion partner that is two genes chosen from: TPM1 and ALK, PRKAR1A and ALK, NCOA1 and ALK, LPP and CASR, MDM2 and EGFR, FGFR3 and ELAVL3, B2M and GNAS, DOCK8 and JAK2, HNF1B and NOTCH1, NFASC and NTRK1, SSBP2 and NTRK1, SQSTM1 and NTRK1, TBL1XR1 and PIK3CA, AKAP13 and RET, FKBP15 and RET, TBL1XR1 and RET, CEP85L and ROS1, CLCN6 and RAF1, TRAK1 and RAF1, PRKACA and AKT1, PRKACA and AKT2, MLL and FYN, and TTC13 and JAK2, SEC16A and NOTCH1, ERC1 and RET, GTF2IRD1 and ALK, HTATSF1 and BRS3, CDH1 and CCDC132, CCDC132 and CDH1, ERBB2 and SLC29A3, MET and TFG; TFG and MET, NOTCH2 and MNDA, IRF2BP2 and NTRK1, EIF2C2 and PTK2, RARA and HOXB3, STAT3 and ETV4, and GFAP and VIM, VIM and GFAP, TOP1 and C17orf64, and TP53 and KIAA0753.

[00692] In the above method the sample is a patient sample.

[00693] The above method can further comprise diagnosing the patient as having bladder urothelial carcinoma, breast carcinoma, endometrial endometrioid adenocarcinoma, colon adenocarcinoma, glioblastoma multiforme, clear cell renal cell carcinoma, papillary renal cell carcinoma, acute myeloid leukemia, brain lower grade glioma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, rectal cutaneous melanoma, or thyroid gland carcinoma when a nucleic acid comprising a sequence selected from SEQ ID NOs: 1-289 is present in the patient sample.

[00694] A kit comprising a set of probes, wherein the set of probes specifically recognize a gene variant chosen from Table 41, Table 44, Table 45, Table 46, Table 47 or a combination thereof.

[00695] A composition comprising a set of probes that specifically recognize a gene variant from Table 41, Table 44, Table 45, Table 46, Table 47 or a combination thereof.

[00696] A set of probes that specifically recognize a gene variant from Table 41, Table 44, Table 45, Table 46, Table 47 or a combination thereof.

[00697] A method of detecting bladder carcinoma, breast carcinoma, cervical cell carcinoma, colon adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, clear cell renal cell carcinoma, kidney renal papillary cell carcinoma, acute myeloid leukemia, brain lower grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, squamous cell lung carcinoma, ovarian serous adenocarcinoma, prostate adenocarcinoma, rectal adenocarcinoma, cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, or uterine corpus endometrioid carcinoma, the method comprising: amplifying a nucleic acid comprising a variant selected from Table 7 and/or Table 11; and detecting the presence of the variant selected from Table 41, Table 44, Table 45, Table 46, Table 47 or a combination thereof; wherein detecting the variant selected from Table 41, Table 44, Table 45, Table 46, Table 47 or a combination thereof, indicates one or more of the above cancers is present in the sample.

[00698] A kit comprising a set of probes that specifically hybridize to a variant selected from Table 41, Table 44, Table 45, Table 46, Table 47 or a combination thereof.

[00699] A set of probes that specifically hybridize to a gene variant selected from Table 41, Table 44, Table 45, Table 46, Table 47 or a combination thereof.

[00700] The method of claim 17, wherein the sample is a patient sample.

[00701] The method of claim 20, further comprising diagnosing the patient as having bladder carcinoma, breast carcinoma, cervical cell carcinoma, colon adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, clear cell renal cell carcinoma, kidney renal papillary cell carcinoma, acute myeloid leukemia, brain lower grade glioma, liver hepatocellular carcinoma, lung adenocarcinoma, squamous cell lung carcinoma, ovarian serous adenocarcinoma, prostate adenocarcinoma, rectal adenocarcinoma, cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, or uterine corpus endometrioid carcinoma when a variant selected from Table 41, Table 44, Table 45, Table 46, Table 47 or a combination thereof is present in the patient sample.

[00702] An isolated genetic construct comprising an antisense polynucleotide sequence of a sequence selected from SEQ ID NOs:1-289.

[00703] The isolated genetic construct can further comprise a promoter operatively linked to the antisense polynucleotide.

[00704] The isolated genetic construct can be a vector. The vector can be a viral vector

[00705] A method comprising contacting a cell comprising a gene fusion with a drug, wherein the gene fusion is disclosed in Table 41 or Table 49 and the drug is one or more of the drugs in Table 43 or Table 49.

[00706] A method comprising administering to a patient with a gene fusion a drug, wherein the gene fusion is disclosed in Table 41 or Table 49.

[00707] The method of claim 27, wherein the patient is diagnosed with cancer.

[00708] The method of claim 28, wherein the cancer is a cancer type disclosed in Table 41 or Table 49.

[00709] A method comprising contacting a nucleic acid sample from a patient with a reaction mixture comprising two primers, wherein a first primer is complementary to one gene and a second primer is complementary to a second gene, wherein the fusion of the first gene and the second gene is detectable by the presence of an amplicon generated by the first primer and the second primer, wherein the fusion breakpoint is one of the breakpoints of Table 42 or 49, and wherein a patient with an amplicon is administered one or more of the drugs in Table 43 or Table 49.

[00710] A method comprising performing a bioassay to detect one or more gene fusions in a sample from a subject, wherein at least one of the gene fusions is selected from those in Table 42 or Table 49, receiving the results of the bioassay into a computer system, processing the results to determine an output, presenting the output on a readable medium, wherein the output identifies therapeutic options recommended for the subject based on the presence or absence of the gene fusions.

[00711] In the above method, the bioassay can include probes specific for one or more of the breakpoints of Table 42 or Table 49.

[00712] A kit, comprising: a set of probes, wherein each probe specifically hybridizes to a nucleic acid comprising a breakpoint from Table 42 or Table 49.

[00713] In the kit, each non-naturally occurring probe can comprise: a nucleic acid sequence configured to specifically hybridize to the nucleic acid comprising the breakpoint from Table 42 or Table 49, and a detectable moiety covalently bonded to the nucleic acid sequence.

[00714] In the kit, each non-naturally occurring probe can comprise: a deoxyribonucleic acid sequence comprising at least one deoxyuridine (dU) residue in place of a deoxythymidine residue.

[00715] A method, comprising: amplifying a nucleic acid comprising at least one gene fusion from Table 41 or Table 49 from a sample; and detecting the presence of the at least one gene fusion by at least one of: contacting the composition with at least one probe, wherein each probe specifically hybridizes to the nucleic acid, or observing the presence of a non-natural or non-native chemical structure in the nucleic acid; wherein detecting the presence of the at least one gene fusion indicates that at least one cancer from Table 41 or Table 49 is present in the sample.

[00716] A system, comprising: a nucleic acid amplifier configured to amplify a nucleic acid comprising at least one gene fusion from Table 41 or Table 49 from a sample, to yield an amplified nucleic acid; a detector configured to detect the presence of the at least one gene fusion in the amplified nucleic acid by at least one of (i) contacting the composition with at least one probe, wherein each probe specifically hybridizes to the nucleic acid, or (ii) observing the presence of a non-natural or non-native chemical structure in the nucleic acid, and further configured to transmit a detection indication; and a computer system configured to receive the detection indication and determine that at least one cancer from Table 41 or Table 49 is present in the sample, based on the detection indication.

[00717] A non-transitory computer readable program storage unit encoded with instructions that, when executed by a computer, perform a method, comprising: receiving an input comprising at least a cancer type and an event type, wherein the cancer type is selected from Table 44, 45, 46 and/or 47 and the event type is selected from Table 44, 45, 46 and/or 47; querying a database for at least one entry comprising a plurality of fields, wherein the plurality of fields comprises at least one of the cancer type and the event type; transmitting an output comprising at least one field of the plurality from the at least one entry, wherein the at least one field comprises at least one gene, at least one druggable gene, at least one drug targeting the at least one druggable gene, or a prognosis.

[00718] A method, comprising: administering to a patient having at least one gene fusion selected from the gene fusions listed in Table 41 at least one drug selected from the drugs listed in Table 43.

[00719] A method, comprising: contacting a nucleic acid sample from a patient with a reaction mixture comprising a first primer complementary to a first gene and a second primer complementary to a second gene, wherein a fusion of the first gene and the second gene is detectable by the presence of an amplicon generated by the first primer and the second primer, wherein the fusion comprises a breakpoint selected from the breakpoints listed in Table 42 or Table 49.

[00720] In the above method, at least one drug can be selected from the drugs listed in Table 43, 45, 46 and/or 47 is administered to the patient, in response to a detection of the fusion.

[00721] A non-transitory computer readable program storage unit encoded with instructions that, when executed by a computer, perform a method, comprising: receiving RNA sequence data from at least one cancer cell line; running at least one gene fusion caller on the sequence data, to identify possible breakpoints between fused genes in the processed data; filtering said possible breakpoints, to retain candidate breakpoints, wherein each candidate breakpoint is in a 5' untranslated region (UTR) or a coding DNA sequence (CDS) of a functional gene region and each candidate breakpoint does not occur in an intron; annotating the candidate breakpoints with at least one annotation useful in determining a relevance of a gene fusion for at least one of cancer diagnosis, cancer prognosis, or cancer treatment, wherein the gene fusion comprises the candidate breakpoint.

[00722] In the non-transitory computer readable program storage unit, at least one gene fusion caller can be selected from TopHat and deFuse.

[00723] A non-transitory computer readable program storage unit encoded with instructions that, when executed by a computer, perform a method, comprising: receiving mutation data from at least one cancer cell line; annotating the mutation data with at least one of variant classification, variant position, or variant change, to yield annotated mutation data; filtering the annotated mutation data, to yield gene region mutation data; classifying the gene region mutation data as hotspot, deleterious, or other; and nominating a gene comprising the gene region mutation as a gain of function, loss of function, or recurrent other gene, based on the relative frequency of mutations in the gene and the classifications of all gene region mutations in the gene.

[00724] In the non-transitory computer readable program storage unit, the variant classification can be selected from splice site, 3' untranslated region (UTR), 5' UTR,

intron, missense, nonsense, nonstop, silent, frame shift insertion, frame shift deletion, in-frame insertion, in-frame deletion, or non-coding exon.

[00725] In the non-transitory computer readable program storage unit, filtering the annotated mutation data can comprise excluding mutations outside of a known transcript and mutations not associated with a known gene.

[00726] In the non-transitory computer readable program storage unit, classifying the mutation as hotspot can comprise determining its presence at the same variant position in at least three cancer cell lines, wherein the variant classification is missense, nonstop, in-frame insertion, in-frame deletion, or non-coding exon.

[00727] In the non-transitory computer readable program storage unit, wherein classifying the mutation as deleterious can comprise observing the variant classification is nonsense, frame shift insertion, or frame shift deletion.

[00728] A method, comprising: detecting one or more gene fusions in a sample from a subject, to yield gene fusion detection data, wherein at least one of the gene fusions is selected from the gene fusions listed in Table 41 or 49, receiving by a computer system the gene fusion detection data, identifying by the computer system at least one therapeutic option recommended for the subject, based on the gene fusion detection data.

[00729] A system, comprising: a detector configured to (i) detect one or more gene fusions in a sample from a subject, to yield gene fusion detection data, wherein at least one of the gene fusions is selected from the gene fusions listed in Table 41 or 49 and (ii) transmit the gene fusion detection data; and a computer system configured to receive the gene fusion detection data and identify at least one therapeutic option recommended for the subject, based on the gene fusion detection data.

[00730] A gene fusion that is prepared by a process comprising: isolating an RNA molecule comprising the gene fusion; and synthesizing a complementary DNA (cDNA) molecule complementary to the isolated RNA molecule.

[00731] Unless otherwise indicated, all numbers expressing quantities of ingredients, properties such as molecular weight, reaction conditions, and so forth used in the specification and claims are to be understood as being modified in all instances by the term "about." Accordingly, unless indicated to the contrary, the numerical parameters set forth in the specification and attached claims are approximations that may vary depending upon the desired properties sought to be obtained by the present invention.

At the very least, and not as an attempt to limit the application of the doctrine of equivalents to the scope of the claims, each numerical parameter should at least be construed in light of the number of reported significant digits and by applying ordinary rounding techniques. Notwithstanding that the numerical ranges and parameters setting forth the broad scope of the invention are approximations, the numerical values set forth in the specific examples are reported as precisely as possible. Any numerical value, however, inherently contains certain errors necessarily resulting from the standard deviation found in their respective testing measurements.

[00732] The terms “a,” “an,” “the” and similar referents used in the context of describing the invention (especially in the context of the following claims) are to be construed to cover both the singular and the plural, unless otherwise indicated herein or clearly contradicted by context. Recitation of ranges of values herein is merely intended to serve as a shorthand method of referring individually to each separate value falling within the range. Unless otherwise indicated herein, each individual value is incorporated into the specification as if it were individually recited herein. All methods described herein can be performed in any suitable order unless otherwise indicated herein or otherwise clearly contradicted by context. The use of any and all examples, or exemplary language (e.g., “such as”) provided herein is intended merely to better illuminate the invention and does not pose a limitation on the scope of the invention otherwise claimed. No language in the specification should be construed as indicating any non-claimed element essential to the practice of the invention.

[00733] Groupings of alternative elements or embodiments of the invention disclosed herein are not to be construed as limitations. Each group member may be referred to and claimed individually or in any combination with other members of the group or other elements found herein. It is anticipated that one or more members of a group may be included in, or deleted from, a group for reasons of convenience and/or patentability. When any such inclusion or deletion occurs, the specification is deemed to contain the group as modified thus fulfilling the written description of all Markush groups used in the appended claims.

[00734] Certain embodiments of this invention are described herein, including the best mode known to the inventors for carrying out the invention. Of course, variations on these described embodiments will become apparent to those of ordinary skill in the art upon reading the foregoing description. The inventor expects skilled artisans to employ such variations as appropriate, and the inventors intend for the invention to be

practiced otherwise than specifically described herein. Accordingly, this invention includes all modifications and equivalents of the subject matter recited in the claims appended hereto as permitted by applicable law. Moreover, any combination of the above-described elements in all possible variations thereof is encompassed by the invention unless otherwise indicated herein or otherwise clearly contradicted by context.

[00735] Furthermore, numerous references have been made to patents and printed publications throughout this specification. Each of the above-cited references and printed publications are individually incorporated herein by reference in their entirety.

[00736] It is to be understood that the embodiments of the invention disclosed herein are illustrative of the principles of the present invention. Other modifications that may be employed are within the scope of the invention. Thus, by way of example, but not of limitation, alternative configurations of the present invention may be utilized in accordance with the teachings herein. Accordingly, the present invention is not limited to that precisely as shown and described.

[00737] Specific embodiments disclosed herein may be further limited in the claims using consisting of or consisting essentially of language. When used in the claims, whether as filed or added per amendment, the transition term "consisting of" excludes any element, step, or ingredient not specified in the claims. The transition term "consisting essentially of" limits the scope of a claim to the specified materials or steps and those that do not materially affect the basic and novel characteristic(s). Embodiments of the invention so claimed are inherently or expressly described and enabled herein.

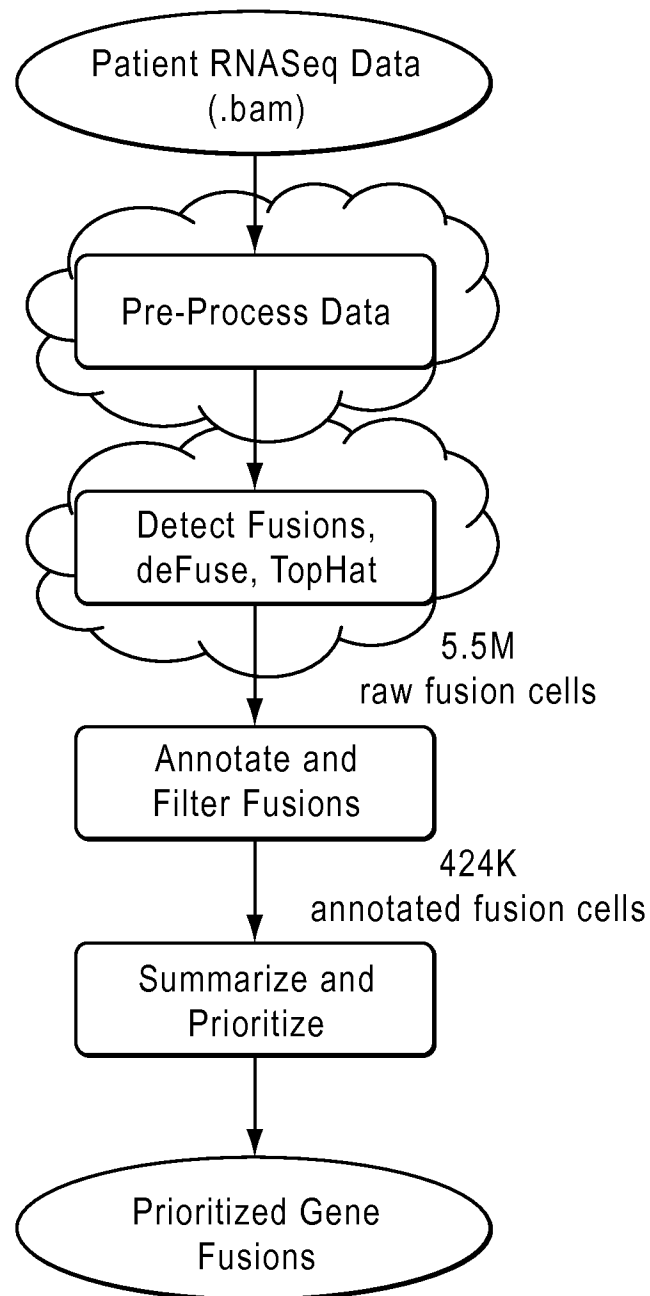
[00738] Reference to sequence identifiers, such as those beginning with NM_, refer to the database accession numbers and the underlying sequences as they were found on April 18, 2013.

What is claimed:

1. A kit comprising a set of probes, wherein the set of probes specifically recognize (a) at least one fusion of two genes chosen from: Tables 42, 49, and a combination thereof; or (b) at least one gene variant from Tables 41, 44, 45, 46, 47, and a combination thereof.
2. A composition comprising a set of probes that specifically recognize (a) at least one fusion of two genes chosen from: Tables 42, 49, and a combination thereof; or (b) at least one gene variant from Tables 41, 44, 45, 46, 47, and a combination thereof.
3. A method of detecting bladder urothelial carcinoma, breast carcinoma, endometrial endometrioid adenocarcinoma, colon adenocarcinoma, glioblastoma multiforme, clear cell renal cell carcinoma, papillary renal cell carcinoma, acute myeloid leukemia, brain lower grade glioma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, rectal cutaneous melanoma, and thyroid gland carcinoma in a sample, the method comprising:
 - amplifying a nucleic acid comprising a sequence selected from SEQ ID NOs: 1-289; and
 - detecting the presence of the nucleic acid comprising a sequence selected from SEQ ID NOs: 1-289;
 - wherein detecting the nucleic acid comprising a sequence selected from SEQ ID NOs: 1-289, indicates that bladder urothelial carcinoma, breast carcinoma, endometrial endometrioid adenocarcinoma, colon adenocarcinoma, glioblastoma multiforme, clear cell renal cell carcinoma, papillary renal cell carcinoma, acute myeloid leukemia, brain lower grade glioma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, rectal cutaneous melanoma, and thyroid gland carcinomas present in the sample.
4. The kit of claim 1, wherein the set of probes specifically hybridize to at least one nucleic acid comprising a break point from Table 42 or 49.
5. The method of claim 3, wherein a gene fusion from Table 49 or 49 is detected:

7. The method of 5, wherein a gene fusion partner is two genes chosen from: TPM1 and ALK, PRKAR1A and ALK, NCOA1 and ALK, LPP and CASR, MDM2 and EGFR, FGFR3 and ELAVL3, B2M and GNAS, DOCK8 and JAK2, HNF1B and NOTCH1, NFASC and NTRK1, SSBP2 and NTRK1, SQSTM1 and NTRK1, TBL1XR1 and PIK3CA, AKAP13 and RET, FKBP15 and RET, TBL1XR1 and RET, CEP85L and ROS1, CLCN6 and RAF1, TRAK1 and RAF1, PRKACA and AKT1, PRKACA and AKT2, MLL and FYN, and TTC13 and JAK2, SEC16A and NOTCH1, ERC1 and RET, GTF2IRD1 and ALK, HTATSF1 and BRS3, CDH1 and CCDC132, CCDC132 and CDH1, ERBB2 and SLC29A3, MET and TFG; TFG and MET, NOTCH2 and MNDA, IRF2BP2 and NTRK1, EIF2C2 and PTK2, RARA and HOXB3, STAT3 and ETV4, and GFAP and VIM, VIM and GFAP, TOP1 and C17orf64, and TP53 and KIAA0753.
8. The method of claim 5, wherein the sample is a patient sample.
9. The method of claim 3, further comprising diagnosing the patient as having bladder urothelial carcinoma, breast carcinoma, endometrial endometrioid adenocarcinoma, colon adenocarcinoma, glioblastoma multiforme, clear cell renal cell carcinoma, papillary renal cell carcinoma, acute myeloid leukemia, brain lower grade glioma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, rectal cutaneous melanoma, or thyroid gland carcinoma when a nucleic acid comprising a sequence selected from SEQ ID NOs: 1-289 is present in the patient sample.
10. The method of claim 9, further comprising administering to a patient with a gene fusion a drug, wherein the gene fusion is disclosed in Table 41 or Table 49.

1/12



Gene Fusion RNASeq data
processing workflow

FIG. 1

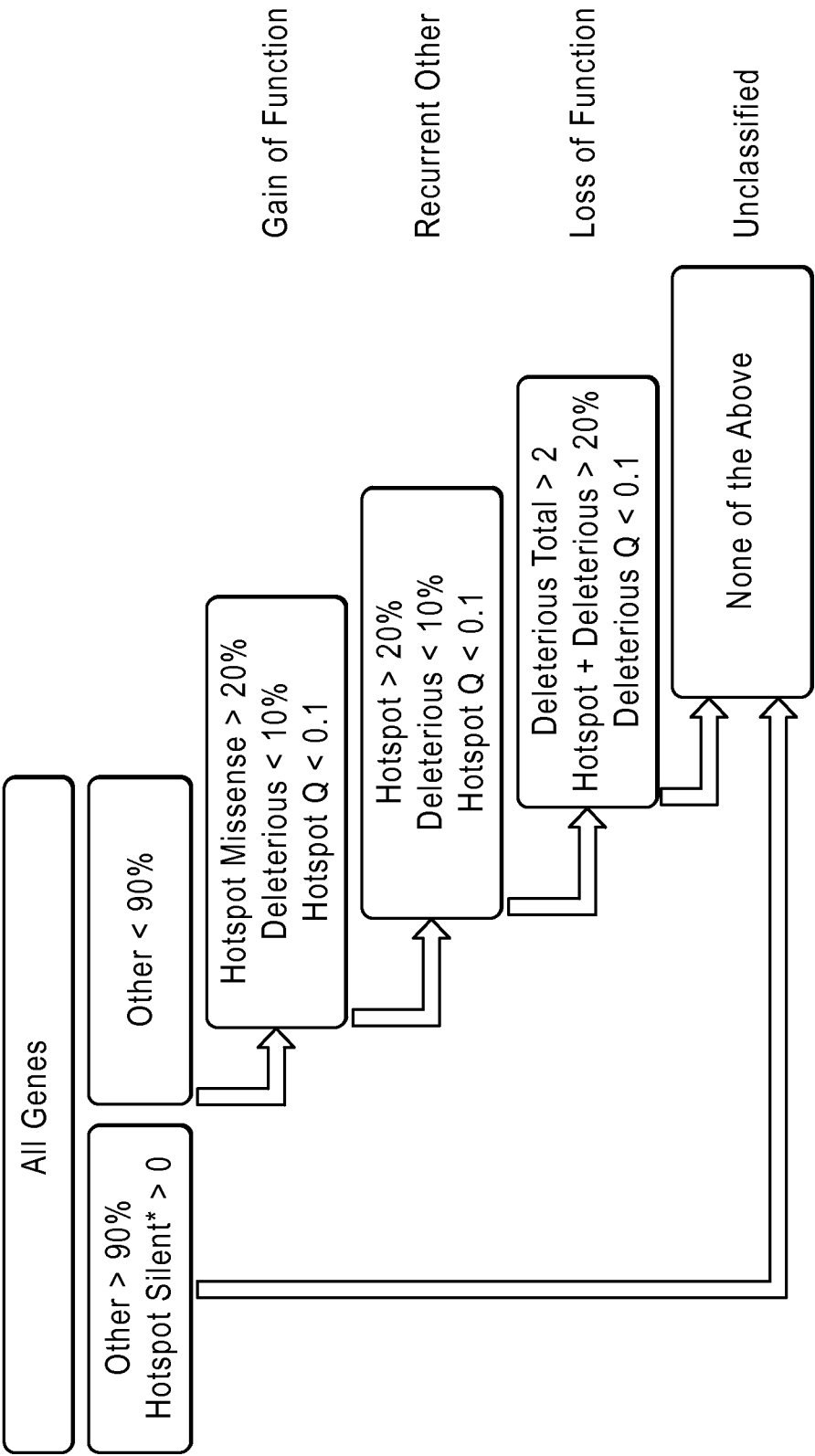


FIG. 2

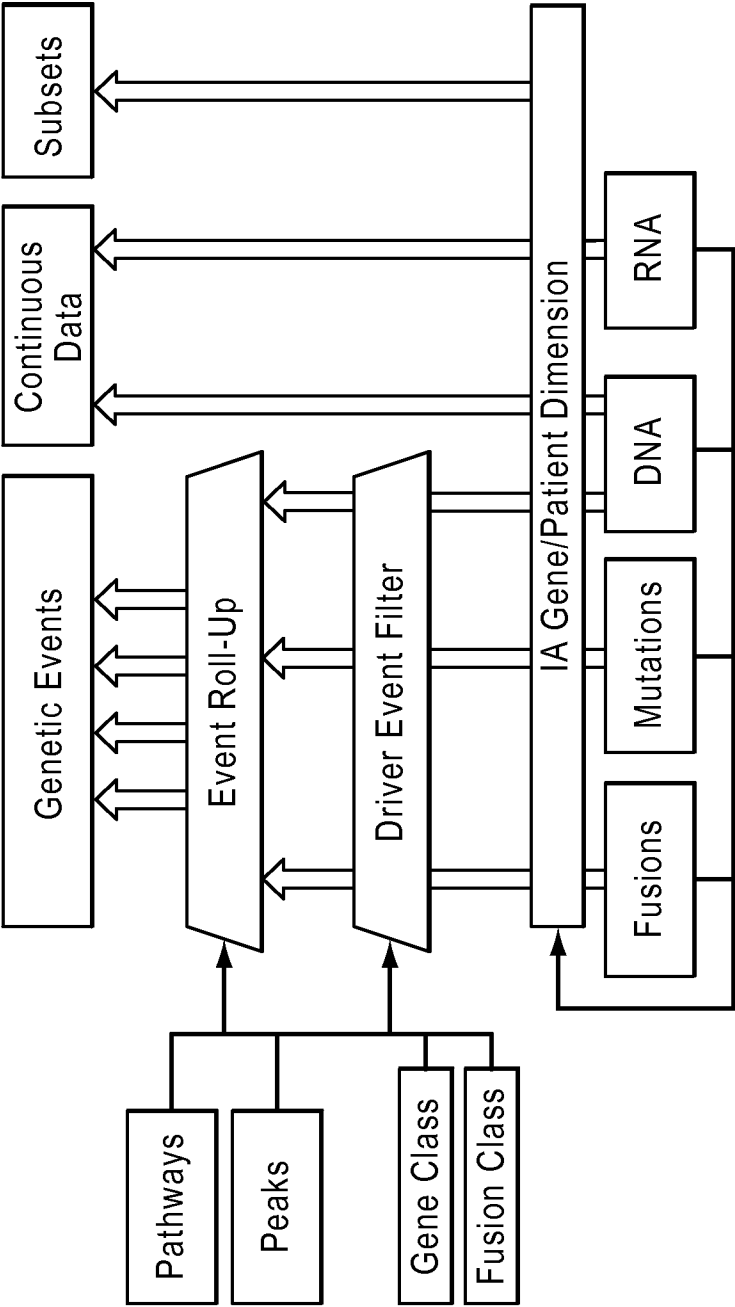


FIG. 3

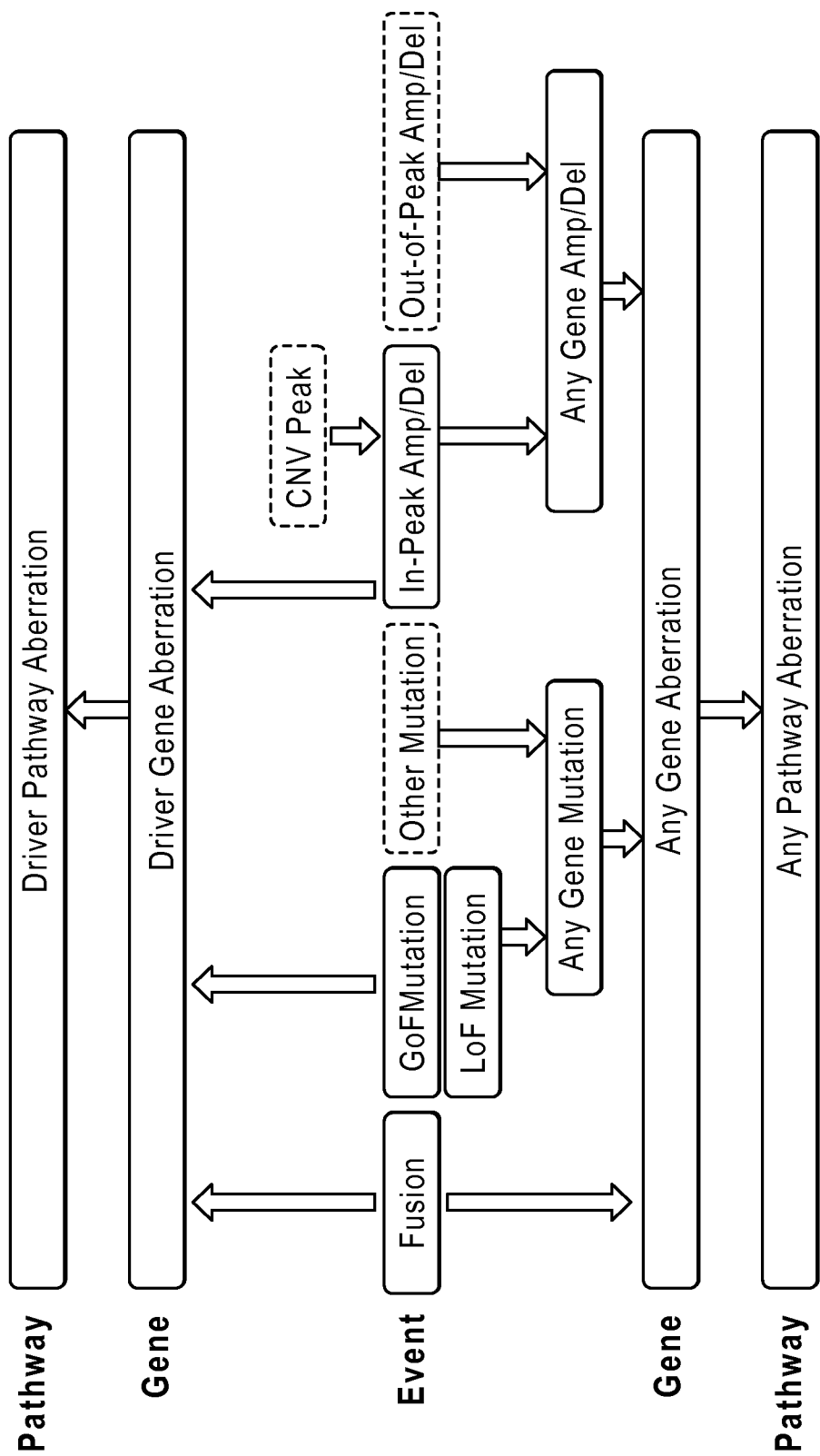


FIG. 4

TP53 wildtype (WT) expression signature is significantly elevated in
TP53 WT breast cancer compared to breast cancer samples harboring a
TP53 point mutation

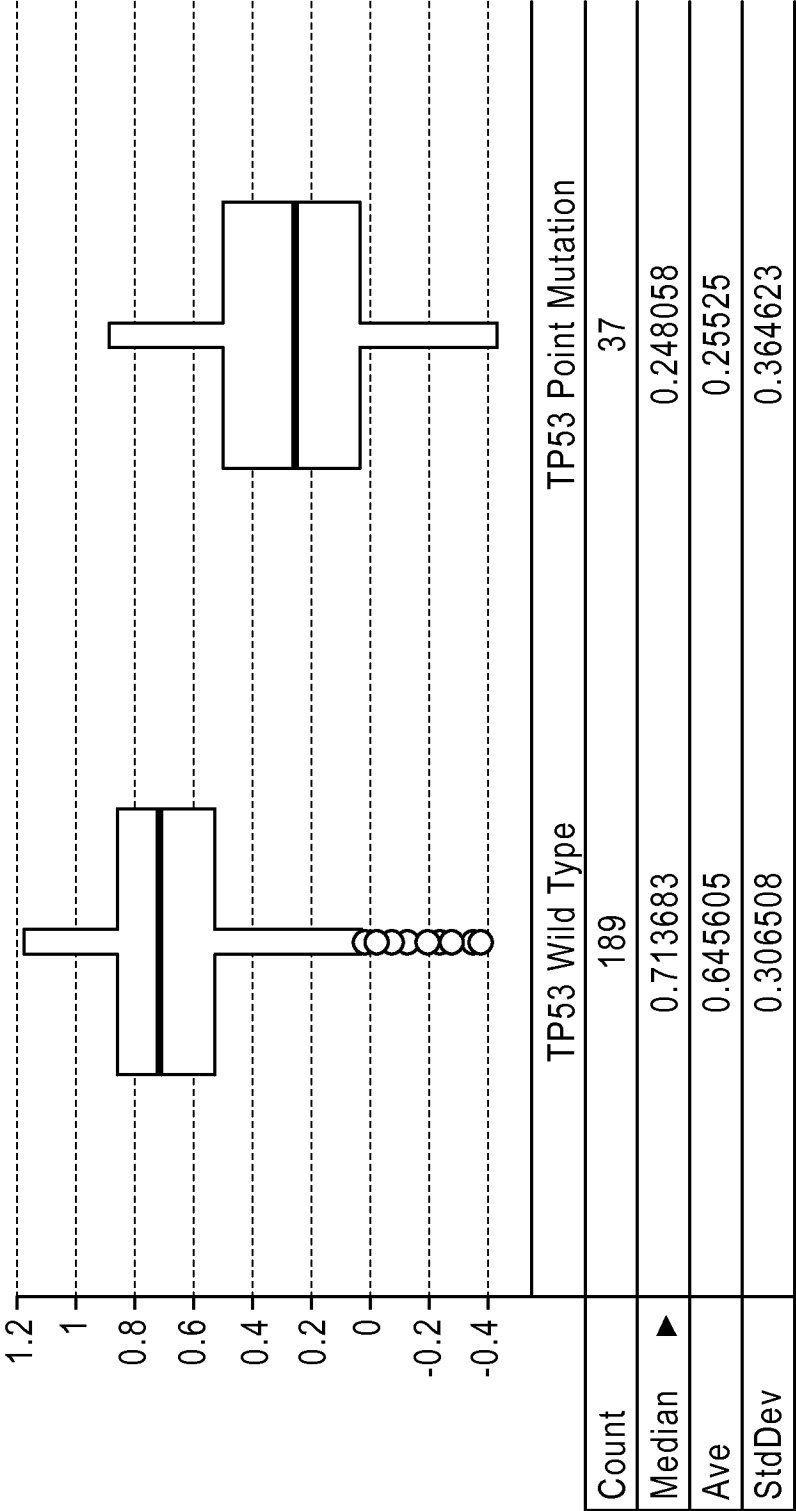


FIG. 5

TP53 wildtype (WT) expression signature is significantly elevated in
TP53 WT lung cancer compared to lung cancer samples harboring a
TP53 mutation

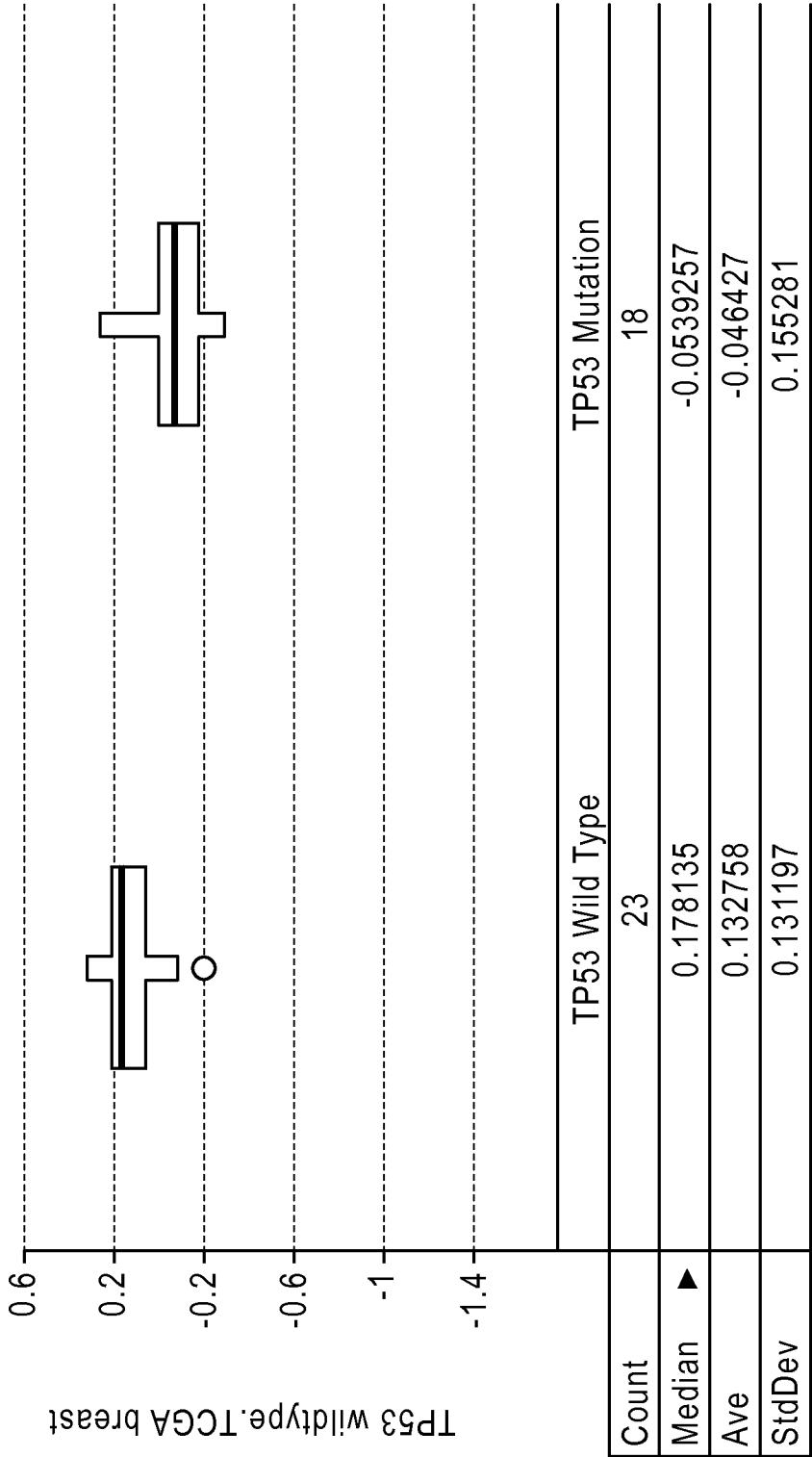


FIG. 6

TP53 wildtype (WT) expression signature is significantly elevated in
TP53 WT ovarian cancer compared to ovarian cancer samples
harboring a TP53 mutation

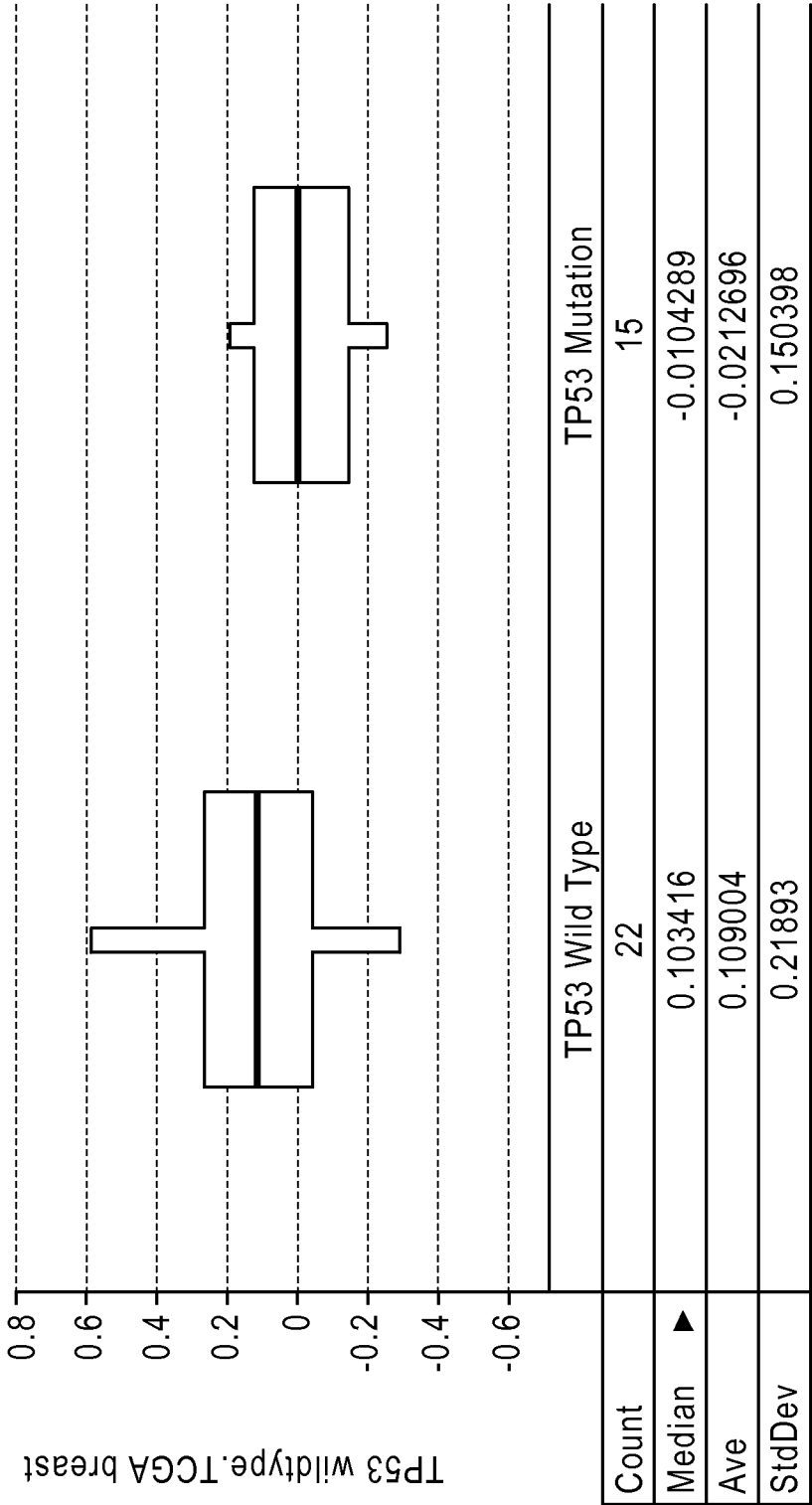
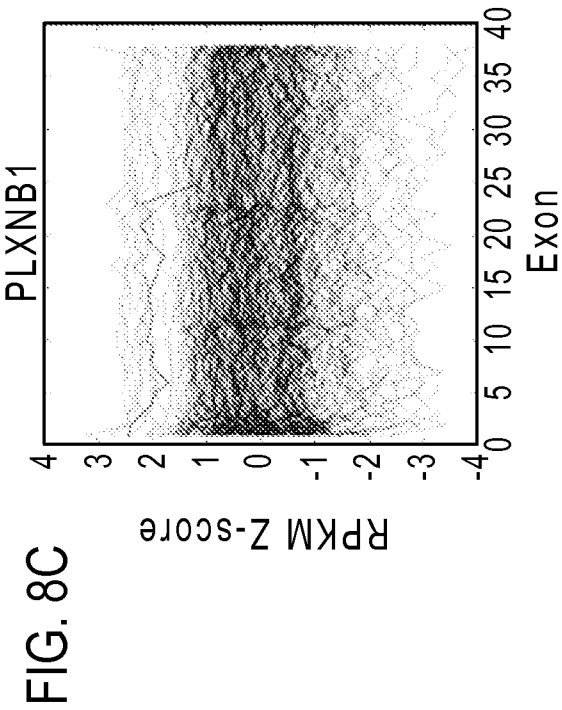
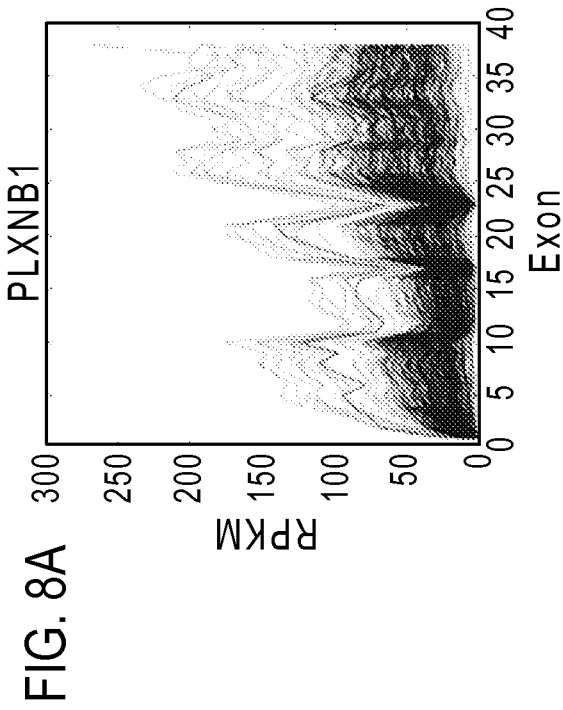
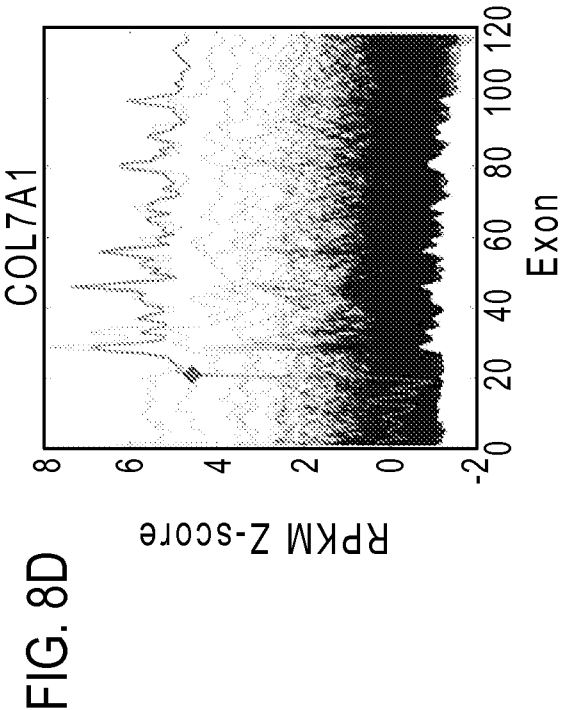
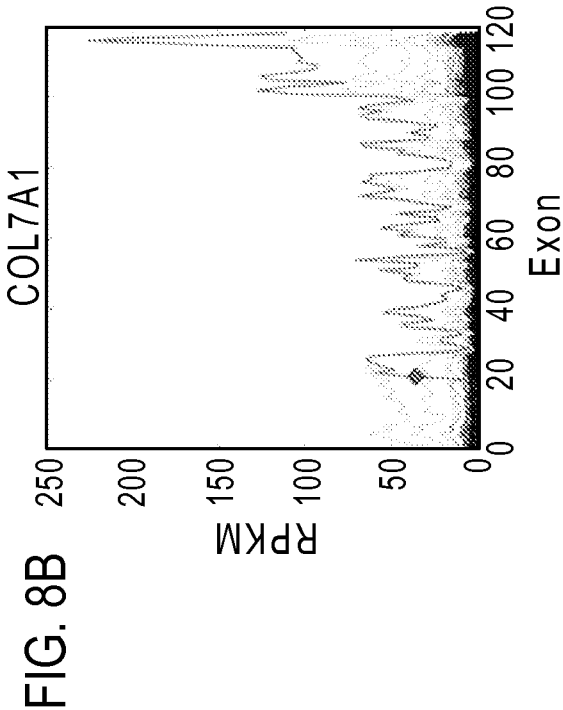


FIG. 7



9/12

Variant position	Human Genome Build and Nucleic Acid Start Positions of Variant Alleles	Pan-cancer Patient Count
p.R273	36_7517845, 36_7517846, 37_7577120, 37_7577121	134
p.R248	36_7518263, 36_7518264, 37_7577538, 37_7577539	95
p.R175	36_7519131, 36_7519132, 37_7578406, 37_7578407	69
p.G245	36_7518272, 36_7518273, 37_7577547, 37_7577548	45
p.R213	36_7518937, 37_7578212	41
p.H179	36_7519118, 36_7519119, 37_7578393, 37_7578394, 37_7578395	40
p.Y220	36_7518915, 37_7578190, 37_7578191	38
p.H193	36_7518996, 36_7518997, 37_7578271, 37_7578272	34
p.R282	36_7517819, 37_7577093, 37_7577094	34
p.R196	36_7518988, 37_7578263	33
p.R249	36_7518259, 36_7518261, 37_7577534, 37_7577535, 37_7577536	30
p.R158	36_7519183, 37_7578457, 37_7578458	27
p.R195	36_7518990, 36_7518991, 37_7578264, 37_7578265, 37_7578266	24
p.C176	36_7519128, 37_7578403, 37_7578404	23
p.G266	36_7517866, 36_7517867, 37_7577141, 37_7577142	22
p.P278	36_7517830, 36_7517831, 37_7577105, 37_7577106	22
p.T125	36_7520037, 36_7520038, 37_7579312, 37_7579313	21
p.C238	36_7518293, 37_7577567, 37_7577568, 37_7577569	20
p.V157	36_7519186, 37_7578460, 37_7578461	19
p.V173	36_7519138, 37_7578412, 37_7578413	19
p.R306	36_7517747, 37_7577022	18
p.R280	36_7517824, 36_7517825, 37_7577098, 37_7577099, 37_7577100	16
p.Y163	36_7519167, 36_7519168, 37_7578442, 37_7578443	16
p.K132	36_7519259, 36_7519260, 36_7519261, 37_7578534, 37_7578535, 36_7578536	15
p.S241	36_7518284, 37_7577559, 37_7577560	15
p.Y234	36_7518305, 36_7518306, 37_7577580	15
p.R337	36_7514743, 37_7574017, 37_7574018	14
p.V216	36_7518927, 36_7518928, 37_7578202, 37_7578203	14

Frequent TP53 mutations by amino acid position

FIG. 9A

10/12

Variant Classification	Overall Frequency (n=5,309 patients)	Frequency Relative to All Observed TP53 Mutations (n=1,985 total TP53 mutations)
Missense_Mutation	2.52%	6.75%
Missense_Mutation	1.79%	4.79%
Missense_Mutation	1.30%	3.48%
Missense_Mutation	0.85%	2.27%
Nonsense_Mutation	0.77%	2.07%
Missense_Mutation	0.75%	2.02%
Missense_Mutation	0.72%	1.91%
Missense_Mutation	0.64%	1.71%
Missense_Mutation	0.64%	1.71%
Nonsense_Mutation	0.62%	1.66%
Missense_Mutation	0.57%	1.51%
Missense_Mutation	0.51%	1.36%
Missense_Mutation	0.45%	1.21%
Missense_Mutation	0.43%	1.16%
Missense_Mutation	0.41%	1.11%
Missense_Mutation	0.41%	1.11%
Splice_Site	0.40%	1.06%
Missense_Mutation	0.38%	1.01%
Missense_Mutation	0.36%	0.96%
Missense_Mutation	0.36%	0.96%
Nonsense_Mutation	0.34%	0.91%
Missense_Mutation	0.30%	0.81%
Missense_Mutation	0.30%	0.81%
Missense_Mutation	0.28%	0.76%
Missense_Mutation	0.28%	0.76%
Missense_Mutation	0.28%	0.76%
Missense_Mutation	0.26%	0.71%
Missense_Mutation	0.26%	0.71%

Frequent TP53 mutations by amino acid position

FIG. 9B

11/12

Variant position	Human Genome Build and Nucleic Acid Start Positions of Variant Alleles	Patient Count
p.P177	36_7519107, 37_7578384	4
p.L255	36_7518239, 37_7577515	2
p.L252	37_7577519, 37_7577525	2
p.P128	37_7578540, 37_7578542	2
p.P191	36_7519000	2
p.T155	37_7578446, 37_7578456	2
p.A138	37_7578505	1
p.A161	37_7578444	1
p.C277	37_7577107	1
p.F113	37_7579346	1
p.G112	37_7579343	1
p.I162	37_7578443	1
p.I162	37_7578443	1
p.K139	37_7578506	1
p.L264	36_7517871	1
p.M237	36_7518288	1
p.N239	37_7577558	1
p.R110	37_7579356	1
p.R267	37_7577120	1
p.R273	37_7577102	1
p.T253	36_7518239	1
p.V218	37_7578195	1
p.V274	37_7577115	1

TP53 in-frame insertion and deletion mutations were rare

FIG. 10A

12/12

Variant Classification	Overall Frequency (n=5,309 patients)	Frequency Relative to All Observed TP53 Mutations (n=1,985 total TP53 mutations)
In_Frame_Del	0.08%	0.20%
In_Frame_Del	0.04%	0.10%
In_Frame_Del	0.04%	0.10%
In_Frame_Del	0.04%	0.10%
In_Frame_Del	0.04%	0.10%
In_Frame_Del	0.04%	0.10%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Ins	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Ins	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Del	0.02%	0.05%
In_Frame_Ins	0.02%	0.05%

TP53 in-frame insertion and deletion mutations were rare

FIG. 10B