(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 April 2012 (12.04.2012)

PCT

(10) International Publication Number
WO 2012/048113 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G01N 33/15* (2006.01)
C12N 15/12 (2006.01)

(21) International Application Number:
PCT/US2011/055108

(22) International Filing Date:
6 October 2011 (06.10.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/390,956	7 October 2010 (07.10.2010)	US
61/493,800	6 June 2011 (06.06.2011)	US

(71) Applicant *(for all designated States except US)*: **THE GENERAL HOSPITAL CORPORATION** [US/US]; 55 Fruit Street, Boston, Massachusetts 02114 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **TING, David, Tsai** [US/US]; 636 Beacon St. #506, Boston, Massachusetts 02215 (US). **HABER, Daniel, A.** [US/US]; 34 Monadnock Road, Chestnut Hill, Massachusetts 02467 (US). **MAHESWARAN, Shyamala** [US/US]; 42 Eastern Ave., Lexington, Massachusetts 02421 (US). **LIPSON, Doron** [IL/US]; 142 Middlesex Rd., Newton, Massachusetts 02467 (US).

(74) Agents: **KUGLER DEYOUNG, Janice** et al.; Fish & Richardson P.C., P.O. Box 1022, Minneapolis, Minnesota 55440-1022 (US).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, 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, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, 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, ML, MR, NE, SN, TD, TG).

Published:

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

(54) **Title**: BIOMARKERS OF CANCER

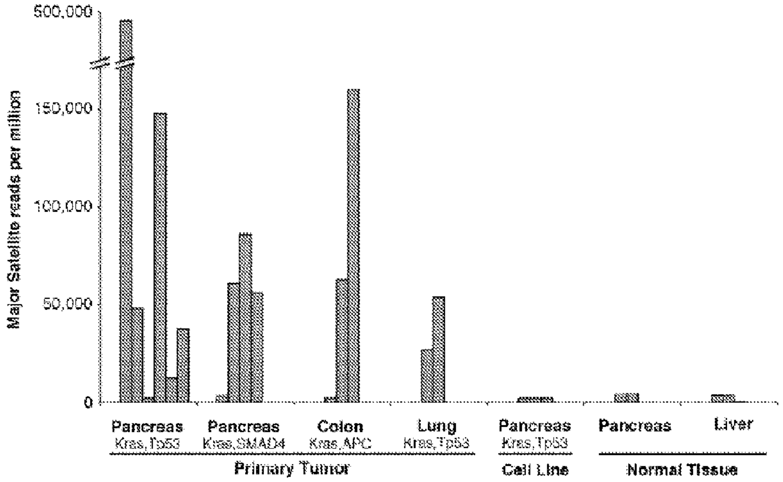


FIG. 1A

(57) **Abstract**: Methods for diagnosing cancer based on detecting the presence of increased levels of expression of satellite repeats and/or Line-1.

WO 2012/048113 A2

Biomarkers of Cancer

CLAIM OF PRIORITY

This application claims the benefit of U.S. Provisional Patent Application Serial Nos. 61/390,956, filed on October 7, 2010, and 61/493,800, filed on June 6, 2011. The entire contents of the foregoing are hereby incorporated by reference.

TECHNICAL FIELD

This invention relates to methods of diagnosing cancer, based on detecting the presence of increased levels of expression of satellite repeats and/or Line-1.

BACKGROUND

Genome-wide sequencing approaches have revealed an increasing set of transcribed non-coding sequences, including "pervasive transcription" by heterochromatic regions of the genome linked to transcriptional silencing and chromosomal integrity (J. Berretta, A. Morillon, EMBO Rep 10, 973 (Sep, 2009); A. Jacquier, Nat Rev Genet 10, 833 (Dec, 2009)). In the mouse, heterochromatin is comprised of centric (minor) and pericentric (major) satellite repeats that are required for formation of the mitotic spindle complex and faithful chromosome segregation (M. Guenatri, D. Bailly, C. Maison, G. Almouzni, J Cell Biol 166, 493 (Aug 16, 2004)), whereas human satellite repeats have been divided into multiple classes with similar functions (J. Jurka et al., Cytogenet Genome Res 110, 462 (2005)). Bidirectional transcription of satellites in yeast maintains silencing of centromeric DNA through the Dicer mediated RNA-induced transcriptional silencing (RITS) and through a recently identified Dicer-independent pathway(M. Halic, D. Moazed, Cell 140, 504 (Feb 19)), although centromeric satellite silencing mechanisms in mammals are less well defined (A. A. Aravin, G. J. Hannon, J. Brennecke, Science 318, 761 (Nov 2, 2007)). Accumulation of satellite transcripts in mouse and human cell lines results from defects in *DICER1* (C. Kanellopoulou et al., Genes Dev 19, 489 (Feb 15, 2005); T. Fukagawa et al., Nat Cell Biol 6, 784 (Aug, 2004)) and from DNA demethylation, heat shock, or the induction of apoptosis (H. Bouzinba-Segard, A. Guais, C. Francastel, Proc Natl Acad Sci U S A 103, 8709 (Jun 6, 2006); R. Valgardsdottir et al., Nucleic Acids Res 36, 423 (Feb, 2008)). Stress-induced transcription of satellites in cultured cells has also been linked to the activation of retroelements encoding RNA polymerase activity such as *LINE-1 (L1TD1)*(D. Ugarkovic, EMBO Rep 6, 1035 (Nov, 2005); D. M. Carone et al., Chromosoma 118, 113 (Feb, 2009)). Despite these *in vitro* models, the global expression of repetitive ncRNAs in primary tumors has not been analyzed, due to the bias of microarray platforms toward annotated coding sequences and the specific exclusion of repeat sequences from standard analytic programs.

SUMMARY

The present invention is based, at least in part, on the identification of massive expression of satellite repeats in tumor cells, and of increased levels of Line-1, e.g., in tumor cells including circulating tumor cells (CTCs). Described herein are methods for diagnosing cancer, e.g., solid malignancies of epithelial origin such as pancreatic, lung, breast, prostate, renal, ovarian or colon cancer, based on the presence of increased levels of expression of satellite repeats and/or Line-1.

In a first aspect, the invention provides methods, e.g., in vitro methods, for detecting the presence of cancer in a subject, including determining a level of LINE-1 in a sample from the subject to obtain a test value; and comparing the test value to a reference value, wherein a test value compared to the reference value indicates whether the subject has cancer.

In some embodiments, the reference value represents a threshold level of LINE-1, wherein the presence of a level of LINE-1 in the subject that is above the reference value indicates that the subject has cancer, and the presence of a level of LINE-1 in the subject that is below the reference value indicates that the subject is unlikely to have cancer.

In a first aspect, the invention provides methods, e.g., in vitro methods, for detecting the presence of cancer in a subject, including determining a level of satellite transcripts in a sample from the subject to obtain a test value; and comparing the test value to a reference value, wherein a test value compared to the reference value indicates whether the subject has cancer.

In some embodiments, the satellite transcripts comprise one or more of ALR, HSATII, GSATII, TAR1, and SST1. In some embodiments, the satellite transcript is

ALR and/or HSATII, and the presence of a level of ALR and/or HSATII satellite transcripts above the reference level indicates that the subject has a tumor.

In some embodiments, the satellite transcript is GSATII, TAR1 and/or SST1, and the presence of a level of GSATII, TAR1 and/or SST1 satellite transcripts below the reference level indicates that subject has a tumor.

In another aspect, the invention provides methods, e.g., in vitro methods, for evaluating the efficacy of a treatment for cancer in a subject. The methods include determining a level of LINE-1 in a first sample from the subject to obtain a first value; administering a treatment for cancer to the subject; determining a level of LINE-1 in a subsequent sample obtained from the subject at a later time, to obtain a treatment value; and comparing the first value to the treatment value. A treatment value that is below the first value indicates that the treatment is effective.

In yet another aspect, the invention provides methods, e.g., in vitro methods, for evaluating the efficacy of a treatment for cancer in a subject. The methods include determining a level of satellite transcripts in a first sample from the subject to obtain a first value; administering a treatment for cancer to the subject; determining a level of satellite transcripts in a subsequent sample obtained from the subject at a later time, to obtain a treatment value; and comparing the first value to the treatment value, wherein a treatment value that is below the first value indicates that the treatment is effective.

In some embodiments, the satellite transcripts comprise one or more of ALR, HSATII, GSATII, TAR1, and SST1.

In some embodiments, the first and second samples are known or suspected to comprise tumor cells, e.g., blood samples known or suspected of comprising circulating tumor cells (CTCs), or biopsy samples known or suspected of comprising tumor cells. In some embodiments, the sample comprises free RNA in serum or RNA within exosomes in blood.

In some embodiments, the treatment includes administration of a surgical intervention, chemotherapy, radiation therapy, or a combination thereof.

In some embodiments of the methods described herein, the subject is a mammal, e.g., a human or veterinary subject, e.g., experimental animal.

In some embodiments of the methods described herein, the cancer is a solid tumor of epithelial origin, e.g., pancreatic, lung, breast, prostate, renal, ovarian or colon cancer.

In some embodiments, the methods described herein include measuring a level of LINE-1 transcript.

In some embodiments of the methods described herein, the level of a LINE-1 transcript or satellite is determined using a branched DNA assay.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Methods and materials are described herein for use in the present invention; other, suitable methods and materials known in the art can also be used. The materials, methods, and examples are illustrative only and not intended to be limiting. All publications, patent applications, patents, sequences, database entries, and other references mentioned herein are incorporated by reference in their entirety. In case of conflict, the present specification, including definitions, will control.

Other features and advantages of the invention will be apparent from the following detailed description and figures, and from the claims.

DESCRIPTION OF DRAWINGS

FIG. 1A is a bar graph showing levels of major satellite in percent of all genomic aligned reads among different tumors, cell lines, and tissues. Genotype of primary tumors and cell lines indicated below each tumor type and cell line. (Kras = KrasG12D; Tp53, SMAD4, and APC represent genes deleted)

FIG. 1B is a graphical representation of sequence read contributions from major satellite among all primary tumors, cancer cell lines, and normal tissues.

FIG. 2A shows the results of Northern blot analysis of three KrasG12D, Tp53lox/+ pancreatic primary tumors (Tumors 1-3) and a stable cell line (CL3) derived from Tumor 3.

FIG. 2B shows the results of Northern blot analysis of CL3 before (0) and after (+) treatment with the DNA hypomethylating agent 5-azacitadine (AZA).

FIG. 2C shows the results of Northern blot analysis of total RNA from multiple adult and fetal mouse tissues. All Northern blots exposed for approximately 30 minutes.

FIG. 2D is a pair of photomicrographs showing the results of RNA in-situ hybridization (ISH) of normal pancreas (left) and primary pancreatic ductal adenocarcinoma (right), hybridized with a 1 kb major satellite repeat probe.

FIG. 2E is a set of three photomicrographs showing the results of ISH analysis of preneoplastic PanIN (P) lesion, adjacent to PDAC (T) and normal pancreas (N), showing positive staining in PanIN, with increased expression in full carcinoma. Higher magnification (40x) of PanIN (left) and PDAC (right) lesions.

FIG. 2F is a set of three photomicrographs showing marked expression of satellites in PDAC cells metastatic to liver, which itself does not express satellites (left). Large, glandular metastatic tumor deposits are readily identified by standard histological evaluation and stain for satellite (middle). Satellite ISH is sensitive enough to detect micrometastases in liver parenchyma not easily appreciated by standard histological analysis (right; arrowheads). All images at 20x magnification (scale bar = 100 μm).

FIG. 3A is a bar graph showing the Total satellite expression in human pancreatic ductal adenocarcinoma (PDAC), normal pancreas, other cancers (L – lung, K – kidney, O – ovary, P – prostate), and other normal human tissues (1 - fetal brain, 2 - brain, 3 - colon, 4 - fetal liver, 5 - liver, 6 - lung, 7 - kidney, 8 - placenta, 9 - prostate, and 10 - uterus) quantitated by DGE. Satellite expression is shown as transcripts per million aligned to human genome.

FIG. 3B is a bar graph showing a breakdown of satellite repeat classes as percent of total satellites in human PDAC (Black, n = 15) and normal human tissues (White, n = 12) sequenced. Satellites are ordered from highest absolute difference in tumors to highest in normal tissue (left to right). Error bars represents standard error of the mean. Fold differential of top three cancer (left, black bars) and normal (right, white bars) tissue satellite classes shown (Bar graph, center).

FIG. 4A shows the results of multiple linear correlation analysis of major satellite to other cellular transcripts among all mouse tumors and normal tissues as depicted by a heat map. X-axis is samples ordered by expression of major satellite and y-axis is genes ordered by linear correlation to major satellite expression. Light grey (High) and dark grey (Low) color is log2 (reads per million). Major satellite expression as percent genomic aligned reads (y-axis) rank ordered by satellite reads (x-axis) with expanded view of top genes with highest linearity ($R \geq 0.85$) with satellite levels.

FIG. 4B is a dot graph showing the Median distance of transcriptional start sites of all genes to *Line-1* elements ordered by linearity to satellite expression (Dark gray; highest linearity to the left) or by random (Light gray). Plotted by genes binned in 100s.

FIG. 4C is a dot graph showing Top genes with highest linearity ($R > 0.85$) defining satellite correlated genes or SCGs plotted by frequency against distance of transcriptional start site to *LINE-1* elements (Dark gray) compared to the expected frequency of these genes (Light gray).

FIG. 4D is a set of four photomicrographs showing the results of immunohistochemistry of mouse PDAC (KrasG12D, Tp53 lox/+) for the neuroendocrine marker chromogranin A. Tumors are depicted as a function of increasing chromogranin A staining (dark grey), with the relative level of major satellite expression noted for each tumor at the bottom of each image (percentage of all transcripts)

FIG. 5 is a bar graph indicating fold change expression of the indicated genes in CTC Device vs. control device. The subjects were newly diagnosed metastatic pancreatic adenocarcinoma patients. LINE-1 expression was seen in all patients at some point.

FIG. 6A is an image of RNA in situ hybridization (RNA-ISH) of human satellite HSATII in a human preneoplastic PanIN (P) lesion with adjacent non-cancerous stroma tissue (N) in formalin fixed paraffin embedded tissue (Top image) and fine needle aspirate biopsy of PDAC (T) and normal adjacent leukocytes (N). (Dark grey dots = HSATII). Scale bar = 100 μm

FIG. 6B is an image of RNA in situ hybridization of HSATII using Affymetrix ViewRNA of a potential human pancreatic circulating tumor cell captured on the HB-chip. HSATII (lightest areas; yellow in original), DAPI nuclear stain (medium grey areas, blue in original). Scale bar = 20 μm.

DETAILED DESCRIPTION

The present invention is based, at least in part, on the identification of a massive generation of LINE-1 protein and bidirectional ncRNAs from the major satellite repeat in mouse tumor models and from ALR and HSATII satellite repeats in human cancers. The exceptional magnitude of satellite levels in these cancers is unprecedented. This is likely to result from a general derepression of chromosomal marks affecting both satellites and LINE-1 retrotransposons, with proximity to LINE-1 activation potentially affecting the expression of selected cellular mRNAs. Together, the very high expression of satellites may affect chromosomal integrity and genetic stability, while the co-deregulated coding sequences may affect cell fates and biological behavior of cancer cells. In addition, the finding of massive expression of specific satellite subsets in human pancreatic cancer provides a novel biomarker for application in early detection of cancer. Finally, levels of *LINE-1* are increased in circulating tumor cells from subjects with newly diagnosed metastatic pancreatic adenocarcinoma. Thus the present methods are useful in the early detection of cancer, and can be used to predict clinical outcomes.

Diagnosing Cancer Using Transcript Biomarkers

The methods described herein can be used to diagnose the presence of, and monitor the efficacy of a treatment for, cancer, e.g., solid tumors of epithelial origin, e.g., pancreatic, lung, breast, prostate, renal, ovarian or colon cancer, in a subject.

As used herein, the term “hyperproliferative” refer to cells having the capacity for autonomous growth, i.e., an abnormal state or condition characterized by rapidly proliferating cell growth. Hyperproliferative disease states may be categorized as pathologic, i.e., characterizing or constituting a disease state, or may be categorized as non-pathologic, i.e., a deviation from normal but not associated with a disease state. The term is meant to include all types of cancerous growths or oncogenic processes, metastatic tissues or malignantly transformed cells, tissues, or organs, irrespective of histopathologic type or stage of invasiveness. A “tumor” is an abnormal growth of hyperproliferative cells. “Cancer” refers to pathologic disease states, e.g., characterized by malignant tumor growth.

As demonstrated herein, the presence of cancer, e.g., solid tumors of epithelial origin, e.g., as defined by the ICD-O (International Classification of Diseases – Oncology) code (revision 3), section (8010-8790), e.g., early stage cancer, is associated with the presence of a massive levels of satellite due to increase in transcription and processing of satellite repeats in pancreatic cancer cells, and of increased levels of LINE-1 expression in circulating tumor cells. Thus the methods can include the detection of expression levels of satellite repeats in a sample comprising cells known or suspected of being tumor cells, e.g., cells from solid tumors of epithelial origin, e.g., pancreatic, lung, breast, prostate, renal, ovarian or colon cancer cells. Alternatively or in addition, the methods can include the detection of increased levels of LINE-1 in a sample, e.g., a sample known or suspected of including tumor cells, e.g., circulating tumor cells (CTCs), e.g., using a microfluidic device as described herein.

Cancers of epithelial origin can include pancreatic cancer (e.g., pancreatic adenocarcinoma or intraductal papillary mucinous carcinoma (IPMN, pancreatic mass)), lung cancer (e.g., non-small cell lung cancer), prostate cancer, breast cancer, renal cancer, ovarian cancer, or colon cancer. For example, the present methods can be used to distinguish between benign IPMN, for which surveillance is the standard treatment, and malignant IPMN, which require resection, a procedure associated with significant morbidity and a small but significant possibility of death. In some embodiments, in a subject diagnosed with IPMN, the methods described herein can be used for surveillance/monitoring of the subject, e.g., the methods can be repeated at selected intervals (e.g., every 3, 6, 12, or 24 months) to determine whether a benign IPMN has become a malignant IPMN warranting surgical intervention. In addition, in some embodiments the methods can be used to distinguish bronchioloalveolar carcinomas from reactive processes (e.g., postpneumonic reactive processes) in samples from subjects suspected of having non-small cell lung cancer. In some embodiments, in a sample from a subject who is suspected of having breast cancer, the methods can be used to distinguish ductal hyperplasia from atypical ductal hyperplasia and ductal carcinoma in situ (DCIS). The two latter categories receive resection/radiation; the former does not require intervention. In some embodiments, in subjects suspected of having prostate cancer, the methods can be used to distinguish between atypical small acinar proliferation and malignant cancer. In some embodiments, in subjects suspected of having bladder cancer, the methods can be used to detect, e.g., transitional cell carcinoma (TCC), e.g., in urine specimens. In some embodiments, in subjects diagnosed with Barrett's Esophagus (Sharma, N Engl J Med. 2009, 24; 361(26):2548-56. Erratum in: N Engl J Med. 2010 Apr 15;362(15):1450), the methods can be used for distinguishing dysplasia in Barrett's esophagus from a reactive process. The clinical implications are significant, as a diagnosis of dysplasia demands a therapeutic intervention. Other embodiments include, but are not limited to, diagnosis of well differentiated hepatocellular carcinoma, ampullary and bile duct carcinoma, glioma vs. reactive gliosis, melanoma vs. dermal nevus, low grade sarcoma, and pancreatic endocrine tumors, *inter alia.*

Therefore, included herein are methods for diagnosing cancer, e.g., tumors of epithelial origin, e.g., pancreatic, lung, breast, prostate, renal, ovarian or colon cancer, in a subject. In some embodiments, the methods include obtaining a sample from a subject, and evaluating the presence and/or level of LINE-1 or satellites in the sample, and comparing the presence and/or level with one or more references, e.g., a control reference that represents a normal level of LINE-1 or satellites, e.g., a level in an unaffected subject or a normal cell from the same subject, and/or a disease reference that represents a level of LINE-1 or satellites associated with cancer, e.g., a level in a subject having pancreatic, lung, breast, prostate, renal, ovarian or colon cancer.

The present methods can also be used to determine the stage of a cancer, e.g., whether a sample includes cells that are from a precancerous lesion, an early stage tumor, or an advanced tumor. For example, the present methods can be used to determine whether a subject has a precancerous pancreatic, breast, or prostate lesion. Where the markers used are LINE-1, or satellite transcript ALR and/or HSATII, increasing levels are correlated with advancing stage. For satellite transcripts GSATII, TAR1 and/or SST1, decreasing levels are correlated with increasing stage. Additionally, levels of LINE-1 and satellite ALR and/or HSATII may be prognostic and predictive to clinical outcomes.

Samples

In some embodiments of the present methods, the sample is or includes blood, serum, and/or plasma, or a portion or subfraction thereof, e.g., free RNA in serum or RNA within exosomes in blood. In some embodiments, the sample comprises (or is suspected of comprising) CTCs. In some embodiments, the sample is or includes urine or a portion or subfraction thereof. In some embodiments, the sample includes known or suspected tumor cells, e.g., is a biopsy sample, e.g., a fine needle aspirate (FNA), endoscopic biopsy, or core needle biopsy; in some embodiments the sample comprises cells from the pancreatic, lung, breast, prostate, renal, ovarian or colon of the subject. In some embodiments, the sample comprises lung cells obtained from a sputum sample or from the lung of the subject by brushing, washing, bronchoscopic biopsy, transbronchial biopsy, or FNA, e.g., bronchoscopic, fluoroscopic, or CT-guided FNA (such methods can also be used to obtain samples from other tissues as well). In some embodiments, the sample is frozen, fixed and/or permeabilized, e.g., is an formalin-fixed paraffin-embedded (FFPE) sample.

Satellite Expression Levels

In some embodiments, the level of satellite transcripts is detected, e.g., in a sample known or suspected to include tumor cells. In some embodiments, the level of satellite transcripts in a known or suspected tumor cell, e.g., a test cell, is compared to a reference level.

In some embodiments, the methods include detecting levels of alpha (ALR) satellite transcripts (D. Lipson et al., Nat Biotechnol 27, 652 (Jul, 2009)) or HSATII satellite transcripts (J. Jurka et al., Cytogenet Genome Res 110, 462 (2005)); in some embodiments, those levels are compared to a reference. In some embodiments, the reference level is a level of ALR and/or HSATII satellite transcripts in a normal (non-cancerous) cell, e.g., a normal cell from the same subject, or a reference level determined from a cohort of normal cells; the presence of levels of ALR and/or HSATII in the test cell above those in the normal cell indicate that the test cell is a tumor cell (e.g., the subject from whom the test cell came has or can be diagnosed with cancer). In some embodiments, the reference level of ALR and/or HSATII transcripts is a threshold level, and the presence of a level of ALR and/or HSATII satellite transcripts above the threshold level indicates that the cell is a tumor cell (e.g., the subject from whom the test cell came has or can be diagnosed with cancer).

In some embodiments, the methods include detecting levels of GSATII, TAR1 and/or SST1 transcripts; in some embodiments, those levels are compared to a reference. In some embodiments, the reference level is a level of GSATII, TAR1 and/or SST1 satellite transcripts in a normal (non-cancerous) cell, e.g., a normal cell from the same subject, or a reference level determined from a cohort of normal cells; the presence of levels of GSATII, TAR1 and/or SST1 in the test cell below those in the normal cell indicate that the test cell is a tumor cell (e.g., the subject from whom the test cell came has or can be diagnosed with cancer). In some embodiments, the reference level of

GSATII, TAR1 and/or SST1 transcripts is a threshold level, and the presence of a level of GSATII, TAR1 and/or SST1 satellite transcripts below the threshold level indicates that the cell is a tumor cell (e.g., the subject from whom the test cell came has or can be diagnosed with cancer).

In some embodiments, the levels of the satellite transcripts are normalized to a relatively non-variant transcript such as GAPDH, actin, or tubulin, e.g., the level of expression of the satellite is compared to the non-variant transcript. For example, a ratio of expression levels can be calculated, and the ratio can be compared to the ratio in a normal (non-cancerous) cell. For example, in some embodiments the presence of a ratio of ALR:GAPDH of over 10:1, e.g., over 50:1, over 100:1, or over 150:1, indicates that the test cell is a cancer cell; in some embodiments the presence of a ratio of ALR:GAPDH of about 3:1 or 5:1 indicates that the test cell is a normal cell. In some embodiments, the presence of a ratio of HSATII satellites:GAPDH transcripts of over 10:1, e.g., 20:1, 30:1, 40:1, or 45:1, indicates that the test cell is a cancer cell. In some embodiments, the presence of significant (e.g., more than about 100 transcripts per million aligned) levels of HSATII indicates that the test cell is a cancer cell. In some embodiments, the absence or presence of very low levels (e.g., less than about 20 transcripts per million aligned) of HSATII indicates that the test cell is a normal cell.

Below are exemplary reference sequences that can be used for ALR (including its variants) and HSATII transcripts:

```
>ALR    SAT     Homo sapiens
aattctcagtaacttccttgtgttgtgtgtattcaactcacagagttgaacgatcctttacacagagcag
acttgaaacactctttttgtggaatttgcaagtggagatttcagccgctttgaggtcaatggtagaatag
gaaatatcttcctatagaaactagacagaat (SEQ ID NO:1)

>ALR1   SAT     Homo sapiens
tcattctcagaaactrctttgtgatgtgtgcrttcaactcacagagtttaacctttcttttgatagagca
gtttggaaacactctgtttgtaaagtctgcaagtggatatttggacctctttgaggccttcgttggaaac
gggatttcttcatataatgctagacagaaga (SEQ ID NO:2)

>ALR2   SAT     Homo sapiens
agctttctgagaaactgctttgtgatgtgtgcattcatctcacagagttaaacctttcttttgattcagc
agtttggaaacactgtttttgtagaatctgtgaagggatatttgggagctcattgaggcctatggtgaaa
aagaaaatatcttcagataaaaactagaaggaagctatc (SEQ ID NO:3)

>ALRa   SAT     Primates
ctatctgagaaactgctttgtgatgtgtgcattcatctcacagagttaaacctttcttttgattcagcag
tttggaaacactgtttttgtagaatctgcgaagggacatttgggagctcattgaggcctatggtgaaaaa
gcgaatatccccagataaaaactagaaagaag (SEQ ID NO:4)
```

```
>ALRa_  SAT     Primates
ttgtagaatctgcgaagggacatttgggagctcattgaggcctatggtgaaaaagcgaatatccccagat
aaaaactagaaagaagctatctgagaaactgctttgtgatgtgtgcattcatctcacagagttaaacctt
tcttttgattcagcagtttggaaacactgttt (SEQ ID NO:5)

>ALRb   SAT     Primates
ttgtggaatttgcaagtggagatttcaagcgctttgaggccaawnktagaaaaggaaatatcttcgtata
aaaactagacagaataattctcagtaacttctttgtgttgtgtgtattcaactcacagagttgaaccttc
ctttagacagagcagatttgaaacactcttt (SEQ ID NO:6)

>ALR_   SAT     Primates
ttgtagaatctgcaagtggatatttggasckctttgaggmcttcgktggaaacgggaatatcttcacata
aaaactagacagaagcattctcagaaacttctttgtgatgtttgcattcaactcacagagttgaacmttc
cttttgatagagcagttttgaaacactcttt (SEQ ID NO:7)

>HSATII SAT     Primates
ccattcgattccattcgatgattccattcgattccattcgatgatgattccattcgattccattcgatga
ttccattcgattccattcgatgatgattccattcgattccattcgatgattccattcgattccattcgat
gatgattccattcgattccattcgatgatt (SEQ ID NO:8)
```

Line-1 Levels

Long interspersed nucleotide element (LINE) non-LTR retrotransposons (Singer, Cell 28 (3): 433–4 (1982)) are a group of genetic elements that are found in large numbers in eukaryotic genomes, and generate insertion mutations, contribute to genomic instability and innovation, and can alter gene expression.

The canonical, full-length LINE-1 element is about 6 kilobases (kb) in length and includes a 5′ untranslated region (UTR) with an internal RNA polymerase II promoter (Swergold, Mol Cell Biol. 10(12):6718-29 (1990)), two open reading frames (designated ORF1 and ORF2) and a 3′ UTR containing a polyadenylation signal ending with an oligo dA-rich tail of variable length (Babushok and Kazazian, Hum Mutat. 28(6):527-39 (2007)). Although there are over 500,000 L1 elements inserted in the human genome, only about 80-100 copies are retrotransposition-competent (Brouha et al., Proc Natl Acad Sci U S A. 100(9):5280-5. (2003)). For additional details, see Cordaux and Batzer, Nat Rev Genet. 10(10): 691–703 (2009)).

Exemplary LINE-1 sequences include GenBank Ref. No. NM_001164835.1 (nucleic acid) and NP_001158307.1 (protein) for variant (1); and GenBank Ref. No. NM_019079.4 (nucleic acid) and NP_061952.3 (protein) for variant 2, which is the shorter transcript. Variant 2 differs in the 5' UTR compared to variant 1, but both variants 1 and 2 encode the same protein. See also Gene ID: 54596.

In some embodiments, the methods for diagnosing cancer described herein include determining a level of LINE-1 mRNA in a cell, e.g., in CTCs present in blood of a subject to obtain a LINE-1 value, and comparing the value to an appropriate reference value, e.g., a value that represents a threshold level, above which the subject can be diagnosed with cancer. The reference can also be a range of values, e.g., that indicate severity or stage of the cancer in the subject. A suitable reference value can be determined by methods known in the art.

In some embodiments, the reference level is a level of LINE-1 transcripts in a normal (non-cancerous) cell, e.g., a normal cell from the same subject, or a reference level determined from a cohort of normal cells; the presence of levels of LINE-1 in the test cell above those in the normal cell indicate that the test cell is a tumor cell (e.g., the subject from whom the test cell came has or can be diagnosed with cancer). In some embodiments, the reference level of LINE-1 transcripts is a threshold level, and the presence of a level of LINE-1 transcripts above the threshold level indicates that the cell is a tumor cell (e.g., the subject from whom the test cell came has or can be diagnosed with cancer).

Methods of Detection

Any methods known in the art can be used to detect and/or quantify levels of a biomarker as described herein. For example, the level of a satellite transcript or LINE-1 mRNA (transcript) can be evaluated using methods known in the art, e.g., Northern blot, RNA in situ hybridization (RNA-ISH), RNA expression assays, e.g., microarray analysis, RT-PCR, deep sequencing, cloning, Northern blot, and quantitative real time polymerase chain reaction (qRT-PCR). Analytical techniques to determine RNA expression are known. See, e.g., Sambrook et al., Molecular Cloning: A Laboratory Manual, 3rd Ed., Cold Spring Harbor Press, Cold Spring Harbor, NY (2001).

In some embodiments, the level of the LINE-1 protein is detected. The presence and/or level of a protein can be evaluated using methods known in the art, e.g., using quantitative immunoassay methods such as enzyme linked immunosorbent assays (ELISAs), immunoprecipitations, immunofluorescence, immunohistochemistry, enzyme immunoassay (EIA), radioimmunoassay (RIA), and Western blot analysis.

In some embodiments, the methods include contacting an agent that selectively binds to a biomarker, e.g., to a satellite transcript or LINE-1 mRNA or protein (such as an oligonucleotide probe, an antibody or antigen-binding portion thereof) with a sample, to evaluate the level of the biomarker in the sample. In some embodiments, the agent bears a detectable label. The term "labeled," with regard to an agent encompasses direct labeling of the agent by coupling (i.e., physically linking) a detectable substance to the agent, as well as indirect labeling of the agent by reactivity with a detectable substance. Examples of detectable substances are known in the art and include chemiluminescent, fluorescent, radioactive, or colorimetric labels. For example, detectable substances can include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, beta-galactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride, quantum dots, or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I, ^{131}I, ^{35}S or ^{3}H. In general, where a protein is to be detected, antibodies can be used. Antibodies can be polyclonal, or more preferably, monoclonal. An intact antibody, or an antigen-binding fragment thereof (e.g., Fab or F(ab')$_2$) can be used.

In some embodiments, high throughput methods, e.g., protein or gene chips as are known in the art (see, e.g., Ch. 12, "Genomics," in Griffiths et al., Eds. Modern genetic Analysis, 1999,W. H. Freeman and Company; Ekins and Chu, Trends in Biotechnology, 1999;17:217-218; MacBeath and Schreiber, Science 2000, 289(5485):1760-1763; Simpson, Proteins and Proteomics: A Laboratory Manual, Cold Spring Harbor Laboratory Press; 2002; Hardiman, Microarrays Methods and Applications: Nuts & Bolts, DNA Press, 2003), can be used to detect the presence and/or level of satellites or LINE-1.

In some embodiments, the methods include using a modified RNA in situ hybridization technique using a branched-chain DNA assay to directly detect and evaluate the level of biomarker mRNA in the sample (see, e.g., Luo et al., US Pat. No.

7,803,541B2, 2010; Canales et al., Nature Biotechnology 24(9):1115-1122 (2006); Nguyen et al., Single Molecule in situ Detection and Direct Quantiication of miRNA in Cells and FFPE Tissues, poster available at panomics.com/index.php?id=product_87). A kit for performing this assay is commercially-available from Affymetrix (ViewRNA).

Detection of LINE-1 and Satellite Transcripts in CTCs

In some embodiments, microfluidic (e.g., "lab-on-a-chip") devices can be used in the present methods. Such devices have been successfully used for microfluidic flow cytometry, continuous size-based separation, and chromatographic separation. In general, methods in which expression of satellites or LINE-1 is detected in circulating tumor cells (CTCs) can be used for the early detection of cancer, e.g., early detection of tumors of epithelial origin, e.g., pancreatic, lung, breast, prostate, renal, ovarian or colon cancer.

The devices can be used for separating CTCs from a mixture of cells, or preparing an enriched population of CTCs. In particular, such devices can be used for the isolation of CTCs from complex mixtures such as whole blood.

A variety of approaches can be used to separate CTCs from a heterogeneous sample. For example, a device can include an array of multiple posts arranged in a hexagonal packing pattern in a microfluidic channel upstream of a block barrier. The posts and the block barrier can be functionalized with different binding moieties. For example, the posts can be functionalized with anti-EPCAM antibody to capture circulating tumor cells (CTCs); see, e.g., Nagrath et al., Nature 450:1235-1239 (2007), optionally with downstream block barriers functionalized with to capture LINE-1 nucleic acids or proteins, or satellites. See, e.g., (S. Maheswaran et al., N Engl J Med. 359, 366 (Jul 24, 2008); S. Nagrath et al., Nature. 450, 1235 (Dec 20, 2007); S. L. Stott et al., Sci Transl Med 2, 25ra23 (Mar 31)) and the applications and references listed herein.

Processes for enriching specific particles from a sample are generally based on sequential processing steps, each of which reduces the number of undesired cells/particles in the mixture, but one processing step may suffice in some embodiments. Devices for carrying out various processing steps can be separate or integrated into one microfluidic system. The devices include devices for cell/particle binding, devices for cell lysis, devices for arraying cells, and devices for particle separation, e.g., based on size, shape, and/or deformability or other criteria. In certain embodiments, processing steps are used to reduce the number of cells prior to introducing them into the device or system. In some embodiments, the devices retain at least 75%, e.g., 80%, 90%, 95%, 98%, or 99% of the desired cells compared to the initial sample mixture, while enriching the population of desired cells by a factor of at least 100, e.g., by 1000, 10,000, 100,000, or even 1,000,000 relative to one or more non-desired cell types.

Some devices for the separation of particles rely on size-based separation with or without simultaneous cell binding. Some size-based separation devices include one or more arrays of obstacles that cause lateral displacement of CTCs and other components of fluids, thereby offering mechanisms of enriching or otherwise processing such components. The array(s) of obstacles for separating particles according to size typically define a network of gaps, wherein a fluid passing through a gap is divided unequally into subsequent gaps. Both sieve and array sized-based separation devices can incorporate selectively permeable obstacles as described above with respect to cell-binding devices.

Devices including an array of obstacles that form a network of gaps can include, for example, a staggered two-dimensional array of obstacles, e.g., such that each successive row is offset by less than half of the period of the previous row. The obstacles can also be arranged in different patterns. Examples of possible obstacle shapes and patterns are discussed in more detail in WO 2004/029221.

In some embodiments, the device can provide separation and/or enrichment of CTCs using array-based size separation methods, e.g., as described in U.S. Pat. Pub. No. 2007/0026413. In general, the devices include one or more arrays of selectively permeable obstacles that cause lateral displacement of large particles such as CTCs and other components suspended in fluid samples, thereby offering mechanisms of enriching or otherwise processing such components, while also offering the possibility of selectively binding other, smaller particles that can penetrate into the voids in the dense matrices of nanotubes that make up the obstacles. Devices that employ such selectively permeable obstacles for size, shape, or deformability based enrichment of particles, including filters, sieves, and enrichment or separation devices, are described in International Publication Nos. 2004/029221 and 2004/113877, Huang et al. Science 304:987-990 (2004), U.S. Publication No. 2004/0144651, U.S. Pat. Nos. 5,837,115 and

6,692,952, and U.S. Application Nos. 60/703,833, 60/704,067, and 11/227,904; devices useful for affinity capture, e.g., those described in International Publication No. 2004/029221 and U.S. Application Ser. No. 11/071,679; devices useful for preferential lysis of cells in a sample, e.g., those described in International Publication No. 2004/029221, U.S. Pat. No. 5,641,628, and U.S. Application No. 60/668,415; devices useful for arraying cells, e.g., those described in International Publication No. 2004/029221, U.S. Pat. No. 6,692,952, and U.S. application Ser. Nos. 10/778,831 and 11/146,581; and devices useful for fluid delivery, e.g., those described in U.S. Application Nos. 11/071,270 and 11/227,469. Two or more devices can be combined in series, e.g., as described in International Publication No. WO 2004/029221. All of the foregoing are incorporated by reference herein.

In some embodiments, a device can contain obstacles that include binding moieties, e.g., monoclonal anti-EpCAM antibodies or fragments thereof, that selectively bind to particular cell types, e.g., cells of epithelial origin, e.g., tumor cells. All of the obstacles of the device can include these binding moieties; alternatively, only a subset of the obstacles include them. Devices can also include additional modules, e.g., a cell counting module or a detection module, which are in fluid communication with the microfluidic channel device. For example, the detection module can be configured to visualize an output sample of the device.

In one example, a detection module can be in fluid communication with a separation or enrichment device. The detection module can operate using any method of detection disclosed herein, or other methods known in the art. For example, the detection module includes a microscope, a cell counter, a magnet, a biocavity laser (see, e.g., Gourley et al., J. Phys. D: Appl. Phys., 36: R228-R239 (2003)), a mass spectrometer, a PCR device, an RT-PCR device, a microarray, RNA in situ hybridization system, or a hyperspectral imaging system (see, e.g., Vo-Dinh et al., IEEE Eng. Med. Biol. Mag., 23:40-49 (2004)). In some embodiments, a computer terminal can be connected to the detection module. For instance, the detection module can detect a label that selectively binds to cells, proteins, or nucleic acids of interest, e.g., LINE-1 DNA, mRNA, or proteins, or satellite DNA or mRNA.

In some embodiments, the microfluidic system includes (i) a device for separation or enrichment of CTCs; (ii) a device for lysis of the enriched CTCs; and (iii) a device for detection of LINE-1 DNA, mRNA, or proteins, or satellite DNA or mRNA.

In some embodiments, a population of CTCs prepared using a microfluidic device as described herein is used for analysis of expression of LINE-1 and/or satellites using known molecular biological techniques, e.g., as described above and in Sambrook, Molecular Cloning: A Laboratory Manual, Third Edition (Cold Spring Harbor Laboratory Press; 3rd edition (January 15, 2001)); and Short Protocols in Molecular Biology, Ausubel et al., eds. (Current Protocols; 52 edition (November 5, 2002)).

In general, devices for detection and/or quantification of expression of satellites or LINE-1 in an enriched population of CTCs are described herein and can be used for the early detection of cancer, e.g., tumors of epithelial origin, e.g., early detection of pancreatic, lung, breast, prostate, renal, ovarian or colon cancer.

Methods of Monitoring Disease Progress or Treatment Efficacy

In some embodiments, once it has been determined that a person has cancer, or has an increased risk of developing cancer, then a treatment, e.g., as known in the art, can be administered. The efficacy of the treatment can be monitored using the methods described herein; an additional sample can be evaluated after (or during) treatment, e.g., after one or more doses of the treatment are administered, and a decrease in the level of LINE-1, and/or ALR and/or HSATII satellites expression, or in the number of LINE-1-, and/or ALR and/or HSATII satellite-expressing cells in a sample, would indicate that the treatment was effective, while no change or an increase in the level of LINE-1, and/or ALR and/or HSATII satellite expression or LINE-1-, and/or ALR and/or HSATII satellite-expressing cells would indicate that the treatment was not effective (the converse would of course be true for levels of GSATII, TAR1 and/or SST1 satellites). The methods can be repeated multiple times during the course of treatment, and/or after the treatment has been concluded, e.g., to monitor potential recurrence of disease.

In some embodiments, e.g., for subjects who have been diagnosed with a benign condition that could lead to cancer, subjects who have been successfully treated for a cancer, or subjects who have an increased risk of cancer, e.g., due to a genetic predisposition or environmental exposure to cancer-causing agents, the methods can be repeated at selected intervals, e.g., at 3, 6, 12, or 24 month intervals, to monitor the disease in the subject for early detection of progression to malignancy or development of cancer in the subject.

EXAMPLES

The invention is further described in the following examples, which do not limit the scope of the invention described in the claims.

Example 1. Major Satellite levels are massively elevated in tumor tissues compared to cell lines and normal tissue

The next generation digital gene expression (DGE) application from Helicos BioSciences (D. Lipson et al., Nat Biotechnol 27, 652 (Jul, 2009)) was utilized to compare expression of tumor markers in primary cancers and their derived metastatic precursors. We first determined DGE profiles of primary mouse pancreatic ductal adenocarcinoma (PDAC) generated through tissue-targeted expression of activated *Kras* and loss of *Tp53* (*Kras*^{G12D}, *Tp53*$^{lox/+}$) (N. Bardeesy et al., Proc Natl Acad Sci U S A 103, 5947 (Apr 11, 2006)). These tumors are histopathological and genetic mimics of human PDAC, which exhibits virtually universal mutant *KRAS* (>90% of cases) and loss of *TP53* (50-60%).

Mice with pancreatic cancer of different genotypes were bred as previously described in the Bardeesy laboratory (Bardeesy et al., Proc Natl Acad Sci U S A 103, 5947 (2006)). Normal wild type mice were purchased from Jackson laboratories. Animals were euthanized as per animal protocol guidelines. Pancreatic tumors and normal tissue were extracted sterilely and then flash frozen with liquid nitrogen. Tissues were stored at -80° C. Cell lines were generated fresh for animals AH367 and AH368 as previously described (Aguirre et al., Genes Dev 17, 3112 (2003)) and established cell lines were cultured in RPMI-1640 + 10% FBS + 1% Pen/Strep (Gibco/Invitrogen). Additional mouse tumors from colon and lung were generously provided by Kevin Haigis (Massachusetts General Hospital) and Kwok-Kin Wong (Dana Farber Cancer Institute).

Fresh frozen tissue was pulverized with a sterile pestle in a microfuge tube on dry ice. Cell lines were cultured and fresh frozen in liquid nitrogen prior to nucleic acid extraction. RNA and DNA from cell lines and fresh frozen tumor and normal tissues were all processed in the same manner. RNA was extracted using the TRIzol® Reagent (Invitrogen) per manufacturer's specifications. DNA from tissue and cell lines was extracted using the QIAamp Mini Kit (QIAGEN) per manufacturer's protocol.

Purified RNA was subjected to Digital Gene Expression (DGE) sample prepping and analysis on the HeliScope™ Single Molecule Sequencer from Helicos BioSciences. This method has been previously described (Lipson et al., Nat Biotechnol 27, 652 (2009)). Briefly, Single stranded cDNA was reverse transcribed from RNA with a dTU25V primer and the Superscript III cDNA synthesis kit (Invitrogen). RNA was digested and single stranded cDNA was purified using a solid phase reversible immobilization (SPRI) technique with Agencourt® AMPure® magnetic beads. Single stranded cDNA was denatured and then a poly-A tail was added to the 3' end using terminal transferase (New England Biolabs).

Purified DNA was subjected to the DNA sequencing sample prepping protocol from Helicos that has been previously described (Pushkarev, N. F. Neff, S. R. Quake, Nat Biotech 27, 847 (2009)). Briefly, genomic DNA was sheared with a Covaris S2 acoustic sonicator producing fragments averaging 200 bps and ranging from 100-500 bps. Sheared DNA was then cleaned with SPRI. DNA was then denatured and a poly-A tail was added to the 3' end using terminal transferase.

Tailed cDNA or DNA were then hybridized to the sequencing flow cell followed by "Fill and Lock" and single molecule sequencing. Gene expression sequence reads were then aligned to the known human or mouse transcriptome libraries using the DGE program. Genomic DNA sequence reads were aligned to the mouse genome and counted to determine copy number of the major mouse satellite (CNV).

The first mouse pancreatic tumor analyzed, AH284, was remarkable in that DGE sequences displayed a 48-52% discrepancy with the annotated mouse transcriptome, compared with a 3-4% difference for normal liver transcripts from the same mouse. Nearly all the discrepant sequences mapped to the pericentric (major) mouse satellite repeat. The satellite transcript accounts for ~49% (495,421 tpm) of all cellular transcripts in the tumor, compared with 0.02-0.4% (196-4,115 tpm) in normal pancreas or liver (Table 1).

Table 1

Total genomic aligned reads with breakdown of major satellite and transcriptome reads. Percentage of total genomic aligned reads in parentheses

	Total Reads	**Major Satellite Reads**	**Transcriptome**
Pancreatic Tumor	18,063,363	8,460,135 (47%)	1,726,768 (10%)
Normal Liver	2,270,669	8,973 (0.4%)	1,718,489 (75%)
Normal Pancreas	492,301	2,026 (0.4%)	63,160 (13%)

Satellite sequence reads were found in both sense and anti-sense directions and are absent from poly-A purified RNA. Tumor AH284 therefore contained massive amounts of a non-polyadenylated dsRNA element, quantitatively determined as >100-fold increased over that present in normal tissue from the same animal. By way of comparison, the levels of satellite transcripts in tumor tissues were about 8,000-fold higher than the abundant mRNA *Gapdh*. A second independent pancreatic tumor nodule from the same mouse showed a lower, albeit still greatly elevated, level of satellite transcript (4.5% of total cellular transcripts).

Analysis of 4 additional pancreatic tumors from (*Kras*^{G12D}, *Tp53*$^{lox/+}$) mice and 4 mice with an alternative pancreatic tumorigenic genotype (*Kras* ^{G12D}, *SMAD4*$^{lox/lox}$) revealed increased satellite expression in 6/8 additional tumors (range 1-15% of all cellular transcripts). In 2/3 mouse colon cancer tumors (*Kras* ^{G12D}, *APC* $^{lox/+}$) and 2/2 lung cancers (*Kras* ^{G12D}, *Tp53* $^{lox/lox}$), satellite expression level ranged from 2-16% of all cellular transcripts. In total, 12/15 (80%) independent mouse tumors had greatly increased levels of satellite expression, compared to normal mouse tissues (Fig. 1A, Table 2).

Table 2

Total genomic reads and percentage of reads aligning to transcriptome and major satellite among multiple mouse tumors, cell lines, and normal tissues.

Mouse ID	Tissue Type	Genotype	Total Genomic Reads	% Transcriptome Reads	% Major Satellite Reads
AH284 Rep 1	Pancreatic Cancer	*Kras*^{G12D}, *Tp53*$^{lox/+}$	18,063,363	9.56%	46.84%
AH284 Rep 2	Pancreatic Cancer	*Kras*^{G12D}, *Tp53*$^{lox/+}$	16,948,693	10.15%	49.54%
AH284 – 2*	Pancreatic Cancer	*Kras*^{G12D}, *Tp53*$^{lox/+}$	1,613,592	48.67%	4.78%
AH287	Pancreatic Cancer	*Kras*^{G12D}, *Tp53*$^{lox/+}$	2,227,850	54.70%	0.07%
AH288	Pancreatic Cancer	*Kras*^{G12D}, *Tp53*$^{lox/+}$	6,780,821	26.57%	14.79%
AH291	Pancreatic Cancer	*Kras*^{G12D}, *Tp53*$^{lox/+}$	1,388,906	43.12%	1.22%
AH294	Pancreatic Cancer	*Kras*^{G12D}, *Tp53*$^{lox/+}$	969,896	37.20%	3.73%
AH323	Pancreatic Cancer	*Kras*^{G12D}, *SMAD4*$^{lox/lox}$	1,887,663	72.73%	0.29%
AH346	Pancreatic Cancer	*Kras*^{G12D}, *SMAD4*$^{lox/lox}$	1,291,648	32.92%	6.07%
AH347	Pancreatic Cancer	*Kras*^{G12D}, *SMAD4*$^{lox/lox}$	1,634,314	38.94%	8.59%
AH348	Pancreatic Cancer	*Kras*^{G12D}, *SMAD4*$^{lox/lox}$	2,030,197	45.84%	5.61%
Colon 1	Colon Cancer - 1	*Kras*^{G12D}, *APC*$^{lox/lox}$	2,954,930	77.49%	0.07%
Colon 1	Colon Cancer - 2	*Kras*^{G12D}, *APC*$^{lox/lox}$	985,510	53.13%	6.27%
Colon 1	Colon Cancer - 3	*Kras*^{G12D}, *APC*$^{lox/lox}$	1,017,319	30.71%	16.02%
KN2128	Lung Cancer	*Kras*^{G12D}, *Tp53*$^{lox/lox}$	2,233,183	60.78%	2.66%
KN2199	Lung Cancer	*Kras*^{G12D}, *Tp53*$^{lox/lox}$	1,653,948	43.21%	5.37%
AH323	PDAC Cell Line	*Kras*^{G12D}, *SMAD4*$^{lox/lox}$	1,958,108	83.13%	0.02%
AH324	PDAC Cell Line	*Kras*^{G12D}, *Tp53*$^{lox/+}$	3,301,108	86.32%	0.04%
NB490	PDAC Cell Line	*Kras*^{G12D}, *Tp53*$^{lox/lox}$	15,378,802	76.85%	0.03%
AH284 Rep 1	Matched Normal Liver	*Kras*^{G12D}, *Tp53*$^{lox/+}$	2,270,669	75.68%	0.40%
AH284 Rep 2	Matched Normal Liver	*Kras*^{G12D}, *Tp53*$^{lox/+}$	1,627,749	56.59%	0.34%
AH284 – 2*	Matched Normal Liver	*Kras*^{G12D}, *Tp53*$^{lox/+}$	644,316	41.10%	0.31%
Colon 1	Matched Normal Liver	*Kras*^{G12D}, *APC*$^{lox/lox}$	1,536,346	86.53%	0.02%
Normal 1	Normal Pancreas	WT	247,582	14.49%	0.41%
Normal 2	Normal Pancreas	WT	244,719	11.15%	0.41%

* AH284-2 was RNA extraction from a different part of the pancreatic tumor and liver

Of note, the composite distribution of all RNA reads among coding, ribosomal and other non-coding transcripts showed significant variation between primary tumors and normal tissues (Fig. 1A), suggesting that the global cellular transcriptional machinery is affected by the massive expression of satellite transcripts in primary tumors.

Immortalized cell lines established from 3 primary pancreatic tumors displayed minimal expression of satellite repeats, suggesting either negative selection pressure during in vitro proliferation or reestablishment of stable satellite silencing mechanisms under in vitro culture conditions (Fig. 1A). Of note in primary tumors overexpressing satellites, the composite distribution of all RNA reads among coding, ribosomal and other non-coding transcripts shows significant variation with that of normal tissues (Fig. 1B), suggesting that the cellular transcriptional machinery is affected by the massive expression of satellite transcripts in these tumors.

Example 2. Major satellite transcripts are of various sizes depending on tissue type and expression levels are linked to genomic methylation and amplification

Northern blot analysis of mouse primary pancreatic tumors was carried out as follows. Northern Blot was performed using the NorthernMax-Gly Kit (Ambion). Total RNA (10 ug) was mixed with equal volume of Glyoxal Load Dye (Ambion) and incubated at 50°C for 30 min. After electrophoresis in a 1% agarose gel, RNA was transferred onto BrightStar-Plus membranes (Ambion) and crosslinked with ultraviolet light. The membrane was prehybridized in ULTRAhyb buffer (Ambion) at 68°C for 30 min. The mouse RNA probe (1100 bp) was prepared using the MAXIscript Kit (Ambion) and was nonisotopically labeled using the BrightStar Psoralen-Biotin Kit (Ambion) according to the manufacturer's instructions. Using 0.1 nM probe, the membrane was hybridized in ULTRAhyb buffer (Ambion) at 68°C for 2 hours. The membrane was washed with a Low Stringency wash at room temperature for 10 min, followed by two High Stringency washes at 68°C for 15 min. For nonisotopic chemiluminescent detection, the BrightStar BioDetect Kit was used according to the manufacturer's instructions.

The results demonstrated that the major satellite-derived transcripts range from 100 bp to 2.5 kb (Fig. 2A), consistent with the predicted cleavage of a large primary transcript comprised of multiple tandem repeats by *Dicer1*(C. Kanellopoulou et al., Genes Dev 19, 489 (Feb 15, 2005); T. Fukagawa et al., Nat Cell Biol 6, 784 (Aug, 2004); H. Bouzinba-Segard, A. Guais, C. Francastel, Proc Natl Acad Sci U S A 103, 8709 (Jun 6, 2006)), whose expression is 2.6-fold higher ($p = 0.0006$, t-test) in mouse pancreatic tumors with satellite expression above the median. An established pancreatic cancer cell line derived from a primary tumor with high satellite expression has very little satellite expression confirming our sequencing results (T3 and CL3; Fig. 2A). Treatment of CL3 with 5-azacytidine leads to massive reexpression of satellite transcripts supporting DNA methylation as a mechanism for stable satellite silencing in vitro (Fig. 2B). Most normal adult mouse tissues, with the exception of lung, show minimal expression of satellite repeats (Fig. 2B). However, expression of the uncleaved 5 kb satellite transcript is evident in embryonic tissues (Fig. 2C). Thus, the aberrant expression of satellite repeats in primary pancreatic tumors does not simply recapitulate an embryonic cell fate, but also reflects altered processing of the primary 5 kb satellite transcript. The single molecule sequencing platform was exceptionally sensitive for quantitation of small repetitive ncRNA fragments, each of which is scored as a unique read. High level expression of the mouse major satellite was evident in all cells within the primary tumor (Fig. 2D), as shown by RNA in situ hybridization (ISH). Remarkably, expression was already elevated in early preneoplastic lesions, pancreatic intraepithelial neoplasia (PanIN), and it increased further upon transition to full pancreatic adenocarcinoma (Fig. 2E). Clearly defined metastatic lesions to the liver ware strongly positive by RNA ISH, as were individual PDAC cells within the liver parenchyma that otherwise would not have been detected by histopathological analysis (Fig. 2F). Low level diffuse expression was evident in liver and lung, as shown by whole mount embryo analysis, but no normal adult or embryonic tissues demonstrated satellite expression comparable to that evident in tumor cells.

To determine whether genomic amplification of satellite repeats also contributes toward the exceptional abundance of these transcripts in mouse pancreatic tumors, the index AH284 tumor was analyzed using next generation DNA digital copy number variation (CNV) analysis as described above for genomic DNA sequencing.

The results, shown in Table 3, indicated that satellite DNA comprised 18.8% of all genome-aligned reads in this tumor, compared with 2.3% of genomic sequences in matched normal liver. The major satellite repeat has previously been estimated at approximately 3% of the normal mouse genome (J. H. Martens et al., EMBO J 24, 800 (Feb 23, 2005)). Thus, in this tumor with >100-fold increased expression of satellite repeats, approximately 8-fold gene amplification of the repeats may contribute to their abnormal expression.

Table 3

CNV analysis of index pancreatic tumor and normal liver from mouse AH284. Major satellite reads as a percentage of all genomic aligned reads (last column)

	Major Satellite Reads	Total Genomic Reads
AH284 Liver	183,327 (2.3%)	7,995,538
AH284 PDAC	2,283,436 (18.8%)	12,124,201

Example 3. Overexpression of satellite transcripts in human pancreatic cancer and other epithelial cancers

To test whether human tumors also overexpress satellite ncRNAs, we extended the DGE analysis to specimens of human pancreatic cancer. Human pancreatic tumor tissues were obtained as excess discarded human material per IRB protocol from the Massachusetts General Hospital. Gross tumor was excised and fresh frozen in liquid nitrogen prior to nucleic acid extraction. Normal pancreas RNA was obtained from two commercial vendors, Clontech and Ambion. The samples were prepared and analyzed as described above in Example 1.

Analysis of 15 PDACs showed a median 21-fold increased expression of total satellite transcripts compared with normal pancreas. A cohort of non-small cell lung cancer, renal cell carcinoma, ovarian cancer, and prostate cancer also had significant levels of satellites and the HSATII satellite. Other normal human tissues, including fetal brain, brain, colon, fetal liver, liver, lung, kidney, placenta, prostate, and uterus have somewhat higher levels of total satellite expression (Table 4, Fig. 3A).

Table 4

SAMPLE ID	Genome	Total Satellite (tpm)	ALR (tpm)	HSATII (tpm)
PDAC 1	4,472,810	25,209	14,688	3,589
PDAC 2	1,668,281	22,001	12,653	3,295
PDAC 3	5,211,399	27,366	15,921	5,057
PDAC 4	1,649,041	23,556	13,428	3,167
PDAC 5	239,483	15,095	8,259	509
PDAC 6	1,520,470	374	195	14
PDAC 7	1,449,321	7,738	4,400	750
PDAC 8	1,950,197	574	316	9
PDAC 9	3,853,773	19,572	12,563	1,731
PDAC 10	2,748,850	28,225	18,767	2,489
PDAC 11	2,848,599	23,163	14,634	2,589
PDAC 12	3,723,326	21,243	12,940	2,122
PDAC 13	1,834,743	24,549	15,342	3,150
PDAC 14	2,481,332	25,650	18,016	2,564
PDAC 15	1,752,081	38,514	25,899	5,210
Normal Pancreas 1	1,196,372	908	284	0
Normal Pancreas 2	975,676	1,043	303	0
Lung Cancer 1	1,549,237	28,658	18,751	4,417
Lung Cancer 2	13,829,845	33,030	26,143	2,555
Kidney Cancer 1	2,104,859	10,814	6,505	1,501
Kidney Cancer 2	4,753,409	5,025	2,739	625
Ovarian Cancer 1	12,596,542	26,658	14,513	3,074
Ovarian Cancer 2	7,290,000	4,089	2,058	403
Prostate Cancer 1	3,376,849	43,730	22,244	9,793
Prostate Cancer 2	12,052,244	23,947	14,201	3,209
Prostate Cancer 3	3,631,148	21,411	12,390	2,804
Normal Fetal Brain	384,453	2,843	1,516	3
Normal Brain	371,161	5,184	2,573	3
Normal Colon	183,855	13,059	7,229	5
Normal Fetal Liver	147,977	11,218	5,879	7
Normal Liver	117,976	7,968	3,730	25
Normal Lung	208,089	15,027	7,857	5
Normal Kidney	144,173	15,218	8,094	7
Normal Placenta	207,929	13,990	7,815	0
Normal Prostate	263,406	8,409	2,228	19
Normal Uterus	477,480	2,702	1,395	2

Subdivision of human satellite among the multiple classes revealed major differences between tumors and all normal tissues. While mouse satellite repeats are broadly subdivided into major and minor satellites, human satellites have been classified more extensively. Of all human satellites, the greatest expression fold differential is evident for the pericentromeric satellite HSATII (mean 2,416 tpm; 10.3% of satellite reads), which is undetectable in normal human pancreas (Fig. 3B). In contrast, normal tissues have much higher representation of GSATII, Beta satellite (BSR), and TAR1 classes (21.1%, 17.3%, and 2.1% of all satellite reads respectively), while these constitute a small minority of satellite reads in pancreatic cancers.

The most abundant class of normally expressed human satellites, alpha (ALR) (Okada et al., Cell 131, 1287 (Dec 28, 2007)) is expressed at 294 tpm in normal human pancreas, but comprises on average 12,535 tpm in human pancreatic adenocarcinomas (43-fold differential expression; 60.3% of satellite reads). Thus, while the overexpression of human ALR repeats is comparable to that of mouse major satellite repeats, it is the less abundant HSATII (49-fold above *GAPDH*), which shows exceptional specificity for human PDAC. The co-expression of *LINE*-1 with satellite transcripts in human pancreatic tumors is also striking, with a mean 16,089 tpm (range 358-38,419).

Beyond ALR repeats, the satellite expression profile of normal pancreas and PDAC are strikingly different; for instance normal pancreatic tissue has a much higher representation of GSATII, TAR1 and SST1 classes (26.4%, 10.6%, and 8.6% of all satellite reads), while these were a small minority of satellite reads in pancreatic cancers. In contrast, cancers express high levels of HSATII satellites (4,000 per 10^6 transcripts; 15% of satellite reads), a subtype whose expression is undetectable in normal pancreas (Fig. 3B). Quantitative comparison of satellite transcription in mouse versus human pancreatic cancers shows that mouse major satellites are expressed a median 466-fold above the abundant *Gapdh* mRNA, while the human ALR and HSATII satellites are respectively expressed 180-fold and 47-fold above *GAPDH.*

<u>Example 4. Cellular transcripts with linear correlation to increasing satellite levels are enriched for stem cell and neural elements that is linked to histone demethylases and RNA processing enzymes</u>

The generation of comprehensive DGE profiles for 25 different mouse tissues of different histologies and genetic backgrounds made it possible to correlate the expression of cellular transcripts with that of satellites across a broad quantitative range. To identify such co-regulated genes, all annotated transcripts quantified by DGE were subjected to linear regression analysis, and transcripts with the highest correlation coefficients to satellite expression were rank ordered.

All mouse sample reads were aligned to a custom made library for the mouse major satellite (sequence from UCSC genome browser). Human samples were aligned to a custom made reference library for all satellite repeats and LINE-1 variants generated from the Repbase library (Pushkarev et al., Nat Biotech 27, 847 (2009)). In addition, all samples were subjected to the DGE program for transcriptome analysis. Reads were normalized per 10^6 genomic aligned reads for all samples.

For linear correlation of mouse major satellite to transcriptome, all tissues and cell lines were rank ordered according to level of major satellite. All annotated genes were then subjected to linear regression analysis across all tissues. Genes were then ordered according to the Pearson coefficient for linear regression and plotted by Matlab.

Analysis of a set of 297 genes with highest linear correlation ($R > 0.85$) revealed 190 annotated cellular mRNAs and a subset of transposable elements (Fig. 4A).

Of all transcripts analyzed with high linear correlation, the autonomous retrotransposon *Line-1* had the highest expression level in mouse samples of diverse tissue types. Mouse pancreatic tumors have a mean Line-1 expression 30,690 tpm (range 183-120,002), representing an average of 330-fold higher levels compared to *Gapdh* (Table 5).

Table 5

	Normalized Reads per 10^6 transcripts	
	LINE-1	**GAPDH**
Mouse PDAC	30,690 (183-120,002)	171 (19-417)
Human PDAC	6,091 (5,153-6,921)	48 (26-67)

The co-expression of *LINE*-1 with satellite transcripts in human pancreatic tumors was also striking, with an average of 6,091 per 10^6 transcripts (127-fold higher than *GAPDH*). Increased expression of the *Tigger* transposable elements 3 and 4 were also correlated with increasing satellite transcription in mouse tumors, but was not seen human tumors.

In addition to retroelements, a subset of cellular mRNAs showed a very high degree of correlation with the levels of satellite repeat expression across diverse mouse tumors (referred to herein as "Satellite Correlated Genes (SCGs)"). Linearly correlated genes with R > 0.85 were mapped using the DAVID program (Dennis, Jr. et al., Genome Biol 4, P3 (2003); Huang et al., Nat Protoc 4, 44 (2009)). These genes were then analyzed with the Functional Annotation clustering program and the UP_TISSUE database to classify each of these mapped genes. Germ/Stem cell genes included genes expressed highly in testis, egg, trophoblast, and neural stem cells. Neural genes included genes expressed highly in brain, spinal cord, and specialized sensory neurons including olfactory, auditory, and visual perception. HOX and Zinc Finger proteins were classified using the INTERPRO database.

Analysis of 190 annotated transcripts using the DAVID gene ontology program identified 120 (63%) of these transcripts as being associated with neural cell fates and 50 (26%) linked with germ/stem cells pathways (Table 6).

Table 6

	Germ/Stem Cell	**Neural**	**HOX Region**	**Zinc Finger Domain**
TOTAL COUNTS	50	120	10	16
% Mapped (190)	26%	63%	5%	8%

In addition, significant enrichment was evident for transcriptional regulators, including HOX related (9, 5%) and zinc finger proteins (16, 8%). This gene set could not be matched to any known gene signature in the GSEA database (Subramanian et al., Proc Natl Acad Sci U S A 102, 15545 (Oct 25, 2005)), but the ontology analysis points towards a neuroendocrine phenotype. Neuroendocrine differentiation has been described in a variety of epithelial malignancies, including pancreatic cancer (Tezel et al., Cancer 89, 2230 (Dec 1, 2000)), and is best characterized in prostate cancer where it is correlated with more aggressive disease (Cindolo et al., Urol Int 79, 287 (2007)). A striking increase in the number of carcinoma cells staining for the characteristic neuroendocrine marker chromogranin A, as a function of higher satellite expression in mouse PDACs

(Fig. 4D), was observed, supporting the link between globally altered expression of ncRNAs and a specific cellular differentiation program.

SCGs were more readily identified in mouse tumors, since the large dynamic range in major satellite expression enabled linear correlation in expression level between satellites and protein-encoding genes. However, human orthologs were identified for 138 of the 190 annotated mouse SGCs, of which 54 (39%) showed >2-fold increased expression in human PDACs compared with normal pancreas (q-value < 0.1). Together, these observations suggest that, as in the mouse genetic model, tumor-associated derepression of satellite-derived repeats is highly correlated with increased expression of *Line-1* and a subset of cellular mRNAs.

Histone modifications, including H3K9 trimethylation (P. A. Cloos, J. Christensen, K. Agger, K. Helin, Genes Dev 22, 1115 (May 1, 2008)), combined with *Dicer1* and Piwi-related protein-mediated ncRNA processing (A. A. Aravin, G. J. Hannon, J. Brennecke, Science 318, 761 (Nov 2, 2007)) have been linked to maintenance of repression of satellite repeats. To search for candidate regulators of satellite derepression in primary tumor specimens, we first measured the quantitative DGE of known epigenetic regulators and RNA processing genes in mouse tumors, as a function of increasing major satellite expression.

A targeted gene expression analysis of demethylases and RNA processing enzymes was carried out in mouse and human PDAC samples. A list of demethylases and RNA processing enzymes were generated from two recent publications (Cloos et al., Genes Dev 22, 1115 (2008); Aravin et al., Science 318, 761 (2007)). Mouse PDACs with *Kras*^{G12D} and *Tp53* loss were used for this analysis. Mouse tumors were separated into high vs low satellite levels using the median satellite expression (7%). A total of 37 genes were evaluated between high and low satellite tumors and fold change was calculated. Analysis of the population means was compared using the 2-tailed student t-test assuming equal variance. Genes that had a significance value of $p < 0.05$ (total of 7 gene) were then used to evaluate human PDAC versus human normal pancreatic tissue. Fold change and a 2-tailed student t-test with equal variance was then applied to human PDAC vs normal pancreatic tissue to find potential gene candidates involved with satellite expression.

Among 37 candidate genes tested, mouse pancreatic tumors with satellite expression above the median had higher expression of the demethylases *Hspbap1, Jmjd1B, Jmjd4, Jarid1d, Jmjd3,* and Fbxl10 as well as the RNA processing enzyme *Dicer1*. Among these, *HSPBAP1, FBXL10* and *DICER1* overexpression was also observed in human pancreatic adenocarcinomas (Table 7, $p < 0.05$, student t-test).

Table 7

List of candidate demethylases and RNA processing enzymes identified in mouse PDAC tumors compared to human PDAC.

	Mouse PDAC High vs Low Satellite		**Human PDAC vs Normal Panc**	
GENE NAME	FOLD Expression	T-test p-value	FOLD Expression	T-test p-value
HSPBAP1	**5.11**	**0.0005**	**11.59**	**0.0069**
DICER1	**2.56**	**0.0006**	**3.00**	**0.0023**
JMJD1B (KDM3B)	-2.16	0.0010	1.16	0.8051
JMJD4	3.10	0.0021	1.18	0.8080
JARID1D (SMCY, KDM5D)	9.84	0.0031	1.62	0.7107
JMJD3 (KDM6B)	1.62	0.0118	3.02	0.1109
FBXL10 (KDM2B)	**1.40**	**0.0412**	**7.96**	**0.0279**

Demethylases and RNA processing enzymes enriched in high vs low major satellite expressing mouse tumors (First two columns) ordered by highest significance (lowest p-value). Human PDAC vs normal tissue fold change and t-test p-value shown for each of these genes (Last two columns). Genes differentially expressed in both human and mouse tumors with a significance value of $p < 0.05$ are highlighted in bold.

The catalytic activity of HSPBAP1 and FBXL10 demethylases has not been extensively characterized, although the former is noteworthy for its contribution to the familial renal cancer *DIRC3-HSPBAP1* fusion (D. Bodmer, M. Schepens, M. J. Eleveld, E. F. Schoenmakers, A. Geurts van Kessel, Genes Chromosomes Cancer 38, 107 (Oct, 2003)) and the latter appears to have some specificity for H3K36me2/me1 and H3K4me3, with an effect on ribosomal RNA expression and cellular proliferation (D. Frescas, D. Guardavaccaro, F. Bassermann, R. Koyama-Nasu, M. Pagano, Nature 450, 309 (Nov 8, 2007)). While current understanding of multicomponent chromatin modifier complexes precludes linking satellite and *Line-1* upregulation in primary tumors with aberrant expression of a single transcript, the relatively small number of genes with consistently altered expression may point to a key subset of epigenetic regulators that contribute to satellite and *Line-1* derepression.

While failure of a common transcriptional silencing mechanism may contribute to repression of *LINE-1* and satellite repeats, the diverse array of cellular transcripts whose overexpression is correlated with these repetitive sequences is less readily explained. Of note, recent findings have demonstrated that *LINE-1* may drive expression of specific cellular mRNAs through its insertion upstream of their transcriptional start sites (T. Kuwabara et al., Nat Neurosci 12, 1097 (Sep, 2009)) or through alterations in flanking chromatin marks (J. A. Bailey, L. Carrel, A. Chakravarti, E. E. Eichler, Proceedings of the National Academy of Sciences of the United States of America 97, 6634 (June 6, 2000, 2000); D. E. Montoya-Durango et al., Mutat Res 665, 20 (Jun 1, 2009)). To test whether such a mechanism might also underlie the co-expression of the 297 cellular mRNAs identified here, we analyzed the genomic distance between their transcriptional start site and *LINE-1* insertions in the genome.

The transcriptional start sites tissues and cell lines were determined (UCSC genome browser (D. Karolchik et al., Nucleic Acids Res 32, D493 (Jan 1, 2004))) as well as the position of all Line-1 elements in the mouse genome with a threshold of 1 Kbp in length. Line-1 closest distance upstream of the transcriptional start sites of all annotated genes with a minimum expression level of 5 transcripts per million were calculated. Genes were then rank ordered according to the Pearson coefficient for linear regression. Genes were binned in 100s and plotted by Excel. Randomization of all genes, followed by binning, and plotting was done as a control.

Focusing on the top linearly correlated genes ($R > 0.85$), these genes were plotted as a frequency plot against distance of Line-1 elements upstream of the transcriptional start site. Enrichment was calculated at 5 Kbp and the Fisher Exact test was used to calculate the test statistic.

Remarkably, there was a striking correlation of *LINE-1* genomic distance to expression of genes with the major satellite (Fig. 4B), and there was highly significant enrichment of our top 297 genes for presence of a *LINE-1* element within 5 Kbp of the transcriptional start site (Enrichment 2.69, $p = 8.18 \times 10^{-7}$, Fisher exact test, Fig. 4C).

Thus, activation of *LINE-1* sequences within the proximity of cellular transcripts may contribute to their overexpression in primary tumors, in striking correlation with the expression of both *LINE-1* and satellite repeats. The consequence of increased expression of these cellular transcripts remains to be defined. However, the high prevalence of genes linked to stem-like and neurogenic fates, along with the frequency of HOX and zinc finger transcriptional regulators raises the possibility that at least a subset of these may contribute to tumor-related phenotypes.

Example 5. LINE-1 is a specific and sensitive marker of CTCs

Satellite levels are most strongly linked with the expression of the autonomous retrotransposon *Line-1*, which has recently been shown to be a major cause for genomic variation in normal and tumor tissues (J. Berretta, A. Morillon, EMBO Rep 10, 973 (Sep, 2009); A. Jacquier, Nat Rev Genet 10, 833 (Dec, 2009); M. Guenatri, D. Bailly, C. Maison, G. Almouzni, J Cell Biol 166, 493 (Aug 16, 2004)). Aberrant expression of cellular transcripts linked to stem cells and neural tissues is also highly correlated with satellite transcript levels, suggesting alteration of cell fate through derepression of a coordinated epigenetic program.

Expression of LINE-1 in circulating tumor cells (CTCs) in newly diagnosed metastatic pancreatic adenocarcinoma patients was evaluated using a CTC device known as a herringbone chip (HB), which combines specific antibody mediated capture against the epithelial cell adhesion molecule (EpCAM) and the high-throughput advantages of microfluidics (Stott et al., Proc Natl Acad Sci U S A, 107(43):18392–18397 2010). Blood was collected from cancer patients that had given consent. Approximately 3 mL of blood was processed on the CTC device and a control device over 2 hours. RNA was extracted from the devices using the Qiagen RNeasy MinElute kit. RNA was then subjected to cDNA synthesis with random primers using the Superscript III first strand synthesis kit (invitrogen). RNA was removed with RNase and cDNA was subject to qRT-PCR using Human LINE-1 Taqman assay (Applied Biosystems). LINE-1 expression was normalized to GAPDH in the CTC and control device. Fold difference between the CTC device and control device was then calculated by standard Ct calculation for qPCR. As shown in Figure 5, LINE-1 expression was seen much more consistently and with higher frequency compared to keratins (Krt) which are typical CTC markers (S. Maheswaran et al., N Engl J Med. 359, 366 (Jul 24, 2008); S. Nagrath et al., Nature. 450, 1235 (Dec 20, 2007)). Preliminary data of HSATII positive cells suggests a 10 fold improvement in sensitivity of CTC detection. This demonstrates that LINE-1 is a specific and sensitive marker for detecting CTCs. The ability of satellites to increase the sensitivity of detecting pancreatic CTCs offers not only a more robust blood based diagnostic for cancer treatment monitoring, but a significant improvement in the ability to use CTCs as an early detection modality. This provides a screening tool for cancer, improving chances of early detection, which can improve the ability to provide curative therapy.

Example 6. HSATII RNA in situ technique using branched DNA detection technology

As noted above, the HSATII satellite is overexpressed in pancreatic cancer and was confirmed to be overexpressed in human preneoplastic pancreatic lesions (Fig. 6A) using a branched DNA detection assay (QuantiGene® ViewRNA Assay, Affymetrix). Breast cancer samples were also tested for HSATII using this method and were found to have significant expression compared to normal breast tissues. Extension of this technique to potential circulating tumor cells captured on the HB-chip (Fig. 6B) has been accomplished indicating that HSATII may be used as a blood based diagnostic for epithelial cancers.

Example 7. Satellites Levels in Serum

Serum was extracted from the blood of 8 metastatic pancreatic cancer patients by using Ficoll buffy coat method. Serum RNA (cell free RNA), which includes exosomes, was purified using the Trizol method and then purfied using Qiagen RNA MinElute columns kits. RNA was then subjected to Helicos DGE sequencing preparation and sequenced on a HeliScope next generation sequencer. Results of this data are summarized in Table 1. As described above, HSATII was specific for cancer and GSATII was found to correlate with normal tissues. Therefore the ratio of HSATII to GSATII was evaluated as a marker for identifying cancer burden and potentially an early detection marker. In this case, one patient who had stable disease had the lowest HSATII/GSATII ratio as predicted (see Table 8). These results suggest that detection of satellites in peripheral blood serum (cell free RNA) can be used as a predictive marker of disease response to therapy.

Table 8.

Patient ID	Clinical status	Total Satellites (tpm)	HSATII (tpm)	GSATII (tpm)	HSATII/GSATII
PDAC 3	PROGRESSION	43,932	1,576	4,904	32%
PDAC 6	PROGRESSION	21,845	735	3,151	23%
PDAC 9	PROGRESSION	39,235	1,867	1,857	101%
PDAC 11	PROGRESSION	28,817	784	3,785	21%
PDAC 12	PROGRESSION	2,472	59	83	71%
PDAC 16	STABLE	43,629	162	6,437	3%
PDAC 18	PROGRESSION	18,034	231	2,450	9%
PDAC 19	PROGRESSION	38,425	399	5,287	8%

Table 8: A total of 8 metastatic cancer patients with clinical status, total satellites, HSATII, and GSATII in transcripts per million aligned to genome (tpm) and the ratio of HSATII/GSATII in cell free RNA sequenced.

However, in a preliminary evaluation of healthy donor serum (cell free) RNA (n=4) HSATII and GSATII did not perform as well as expected, though the presence of . However, other satellites like TAR1 seemed to be better predictors of "cancer" compared to "non-cancer" status as shown in Table 9. TAR1 was significantly different between the two populations with a p value = 0.025.

Table 9.

	Total Satellites (tpm)	HSATII (tpm)	GSATII (tpm)	HSATII/GSATII	TAR1 (tpm)
AVG PDAC	29,549	727	3,494	0.33	100
AVG HD	47,279	6,275	6,081	1.12	51
TTEST	0.114	0.172	0.103	0.25	**0.025**
FOLD	0.625	0.116	0.575	0.30	1.963

Table 9: Average total satellites, HSATII, GSATII, HSATII/GSATII, and TAR1 (tpm) in a total of 8 metastatic PDAC patients and 4 healthy donors with cell free RNA sequenced. Student t-test was used to calculate significance.

OTHER EMBODIMENTS

It is to be understood that while the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims. Other aspects, advantages, and modifications are within the scope of the following claims.

WHAT IS CLAIMED IS:

1. An in vitro method of detecting the presence of cancer in a subject, the method comprising:
   determining a level of LINE-1 in a sample from the subject to obtain a test value; and
   comparing the test value to a reference value,
   wherein a test value compared to the reference value indicates whether the subject has cancer.

2. The method of claim 1, wherein the reference value represents a threshold level of LINE-1, wherein the presence of a level of LINE-1 in the subject that is above the reference value indicates that the subject has cancer, and the presence of a level of LINE-1 in the subject that is below the reference value indicates that the subject is unlikely to have cancer.

3. An in vitro method of detecting the presence of cancer in a subject, the method comprising:
   determining a level of satellite transcripts in a sample from the subject to obtain a test value; and
   comparing the test value to a reference value,
   wherein a test value compared to the reference value indicates whether the subject has cancer.

4. The method of claim 2, wherein the satellite transcripts comprise one or more of ALR, HSATII, GSATII, TAR1, and SST1.

5. The method of claim 4, wherein the satellite transcript is ALR and/or HSATII, and the presence of a level of ALR and/or HSATII satellite transcripts above the reference level indicates that the subject has a tumor.

6. The method of claim 4, wherein the satellite transcript is GSATII, TAR1 and/or SST1, and the presence of a level of GSATII, TAR1 and/or SST1 satellite transcripts below the reference level indicates that subject has a tumor.

7. The method of any of claims 1 to 6, wherein the sample is known or suspected to comprise tumor cells.

8. The method of claim 7, wherein the sample is a blood sample known or suspected of comprising circulating tumor cells (CTCs), or a biopsy sample known or suspected of comprising tumor cells.

9. The method of any of claims 1 to 6, wherein the sample comprises free RNA in serum or RNA within exosomes in blood.

10. The method of any of claims 1 to 8, wherein the subject is a human.

11. The method of any of claims 1 to 10, wherein the cancer is a solid tumor of epithelial origin.

12. An in vitro method of evaluating the efficacy of a treatment for cancer in a subject, the method comprising:
determining a level of LINE-1 in a first sample from the subject to obtain a first value;
administering a treatment for cancer to the subject;
determining a level of LINE-1 in a subsequent sample obtained from the subject at a later time, to obtain a treatment value; and
comparing the first value to the treatment value,
wherein a treatment value that is below the first value indicates that the treatment is effective.

13. An in vitro method of evaluating the efficacy of a treatment for cancer in a subject, the method comprising:
determining a level of satellite transcripts in a first sample from the subject to obtain a first value;
administering a treatment for cancer to the subject;
determining a level of satellite transcripts in a subsequent sample obtained from the subject at a later time, to obtain a treatment value; and
comparing the first value to the treatment value,

wherein a treatment value that is below the first value indicates that the treatment is effective.

14. The method of claim 12 or 13, wherein the satellite transcripts comprise one or more of ALR, HSATII, GSATII, TAR1, and SST1.

15. The method of any of claims 12 to 14, wherein the first and second samples are known or suspected to comprise tumor cells.

16. The method of claim 15, wherein the samples are blood samples known or suspected of comprising circulating tumor cells (CTCs), or biopsy samples known or suspected of comprising tumor cells.

17. The method of any of claims 12 to 16, wherein the sample comprises free RNA in serum or RNA within exosomes in blood.

18. The method of any of claims 12 to 17, wherein the treatment includes administration of a surgical intervention, chemotherapy, radiation therapy, or a combination thereof.

19. The method of claim any of claims 12 to 18, wherein the subject is a human.

20. The method of any of claims 12 to 19, wherein the cancer is a solid tumor of epithelial origin.

21. The method of claim 1, 2, or 12, comprising measuring a level of LINE-1 transcript.

22. The method of any of claims 13 to 21, wherein the level is determined using a branched DNA assay.

23. The method of claims 11 or 20, wherein the solid tumor of epithelial origin is pancreatic, lung, breast, prostate, renal, ovarian or colon cancer.

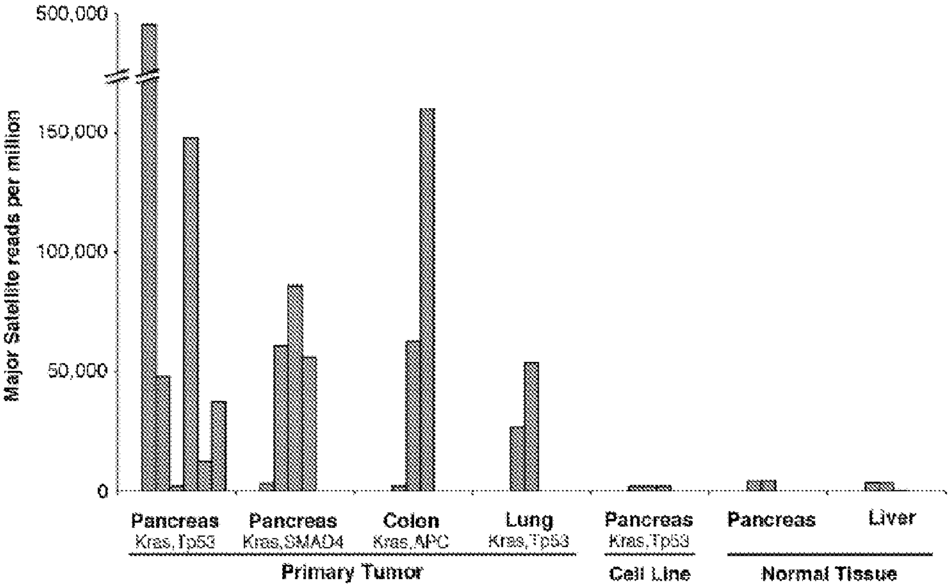

Major Satellite reads per million
500,000
150,000
100,000
50,000
0
Pancreas
Kras,Tp53
Pancreas
Kras,SMAD4
Colon
Kras,APC
Lung
Kras,Tp53
Pancreas
Kras,Tp53
Pancreas
Liver
Primary Tumor
Cell Line
Normal Tissue

FIG. 1A

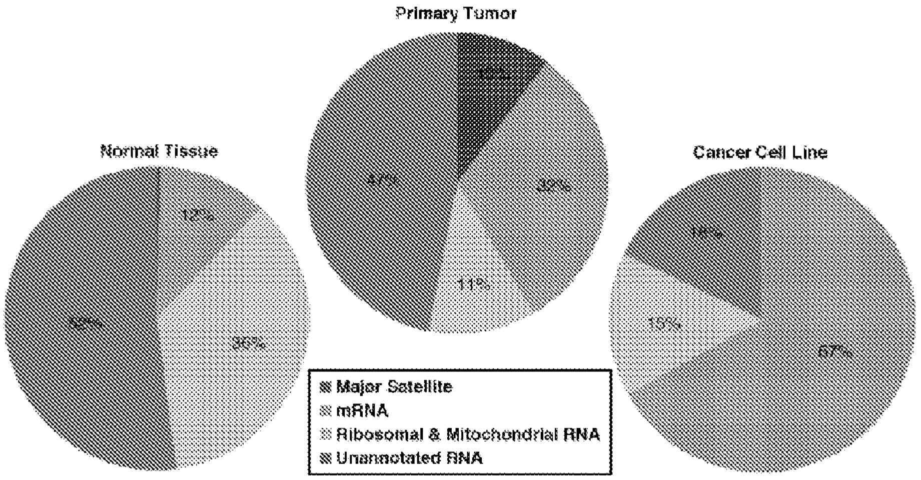
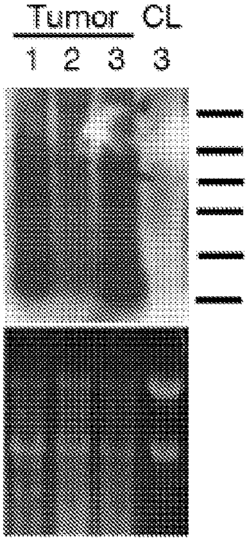

Primary Tumor
10%
32%
11%
47%
Normal Tissue
12%
36%
52%
Cancer Cell Line
18%
15%
67%
▨ Major Satellite
▨ mRNA
▨ Ribosomal & Mitochondrial RNA
▨ Unannotated RNA

FIG. 1B



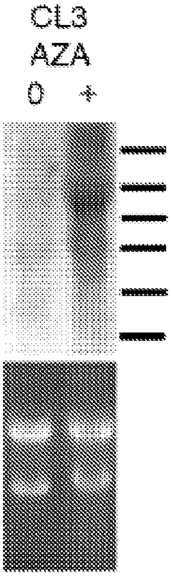


FIG. 2A

FIG. 2B

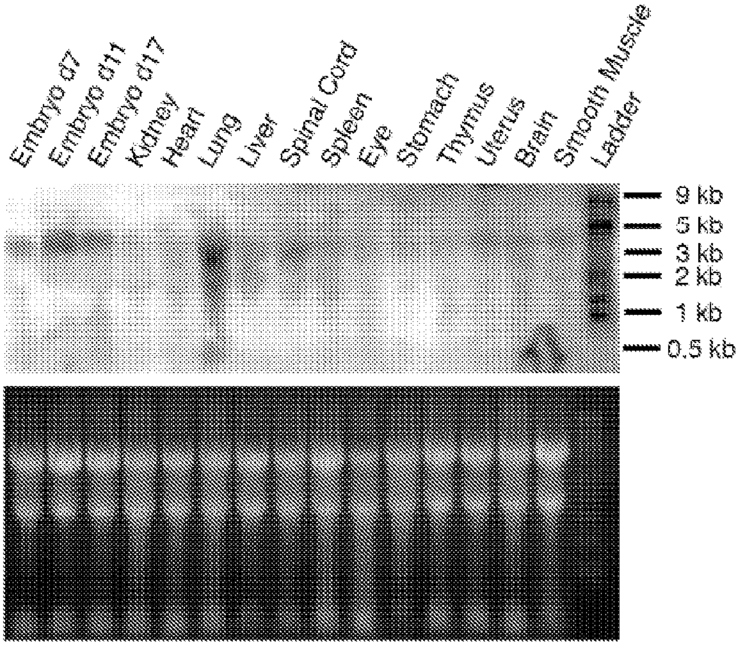


FIG. 2C

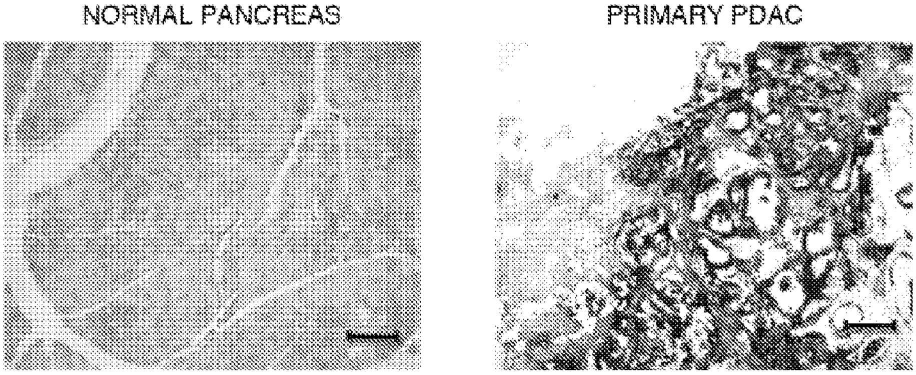


FIG. 2D

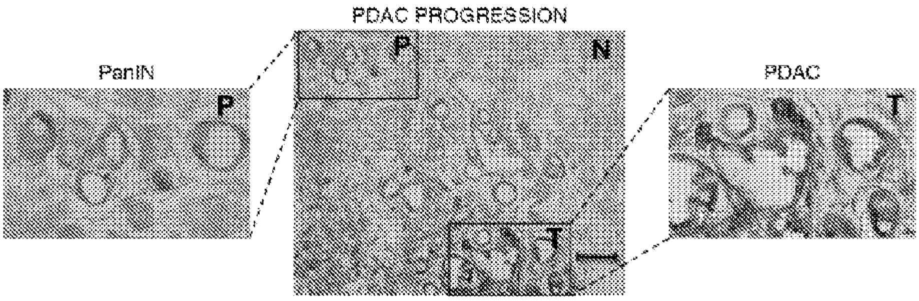

PDAC PROGRESSION
PanIN
P
P
N
T
PDAC
T

FIG. 2E

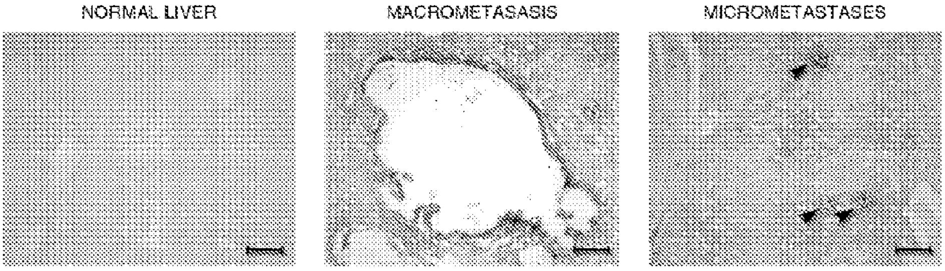

NORMAL LIVER
MACROMETASASIS
MICROMETASTASES

FIG. 2F

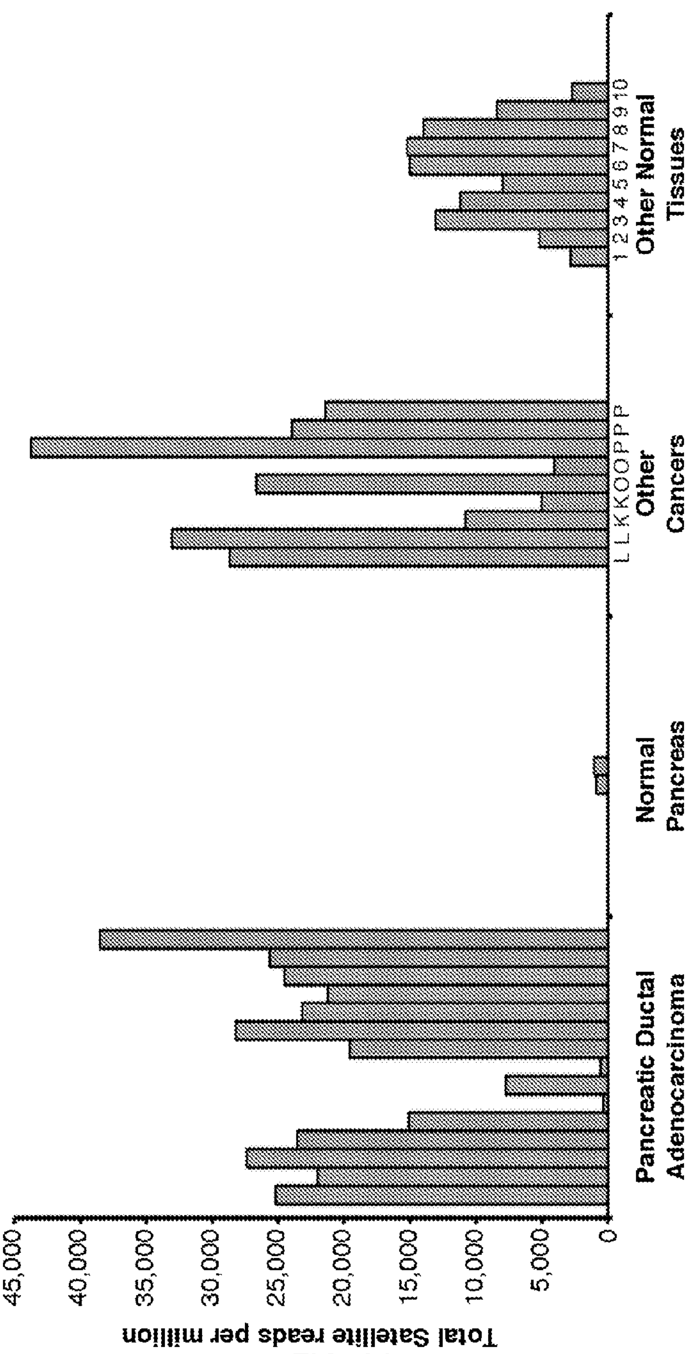

Total Satellite reads per million
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
Pancreatic Ductal Adenocarcinoma
Normal Pancreas
L L K K O O P P P
Other Cancers
1 2 3 4 5 6 7 8 9 10
Other Normal Tissues

FIG. 3A

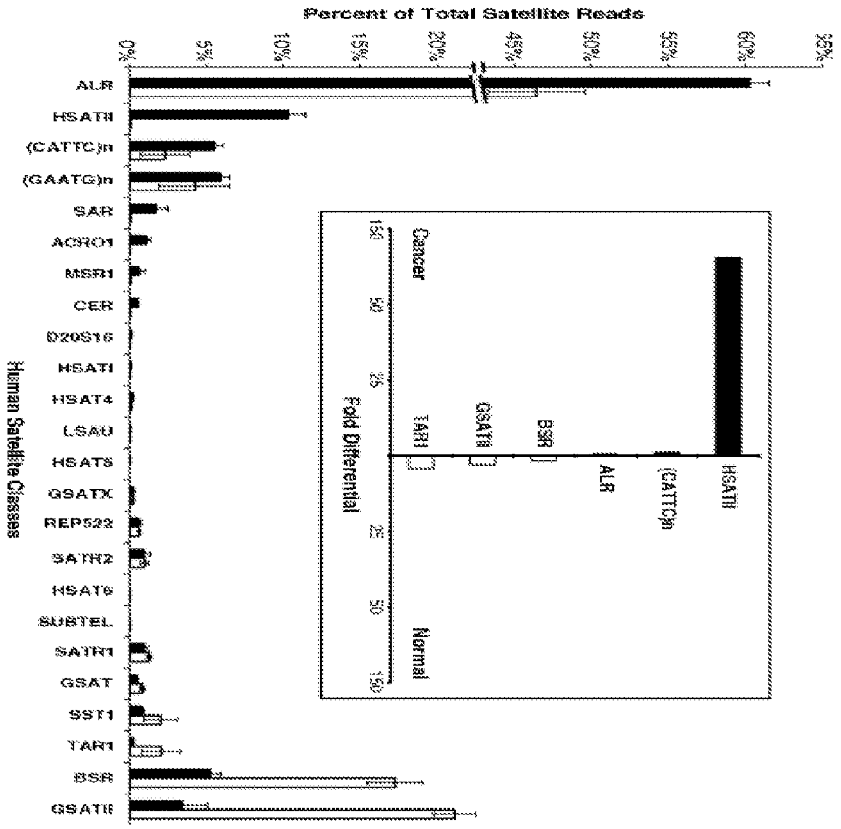

Percent of Total Satellite Reads
0%
5%
10%
15%
20%
45%
50%
55%
60%
65%
Human Satellite Classes
ALR
HSATII
(CATTC)n
(GAATG)n
SAR
ACRO1
MSR1
CER
D20S16
HSATI
HSAT4
LSAU
HSAT5
GSATX
REP522
SATR2
HSAT6
SUBTEL
SATR1
GSAT
SST1
TAR1
BSR
GSATII
Cancer
Normal
Fold Differential
150
50
25
25
50
150
TAR1
GSATII
BSR
ALR
(CATTC)n
HSATII

FIG. 3B

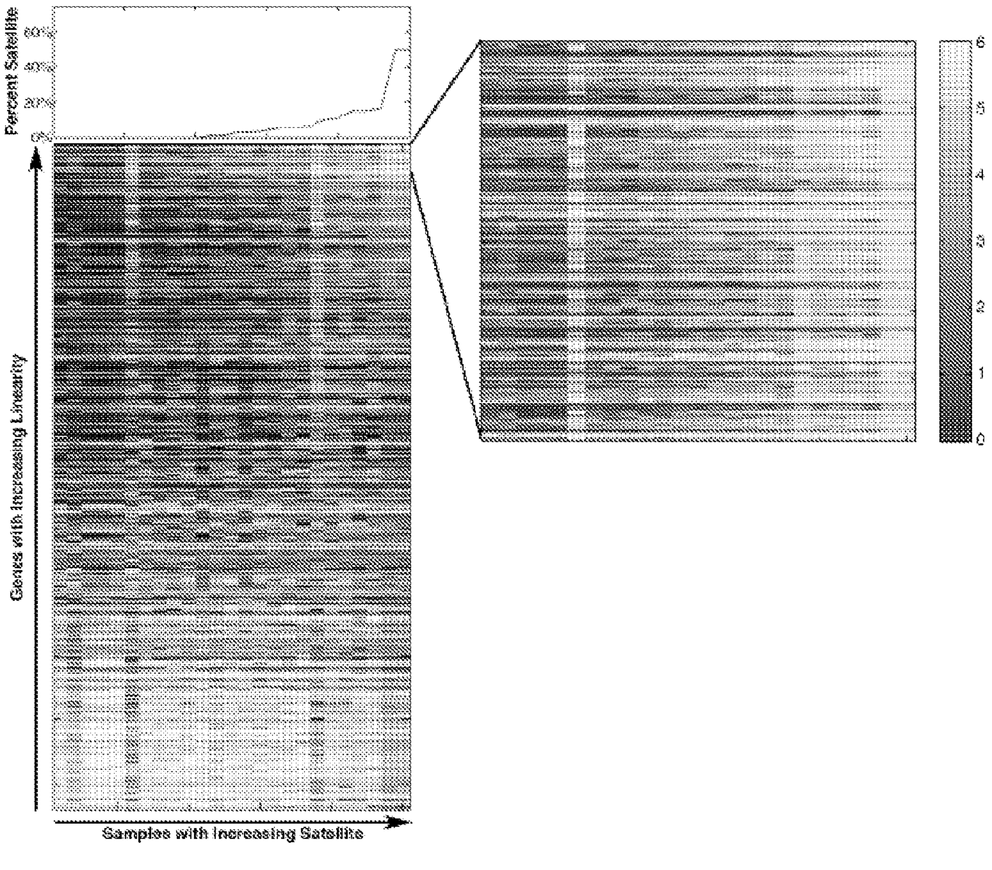

Percent Satellite
60%
40%
20%
0%
Genes with Increasing Linearity
Samples with Increasing Satellite
6
5
4
3
2
1
0

FIG. 4A

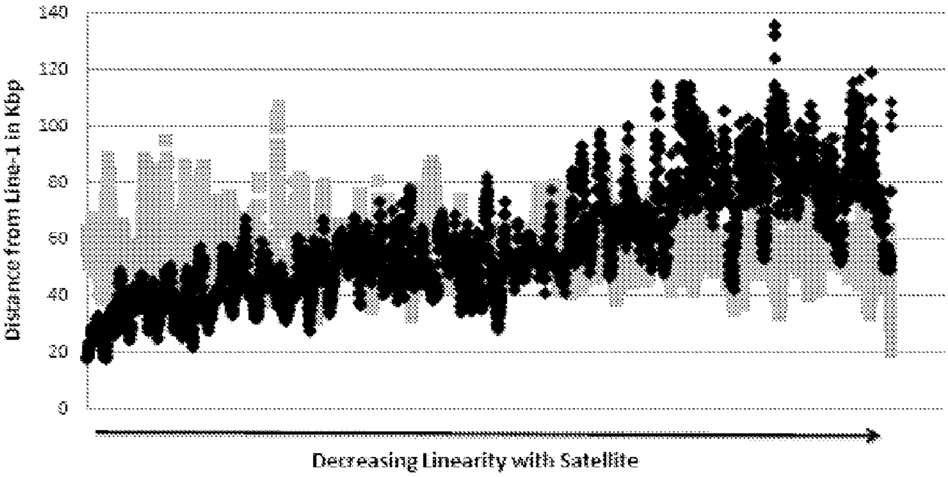


FIG. 4B

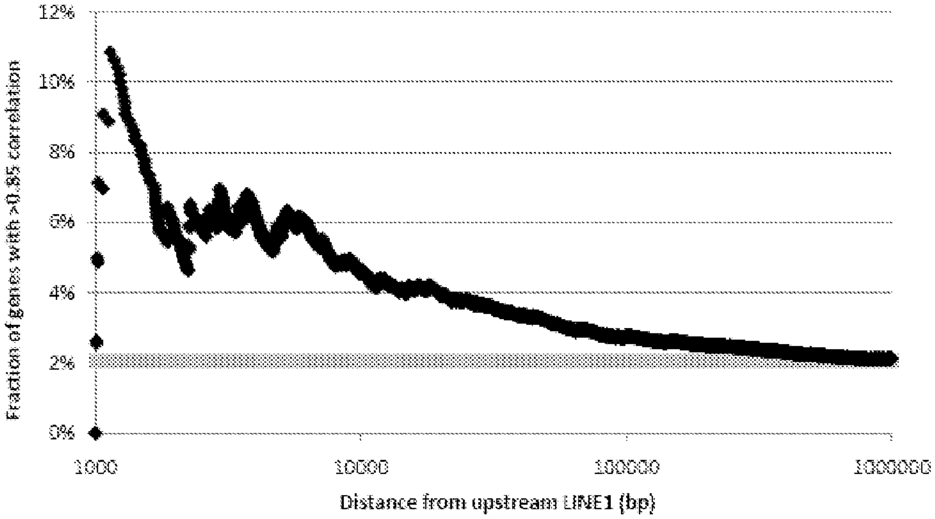


FIG. 4C

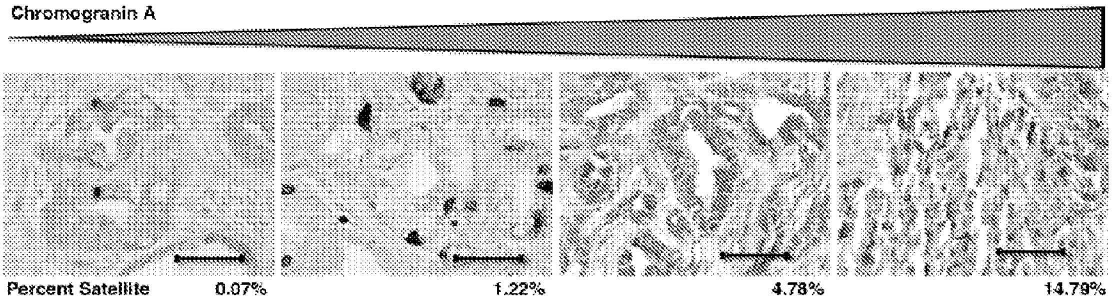


FIG. 4D

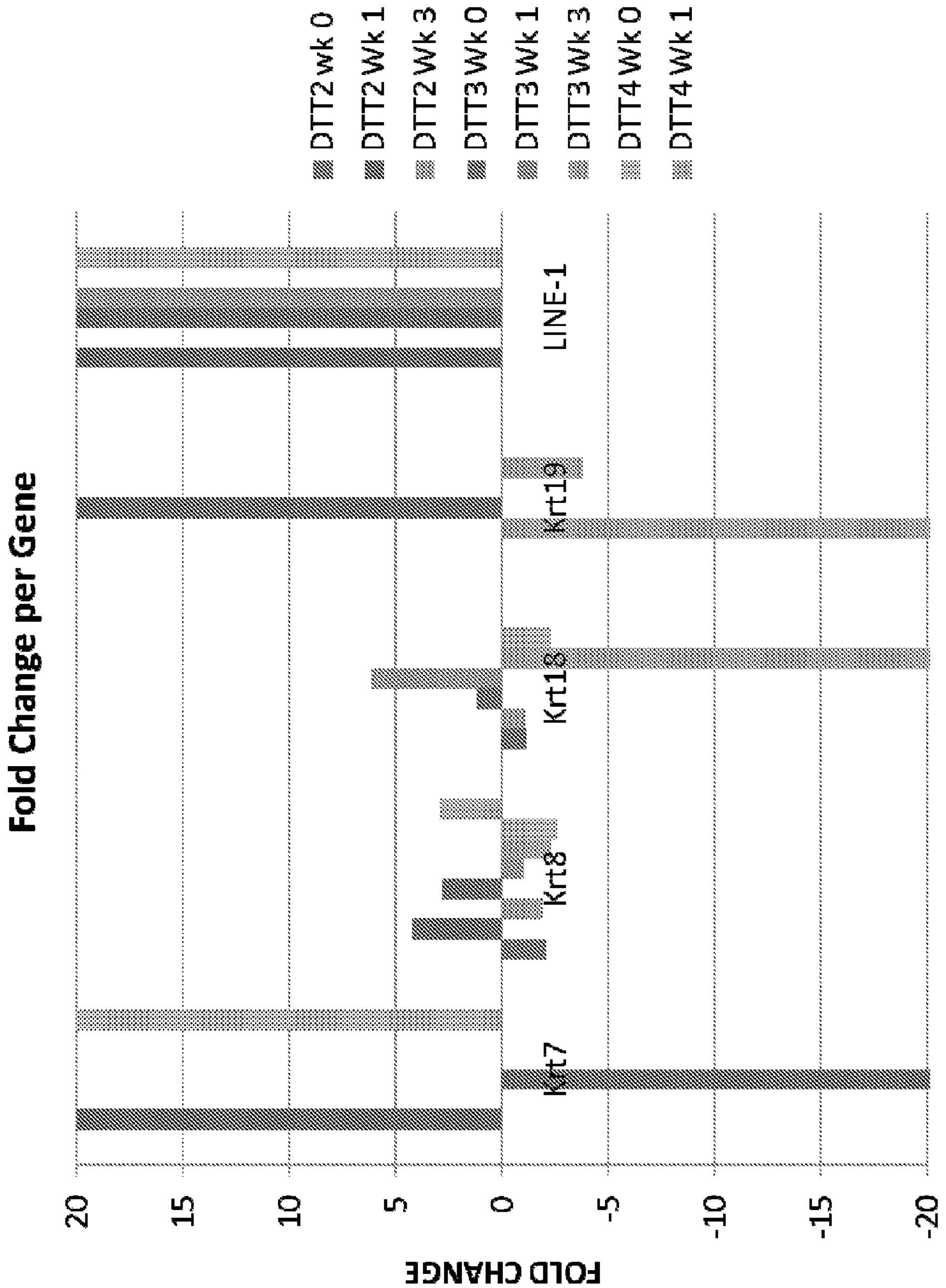
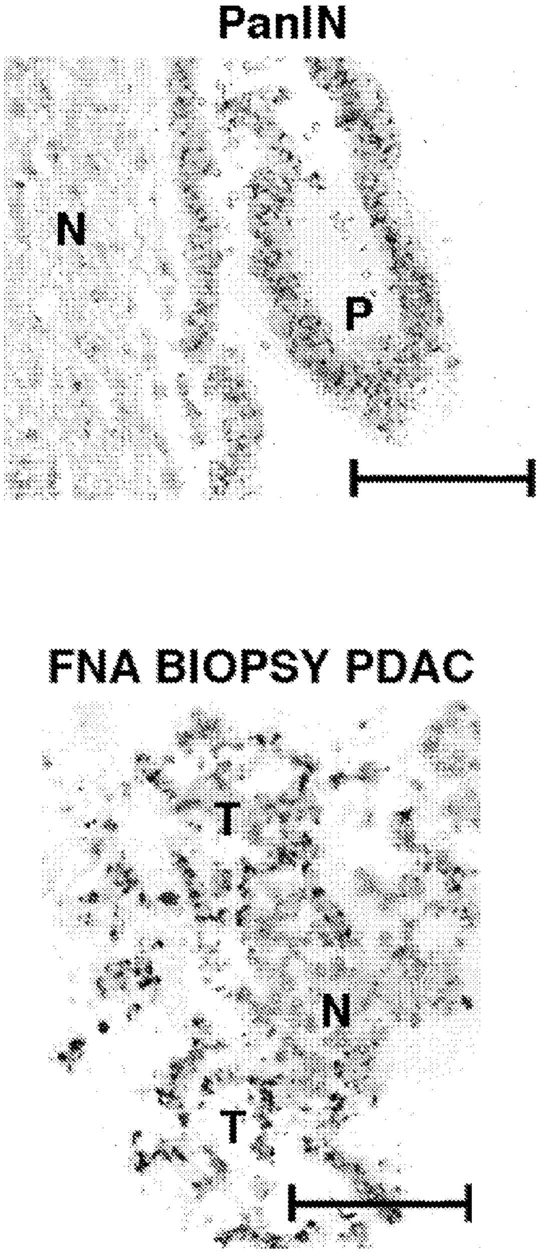

Fold Change per Gene
DTT2 wk 0
DTT2 Wk 1
DTT2 Wk 3
DTT3 Wk 0
DTT3 Wk 1
DTT3 Wk 3
DTT4 Wk 0
DTT4 Wk 1
20
15
10
5
0
-5
-10
-15
-20
FOLD CHANGE
Krt7
Krt8
Krt18
Krt19
LINE-1

FIG. 5



PanIN

FNA BIOPSY PDAC

FIG. 6A

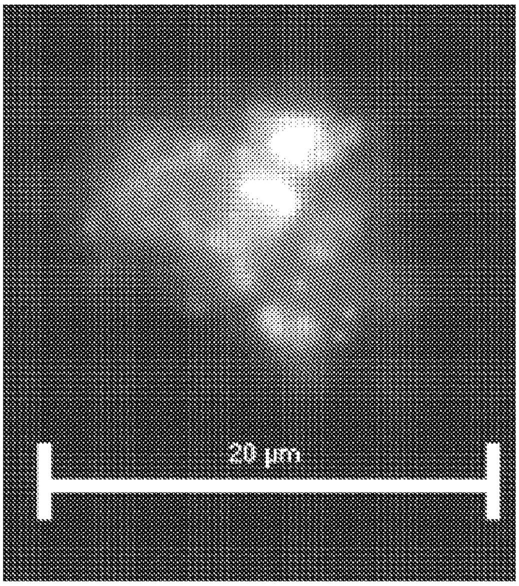


FIG. 6B