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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 June 2009 (25.06.2009)

PCT

(10) International Publication Number
WO 2009/077864 A2

(51) International Patent Classification:
G01N 33/53 (2006.01) *G01N 33/574* (2006.01)

(21) International Application Number:
PCT/IB2008/003836

(22) International Filing Date:
15 December 2008 (15.12.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
07301680.0 14 December 2007 (14.12.2007) EP

(71) Applicant *(for all designated States except US)*: **TRANSMEDI SA** [FR/FR]; 15, rue du Bois de la Champelle, F-54500 Vandoeuvre Les Nancy (FR).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **BIHAIN, Bernard** [FR/FR]; 26, rue de Metz, F-54000 Nancy (FR). **OGIER, Virginie** [FR/FR]; 38 rue Raymond Poincaré, F-54000 Nancy (FR). **BRULLIARD, Marie** [FR/FR]; 45 boulevard Clémenceau, F-54000 Nancy (FR).

(74) Agents: **BECKER, Philippe** et al.; 25 rue Louis Le Grand, F-75002 Paris (FR).

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

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

Published:
— *without international search report and to be republished upon receipt of that report*

WO 2009/077864 A2

(54) **Title:** COMPOSITIONS AND METHODS OF DETECTING TIABS

(57) **Abstract:** The present invention relates to novel methods and products for assessing the physiological status of a subject. More particularly, the invention relates to methods of assessing the presence, risk or stage of a cancer in a subject by measuring the levels of antibodies against particular aberrant protein domains in a sample from the subject, or the presence or number of immune cells bearing TCR specific for such aberrant protein domains. The invention is also suitable to assess the responsiveness of a subject to a treatment, as well as to screen candidate drugs and design novel therapies. The invention may be used in any mammalian subject, particularly in human subjects.

COMPOSITIONS AND METHODS OF DETECTING TIABS

The present invention relates to novel methods and products for assessing the physiological status of a subject. More particularly, the invention relates to methods of assessing the presence, risk or stage of a cancer in a subject by measuring the levels of antibodies against particular aberrant protein domains in a sample from the subject, or the presence or number of immune cells bearing TCR specific for such aberrant protein domains. The invention is also suitable to assess the responsiveness of a subject to a treatment, as well as to screen candidate drugs and design novel therapies. The invention may be used in any mammalian subject, particularly in human subjects.

INTRODUCTION

Cancer is progressively becoming the leading cause of death in Western countries and strong therapeutic benefits, *i.e.*, event-free long-term survival of > 90 % of patients are obtained mostly in individuals diagnosed at early stage [1]. Cancer is a genetic disease with accumulation of mutations in oncogenes and tumor suppressor genes [2]. Genetic testing is useful for identifying individuals at risk for colon, lung, breast, ovary and neuro-endocrine cancers [3,4]. However, clinical management of patients with genetic risks is complex because of the lack of precision as to when a given individual will develop cancer [5].

Gene expression varies widely in cancer cells and analysis of differences in transcription patterns have led to definition of molecular signatures associated with good or bad prognosis [6]. Such signatures may guide and optimize therapeutic strategies, but again the prerequisite is prior identification of the tumor.

Massive efforts towards identification of reliable, early stage, cancer molecular markers detectable in accessible human body fluids are pursued through 3 main directions. First, changes in protein concentrations and/or isoforms between normal and cancer patients are analyzed using various separation and identification procedures [7,8]. Some of these methods are suitable for systematic screening, but the technology is

currently confronted with huge differences in concentrations of abundant plasma proteins (>>> mg/ml) compared to that of protein released by low-level tissue leakage (<< pg/ml). An alternative approach consists of identification through expression profiling of a few specific targets differentially expressed in cancers and development of sensitive assays to monitor their concentrations in plasma [9]. Both strategies have been successfully implemented, but none has yet provided markers sufficiently robust to be useful in systematic clinical screening [10,11]. The second axis of investigation is based on identification and characterization of circulating DNA in plasma. An early report suggested that the simple presence of circulating DNA in serum was diagnostic [12]; however, this is now questionable because healthy individual circulating free DNA concentrations are in the same range as that of cancer patients [13,14]. Maintenance of normal DNA methylation pattern is critical for proper cell function and its loss is among the earliest molecular alteration during carcinogenesis [15,16]. Several groups have reported detection of tumor-associated methylation patterns in serum, but the success rate varied greatly among different teams that used the same biomarkers and technology [17-20]. This is due both to the diversity of tumor DNA methylation patterns and to low abundance of tumor DNA that represents at most 0.12 % of somatically normal haploid genome [21,22]. Thus, detection of cancer somatic mutations in minute amounts of circulating cancer DNA is also too close to background levels to provide robust assays, even when considering recent improvements in sequencing technology [23,24]. The third axis consists of probing the immune system response to cancer by systematic search of auto-antibodies [25,26]. The presence of these antibodies has been established, but the process by which self molecules become immunogenic is not yet understood [27]. Screening of expression libraries constructed from cancer cell mRNA led to identification of a large number of low sensitivity antibodies that, when used in combination of >20, achieved up to 82 % sensitivity in prostate cancer patients [28,29]. An alternative method relies on identification of auto-antibody signature by immunoblotting of 2 D gel electrophoresis [30]. This yields subsequent identification primarily of proteins identified by auto-antibodies independently of cancer status and of a limited number of proteins reacting preferentially with cancer sera [30].

Our strategy is based on a different rationale that directly stems from the results obtained through large scale cancer DNA sequencing programs [31,32]. This important work led to the conclusion that cancer somatic mutations occur at rates, higher than expected, but nevertheless remain rare events: the estimated rate is 3.1 per 10^6 base leading on average to 90 amino-acid substitutions in a given tumor [31]. Virtually all biochemical, biological and clinical attributes are heterogeneous within cancer of the same histological subtype [33]. We have thus sought for alternate mechanisms contributing to cancer cell heterogeneity. We recently showed that cancer cell mRNA sequences contain more base substitutions than that of normal cells [34]. Cancer mRNA base substitution occurs at sites that are 10^4 more commonly encountered than those bearing somatic mutations and do not correspond to single nucleotide polymorphisms (SNP). Thus the differences in mRNA heterogeneity isolated from normal and cancer cells from the same patient can not be explained by differences occurring at the genomic level. Base substitution in cancer mRNA is determined by the composition of DNA context that corresponds to the portion melted by active RNA Polymerase II (Pol II) [34,35]. The substituted base is most frequently identical to that immediately preceding or following the event. *In vitro* data demonstrated forward slipping of Pol II in specific DNA contexts [36,37], and we have therefore proposed that transcription infidelity (TI) explains that a fraction of cancer mRNA are not faithful copies of genomic DNA.

We have expanded this analysis to whole genome and all available human transcripts and confirm that mRNA base substitution is significantly increased (2.5-fold) in cancer. Most importantly, we discovered that single base omission in cancer mRNA is much more dramatically (38-fold) increased. Gaps in mRNA cause the loss of downstream genomic information and can lead to aberrant proteins that might trigger immunological response. We have sought and found, in cancer patients, specific IgG directed against predicted aberrant peptides (PAP) translated from cancer mRNA containing a single base gap. Detection of low abundance diversified IgG provides a novel method for diagnosis of most common forms of human solid tumors. A panel of IgG effectively discriminated patients with non small cell lung cancer (NSCLC) from subjects without cancer.

The present invention thus shows such gaps (and insertions) are dramatically increased in cancer patients and create aberrant but predictable immunogenic proteins which represent very efficient biomarkers.

SUMMARY OF THE INVENTION

An object of this invention relates to a method for detecting the presence, risk or stage of development of a cancer in a subject, the method comprising contacting in vitro a sample from the subject with a polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, wherein the formation of a complex between said polypeptide and an antibody (TIAB) or TCR-bearing cell present in said sample is an indication of the presence, risk or stage of development of a cancer.

A further object of this invention relates to a method of assessing the physiological status of a subject, the method comprising a step of measuring the presence or level of antibodies specific for aberrant protein domains created by transcription infidelity (TIAB) or of TCR-bearing immune cells that bind to such domains in a sample from the subject, wherein a modified level of said TIAB or immune cells in said sample as compared to a reference value is an indication of a physiological disorder.

A further object of the invention is a method of determining the efficacy of a treatment of a cancer, the method comprising (i) determining the level of at least one polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity or the level of TIAB or corresponding TCR-bearing cells, in a sample from the subject and (ii) comparing said level to the level in a sample from said subject taken prior to or at an earlier stage of the treatment.

An other object of the invention is a method of monitoring the progression or the extension of a cancer in a subject, said method comprising (i) contacting a sample obtained from said subject with at least one polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, (ii) determining the level of

TIAB or corresponding TCR-bearing cells in said sample and (iii) comparing said level to reference level. The reference value may be a mean or median value determined from individuals not having a cancer or disease, a reference level obtained from a control patient, a reference level obtained from the subject before cancer onset or with a control polypeptide.

An other object of the invention relates to a method of determining whether an individual is making a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334, comprising contacting a sample obtained from said individual with an agent indicative of the presence of said polypeptide and determining whether said agent binds to said sample.

A further object of this invention is a method of selecting, characterizing, screening or optimizing a biologically active compound, said method comprising placing *in vitro* a test compound in contact with a gene and determining the ability of said test compound to modulate the production, from said gene, of RNA molecules containing transcription infidelity gaps and insertions.

A further object of this invention resides in a method of producing a peptide specific for transcription infidelity, the method comprising :

a) identifying a protein domain resulting from a transcription infidelity gap or insertion;

b) synthesizing a peptide comprising the sequence of said protein domain of a); and

c) optionally verifying, in a biological sample from a mammalian subject, that the peptide binds an antibody.

The invention also relates to any polypeptide comprising the sequence of an aberrant protein domain created by gap or insertion transcription infidelity, or an epitope-containing fragment thereof, especially a polypeptide comprising a sequence selected from SEQ ID NOs: 1 to 3334, or an epitope-containing fragment thereof.

A further object of the invention is an isolated nucleic acid encoding a polypeptide described above or comprising a first nucleotide sequence encoding a

polypeptide selected from the group consisting of SEQ ID NOs: 1-3334 or a sequence complementary thereto and a second nucleotide sequence of 100 or less nucleotides in length, wherein said second nucleotide sequence is adjacent to said first nucleotide sequence in a naturally occurring nucleic acid.

An other object of this invention is a cloning or expression vector comprising a polynucleotide described above and the host cell transformed or transfected with this vector

A further object of this invention is an isolated antibody or portion of an antibody which specifically binds to any polypeptide comprising the sequence of an aberrant protein domain created by gap or insertion transcription infidelity and, particularly, to a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof.

A further object of this invention is an immune cell comprising a TCR specific for any polypeptide comprising the sequence of an aberrant protein domain created by gap or insertion transcription infidelity and, particularly, for a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334 .

The invention also relates to a solid support comprising at least one polypeptide comprising the sequence of an aberrant protein domain created by gap or insertion transcription infidelity, and, particularly, at least one polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof.

The invention further relates to a device or product comprising, immobilized on a support, at least one polypeptide comprising the sequence of an aberrant protein domain created by gap or insertion transcription infidelity, and, particularly, at least one polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof.

An other object of this invention is a kit comprising at least a device or product as defined above and a reagent to perform an immune reaction.

The invention further relates to a method of modulating an immune response in a subject, the method comprising treating the subject to deplete immune cells expressing a TCR specific for a polypeptide as defined above. Such immune cells typically include B cells, dendritic cells or T cells. Depletion may be accomplished by methods known in the art, such as ex vivo depletion using specific ligands.

LEGEND TO THE FIGURES

Fig. 1 Representation of K gaps located within ORF.

Shows for all statistically significant K gaps within ORF the % of deviation measured in the cancer set Y axis and normal set X axis. Red diamonds indicate the 45 K gaps selected for biological evaluation (table 5). Insert shows the number of transcripts affected by the indicated number of statistically significant K gaps located within the ORF.

Fig. 2. Aberrant mRNA detection of a deletion predicted on Cofilin gene in a lung cancer patient c-DNA library.

Fig.2a. Bioinformatics prediction and characteristics of selected gap.
Fig.2b. Cloning strategy.
Fig.2c. qPCR on normal and variant.
Fig.2d. cDNA variant sequence.
Fig.2e. Genomic DNA sequence.

Fig. 3. Detection of IgG recognizing 15 PAP (peptides 1-15 from table 5) in plasma of control and cancer patients bearing the indicated forms of solid tumors.

Fig.3a. Fluorescence intensity signal recorded for each individual sample incubated with biotinylated PAP were subtracted from that recorded in blank streptavidin coated wells. Intensities of this differences are shown in light blue [36] if it corresponds to the

lower half of values recorded in controls. Dark blue ■ corresponds to signals in the highest half of control values. Above controls, but lower half of positive signal are in light red ■. Dark red ■ is for patients in highest half of positive signals. Blank cells show tests that could not be performed due to sample shortage. Controls are 26 healthy individuals, patients with indicated various forms of cancer are ranked from top (early) to bottom (advanced) for each cancer according to staging. Significant p-values of Wilcoxon tests are shown at the bottom of the figure.

Fig.3b. Detection of IgG directed 13 predicted aberrant peptides in plasma of controls and cancer patients bearing the indicated forms of solid tumor.

Fluorescence intensity signal recorded for each individual sample incubated with biotinylated PAP were subtracted from the mean value recorded with PAP 14 and 15 or solely PAP 15 when the information was not available for PAP 14. The raw data are the same as that of figure 3A. Intensities of calculated signal is shown in light blue if it correspond to the lower half of value recorded in controls. Dark blue corresponds to signal in the highest half of control. Above control but lower half of positive signal are in light red. Upper half positive signals are in dark red.

Fig 4. Detection of IgG directed against PAP in sera of NSCLC versus controls without cancer.

Fig 4A. IgG directed against 37 PAP are measured (37 first peptides from table 5). N terminally biotinylated peptide with AA sequence corresponding to 37 PAP were produced and used as baits in streptavidin wells to bind putative immunoglobulins. Samples were tested at 1/100 dilution, after washing IgG bound to PAP were revealed by secondary anti-human IgG Fc domain. Control patients include 25 healthy individuals and 12 patients with COPD (light and dark blue panels) and 49 NSCLC (Study II, Table 4)(red panel). P-values of Wilcoxon tests of cases versus all controls are shown.

Fig 4B. Statistical analysis between 37 controls and 49 lung cancers. Non parametric Wilcoxon test p-values are given for each TIAB.

Fig 5. Lack of detection of IgG directed against canonical peptides (CP)

Fluorescence intensity is measured for IgG directed against peptides corresponding to canonical reading of the genome, i.e., peptides corresponding to translation of mRNA corresponding to RefSeq without gap. Canonical peptide sequences are given in table 6.

Fig 5A. IgG directed against PAP 1 and canonical peptide chosen on the same gene (I_{64} to Q_{93}) are measured. IgG directed against a canonical peptide from albumin (between Q_{128} and P_{143}) are also measured. All patients from figure 4 are included.

Fig 5B. Schematic representation of canonical and predicted aberrant peptide sequences.

Fig 5C. IgG directed against PAP 7, 24 and 28 and their canonical peptides are shown for 11 controls (blue panel) and 16 NSCLC (red panel).

Fig 6. Detection of IgG directed against PAP in sera of *Mus musculus*.

Fig 6A. Bioinformatics analysis of homology between *Homo sapiens* and *Mus musculus* sequences. mRNA and PAP alignments are given for PAP 7, 48 and 62 and show that these sequences are conserved. Protein sequence of negative control CP 7 is also conserved between human and mouse. PAP 2 and 9 are discriminant between controls and NSCLC in human but are not conserved in mouse.

Fig 6B. 12 normal mice (C57Bl/6) were injected in sub-cutaneous with $5*10^5$ LLC1 cells. The day of injection and 1-2-3 weeks after, TIAB directed against PAP 48, 62, 7 and corresponding CP titers were measured. TIAB directed against PAP 2 and 9 titers were measured in 4 mice the day of injection and 3 weeks after. Mean +- SEM are shown.

Fig 7. Combination of IgG titers directed against PAP in sera of lung cancers versus controls without cancer.

Fig 7A. Control patients include 161 healthy individuals (blue panel) and 140 lung cancers (Study III, Table 4) including adenocarcinomas ADK (red panel), squamous (orange panel) and others (yellow panel). Support Vector Machine allows discrimination of controls versus lung cancers with 6 PAP (7, 29, 48, 66, 68, 70). Distance to hyperplane is shown ; patients showing negative values by SVM model are

classified as non cancerous. Patients showing positive values are classified as cancerous.

Fig 7B shows percentage of lung cancer patients classified as positive according to their age.

Fig 7C shows percentage of lung cancer patients classified as positive according to the histopathology of their disease.

Fig 7D shows difference of distances to SVM hyperplane between lung cancer patients that are disease free 3 years after surgery and patients that are deceased or alive with recurrent cancer.

Fig 8. Combination of IgG titers directed against PAP in sera of lung cancers versus breast cancers. Control patients include 20 healthy individuals (blue panel), 20 lung cancers (red panel) and 20 breast cancers (purple panel) (Study IV, Table 4).

Fig 8A shows a combination of PAP that discriminates lung cancers versus controls without cancer.

Fig 8B shows several combinations of PAP that discriminate lung cancers versus breast cancers.

Fig 8C shows a combination of PAP that discriminates breast cancers versus controls without cancer.

DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to novel methods and products for assessing the physiological status of a subject by measuring TIAB levels. More particularly, the invention relates to methods of assessing the presence, risk or stage of a cancer in a subject by measuring TIAB levels in a sample from the subject. The invention is also

suitable to assess the responsiveness of a subject to a treatment, to monitor the progression or the extension of a cancer as well as to screen candidate drugs.

Transcription Infidelity designates a novel mechanism by which several distinct RNA molecules are produced in a cell from a single transcript sequence. This newly identified mechanism potentially affects any gene, is non-random, and follows particular rules, as disclosed in co-pending application n° PCT/EP07/057541, herein incorporated by reference.

The present application shows that transcription infidelity can introduce gaps or insertions in RNA molecules, thereby creating a diversity of detectable aberrant protein sequences from a single gene (TI polypeptide sequences). These TI polypeptide sequences are particularly interesting since they are long enough to contain epitopes against which antibodies may be generated by mammalians. As a result, the expression of such aberrant proteins in a subject can be assessed by measuring the presence of corresponding antibodies or TCR-bearing cells in a sample from the subject.

The present invention now provides a method for predicting and/or identifying the sequence of such aberrant protein domains generated by gap or insertion transcription infidelity events from any gene, as well as methods of producing polypeptides comprising such TI sequences. The invention also discloses more than 2000 gap TI (gTI) polypeptides and more than 1000 insertions TI (iTI) polypeptides, and demonstrates, in human samples, the striking correlation between the presence of antibodies directed against these polypeptide sequences and the presence of a cancer in the subject. More specifically, increased levels of specific IgG directed against predicted aberrant peptide (PAP) are detected in sera of most (>75 %) patients with common forms of solid tumours in excess of normal subjects. All 7 of the common forms of solid tumours (colon, lung, breast, ovarian, uterus, head and neck and melanoma) cause the production of IgG directed against aberrant proteins. Increase specific IgG levels were observed in most subjects with early stage disease, i.e., negative lymph node and no metastasis.

Measuring such antibodies directed against TI polypeptides, termed TIABs (Transcriptional Infidelity AntiBody), or corresponding immune cells bearing a TCR receptor specific for such aberrant domains, therefore represents a novel approach for detecting and monitoring disorders, as well as for drug development.

TIAB

Within the context of the present invention, the term TIAB ("Transcription Infidelity AntiBody") designates an antibody that specifically binds an epitope contained in a protein sequence generated by TI, particularly by gTI or iTI. TIABs more specifically designate antibodies naturally produced by a mammalian against an epitope contained in a protein sequence generated by TI, particularly by gTI and iTI (gap and insertion Transcription Infidelity). TIABs may be of any type, including IgG, IgM, IgA, IgE, IgD, etc. An antibody is "specific" for a particular epitope or sequence when the binding of the antibody to said epitope or sequence can be reliably discriminated from non-specific binding (i.e., from binding to another antigen, particularly to the native protein not containing said domain).

In one aspect, TIAB or portion of TIAB may be attached to a solid support. The attachment maintains the TIAB in a suitable conformation to allow binding of a specific gTI or iTI polypeptide when contacted with a sample containing the same. The attachment may be covalent or non-covalent, directly to the support or through a spacer group. Various techniques have been reported in the art to immobilize an antibody on a support (polymers, ceramic, plastic, glass, silica, etc.). The support may be magnetic, such as magnetic beads, to facilitate e.g., separation.

Immune cells bearing a TCR specific for such TI polypeptides include any cells of the immune system which contain a TCR, such as e.g., T cells, such as CTL, CD4+ lymphocytes, CD8+ lymphocytes and/or Treg cells, as well as antigen-presenting cells : B cells, dendritic cells or macrophages. The term includes, in particular, any TIAB-producing immune cells. Such cells may be cultured in conventional conditions, and

expanded in vitro or ex vivo using TI polypeptides of this invention as a (co-) stimulatory factor.

TI polypeptides and their production

As will be disclosed below, the invention now discloses the sequence of various TI polypeptides and allows the prediction of TI sequences from virtually any gene.

In a first embodiment, the present invention is drawn to an isolated polypeptide comprising a gTI sequence, i.e., a sequence of an aberrant protein domain created by gap transcription infidelity. Specific examples of polypeptides of this invention comprise a sequence selected from SEQ ID NOs: 1 to 2206 (see Table 3a), or an epitope containing fragment thereof.

In a second embodiment, the present invention is drawn to an isolated polypeptide comprising an insertion TI sequence, i.e., a sequence of an aberrant protein domain created by insertion transcription infidelity. Specific examples of polypeptides of this invention comprise a sequence selected from SEQ ID NOs: 2207 – 3334 (see Table 3b), or an epitope containing fragment thereof.

The term "epitope-containing fragment" denotes any fragment containing at least 6 consecutive amino acid residues, preferably at least 8, even more preferably at least 10, most preferably at least 12, which form an immunologic epitope for antibodies or TCR-expressing cells. Such an epitope may be linear or conformational, and specific for B- or T-cells.

A TI polypeptide of this invention typically comprises between 8 and 100 amino acids, preferably between 8 and 50, more preferably between 10 and 40 amino acids. The polypeptides of this invention may be produced by any conventional technique, such as artificial polypeptide synthesis or recombinant technology.

Polypeptides of this invention may optionally comprise additional residues or functions, such as, without limitation, additional amino acid residues, chemical or biological groups, including labels, tags, stabilizer, targeting moieties, purification tags, secretory peptides, functionalizing reactive groups, etc. Such additional residues or functions may be chemically derivatized, added as an amino acid sequence region of a fusion protein, complexed with or otherwise either covalently or non-covalently attached. They may also contain natural or non-natural amino acids. The polypeptide may be in soluble form, or attached to (or complexed with or embedded in) a support, such as a matrix, a column, a bead, a plate, a membrane, a slide, a cell, a lipid, a well, etc.

In a particular embodiment, polypeptides are biotinylated to form complexes with streptavidin.

The polypeptides of this invention may be present as monomers, or as multimers. Also, they may be in linear conformation, or in particular spatial conformation. In this respect, the polypeptides may be included in particular scaffold to display specific configuration.

Polypeptides of the present invention may be used as immunogens in vaccine compositions or to produce specific antibodies. They may also by used to target drugs or other molecules (e.g., labels) to specific sites within an organism. They may also be used as specific reagents to detect or dose specific antibodies or TCR-bearing immune cells from any sample.

In this respect, a particular object of this invention resides in a device or product comprising a polypeptide as defined above attached to a solid support. The attachment is preferably a terminal attachment, thereby maintaining the polypeptide in a suitable conformation to allow binding of a specific antibody when contacted with a sample containing the same. The attachment may be covalent or non-covalent, directly to the support or through a spacer group. Various techniques have been reported in the art to immobilize a peptide on a support (polymers, ceramic, plastic, glass, silica, etc.), as

disclosed for instance in Hall et al., Mechanisms of ageing and development 128 (2007) 161. The support may be magnetic, such as magnetic beads, to facilitate e.g., separation.

The device preferably comprises a plurality of polypeptides of this invention, e.g., arrayed in a pre-defined order, so that several TIABs may be detected or measured with the same device.

The device is typically made of any solid or semi-solid support, such as a titration plate, dish, slide, wells, membrane, bead, column, etc. The support typically comprises at least two polypeptides selected from SEQ ID NO: 1 to 3334, or an epitope-containing fragment thereof, more preferably from the 45 PAP polypeptides of table 5 (included in SEQ ID NO 1-3334).

In a most preferred embodiment, the method or support of the invention uses a combination of at least 2, preferably at least 3 polypeptides comprising the sequence of a distinct PAP polypeptide of Table 5.

In a particular embodiment, the device or method uses at least one, two or three polypeptides selected from PAP 1, 2, 4, 6, 7, 24, 25, 28, 29, 44, or 48 (Table 5).

In another particular embodiment, the method or support of the invention uses a combination of distinct PAP polypeptides of Table 5 selected from:

- polypeptides PAP7, PAP66, PAP70, PAP29, PAP68 and PAP48;
- polypeptides PAP7, PAP48, PAP70 and PAP29;
- polypeptides PAP6, PAP29, PAP70 and PAP82;
- polypeptides PAP6, PAP7, PAP29, PAP48, PAP70 and PAP82 ;
- polypeptides PAP6, PAP29, PAP70 and PAP69;
- polypeptides PAP7, PAP48, PAP70, PAP74 and PAP29; or
- polypeptides PAP7, PAP29 and PAP94.

In a particular embodiment, the device comprises from 2 to 10 polypeptides.

The support may comprise additional objects or biological elements, such as control polypeptides and/or polypeptides having a different immune reactivity.

Formation of an immune complex between the polypeptide and a TIAB may be assessed by known techniques, such as by using a second labelled antibody specific for human antibodies, or by competition reactions, etc.

A further aspect of this invention resides in a kit comprising a device as disclosed above, as well as one or several reagents to perform an immune reaction, i.e. formation and detection of an immune complex.

TI polynucleotides

A further embodiment of this invention relates to a polynucleotide comprising a nucleotide sequence encoding a polypeptide as defined above or a complementary strand thereof. Particularly, this polynucleotide comprising a first nucleotide sequence encoding a polypeptide selected from the group consisting of SEQ ID NOs: 1-3334 or a sequence complementary thereto and a second nucleotide sequence of 100 or less nucleotides in length, wherein said second nucleotide sequence is adjacent to said first nucleotide sequence in a naturally occurring nucleic acid. The length of the second nucleotide sequence which is adjacent to the first nucleotide sequence may be, for example, 75, 50, 25, 10 or 0.

The polynucleotides of the present invention may be DNA or RNA, such as complementary DNA, synthetic DNA, mRNA, or analogs of these containing, for example, modified nucleotides such as 3'alkoxyribonucleotides, methylphosphanates, and the like, and peptide nucleic acids (PNAs), etc. The polynucleotide may be labelled. The polynucleotide may be produced according to techniques well-known per se in the art, such as by chemical synthetic methods, in vitro transcription, or through recombinant DNA methodologies, using sequence information contained in the present application. In particular, the polynucleotide may be produced by chemical

oligonucleotide synthesis, library screening, amplification, ligation, recombinant techniques, and combination(s) thereof.

A specific embodiment of this invention resides in a polynucleotide encoding a polypeptide comprising a sequence selected from SEQ ID 2207-3334 or an epitope-containing fragment thereof.

Polynucleotides of this invention may comprise additional nucleotide sequences, such as regulatory regions, i.e., promoters, enhancers, silencers, terminators, and the like that can be used to cause or regulate expression of a polypeptide.

Polynucleotides of this invention may be used to produce a recombinant polypeptide of this invention. They may also be used to design specific reagents such as primers, probes or antisense molecules (including antisense RNA, iRNA, aptamers, ribozymes, etc.), that specifically detect, bind or affect expression of a polynucleotide encoding a polypeptide as defined above. They may also be used as therapeutic molecules (e.g., as part of an engineered virus, such as, without limitation, an engineered adenovirus or adeno-associated virus vector in gene therapy programs) or to generate recombinant cells or genetically modified non-human animals, which are useful, for instance, in screening compound libraries for agents that modulate the activity of a polypeptide as defined above.

Within the context of this invention, a nucleic acid "probe" refers to a nucleic acid or oligonucleotide having a polynucleotide sequence which is capable of selective hybridization with a transcription infidelity domain or a complement thereof, and which is suitable for detecting the presence (or amount thereof) in a sample containing said domain or complement. Probes are preferably perfectly complementary to a transcription infidelity domain however, certain mismatch may be tolerated. Probes typically comprise single-stranded nucleic acids of between 8 to 1500 nucleotides in length, for instance between 10 and 1000, more preferably between 10 and 800, typically between 20 and 700. It should be understood that longer probes may be used as well. A preferred probe of this invention is a single stranded nucleic acid molecule of

between 8 to 400 nucleotides in length, which can specifically hybridize to a transcription infidelity domain.

The term "primer" designates a nucleic acid or oligonucleotide having a polynucleotide sequence which is capable of selective hybridization with a transcription infidelity domain or a complement thereof, or with a region of a nucleic acid that flanks a transcription infidelity domain, and which is suitable for amplifying all or a portion of said transcription infidelity domain in a sample containing said domain or complement. Typical primers of this invention are single-stranded nucleic acid molecules of about 5 to 60 nucleotides in length, more preferably of about 8 to about 50 nucleotides in length, further preferably of about 10 to 40, 35, 30 or 25 nucleotides in length. Perfect complementarity is preferred, to ensure high specificity. However, certain mismatch may be tolerated, as discussed above for probes.

Another aspect of this invention resides in a vector, such as an expression or cloning vector comprising a polynucleotide as defined above. Such vectors may be selected from plasmids, recombinant viruses, phages, episomes, artificial chromosomes, and the like. Many such vectors are commercially available and may be produced according to recombinant techniques well known in the art, such as the methods set forth in manuals such as Sambrook et al., *Molecular Cloning* (2d ed. Cold Spring Harbor Press 1989), which is hereby incorporated by reference herein in its entirety.

A further aspect of this invention resides in a host cell transformed or transfected with a polynucleotide or a vector as defined above. The host cell may be any cell that can be genetically modified and, preferably, cultivated. The cell can be eukaryotic or prokaryotic, such as a mammalian cell, an insect cell, a plant cell, a yeast, a fungus, a bacterial cell, etc. Typical examples include mammalian primary or established cells (3T3, CHO, Vero, Hela, etc.), as well as yeast cells (e.g., *Saccharomyces* species, *Kluyveromyces*, etc.) and bacteria (e.g., *E. coli*). It should be understood that the invention is not limited with respect to any particular cell type, and can be applied to all kinds of cells, following common general knowledge.

Diagnosis

The present invention allows the performance of detection or diagnostic assays that can be used, e.g., to detect the presence, absence, predisposition, risk or severity of a disease from a sample derived from a subject. In a particular embodiment, the disease is a cancer. The term "diagnostics" shall be construed as including methods of pharmacogenomics, prognostic, and so forth.

In a particular aspect, the invention relates to a method of detecting *in vitro* or *ex vivo* the presence, absence, predisposition, risk or severity of a disease in a subject, preferably a human subject, comprising placing a sample from the subject in contact with a polypeptide as defined above and determining the formation of an immune complex. Most preferably, the polypeptide is immobilized on a support. In a preferred embodiment, the method comprises contacting the sample with a device as disclosed above and determining the formation of immune complexes. Preferably, the polypeptide is selected from SEQ ID NO: 1-3334 or an epitope-containing fragment thereof, and most preferably from the 45 PAP of table 5 (included in SEQ ID NO 1-3334).

In an other aspect, the invention relates to a method of detecting *in vitro* or *ex vivo* the presence, absence, predisposition, risk or severity of a disease in a subject, preferably a human subject, comprising placing a sample from the subject in contact with a TIAB or a portion of a TIAB or a corresponding TCR-bearing cell as defined above and determining the formation of an immune complex. Most preferably, the TIAB or the corresponding TCR-bearing cell is immobilized on a support. In a preferred embodiment, the method comprises contacting the sample with a device as disclosed above and determining the formation of immune complexes. In an other preferred embodiment, the TIAB or the corresponding TCR-bearing cell are specific for a polypeptide selected from SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof, and preferably from the 45 PAP of table 5.

A particular object of this invention resides in a method of detecting the presence, absence, predisposition, risk or severity of cancers in a subject, the method comprising placing *in vitro* or *ex vivo* a sample from the subject in contact with a polypeptide as

defined above and determining the formation of an immune complex. More preferably, the polypeptide is immobilized on a support and selected from SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof, and most preferably from the 45 PAP of table 5.

Another object of the invention relates to a method of detecting *in vitro* or *ex vivo* the presence, absence, predisposition, risk or severity of a disease in a biological sample, preferably, a human biological sample, comprising placing said sample in contact with a polypeptide as defined above and determining the presence of immune cells expressing a TCR specific for such a polypeptide. Preferably, the polypeptide is selected from SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof, and most preferably from the 45 PAP of table 5.

A further aspect of this invention resides in a method of assessing *in vitro* or *ex vivo* the level of transcription infidelity in a subject, preferably, a human subject, comprising placing a sample from the subject in contact with a polypeptide as defined above and determining the formation of an immune complex. Most preferably, the polypeptide is immobilized on a support. In a preferred embodiment, the method comprises contacting the sample with a device as disclosed above and determining the formation of immune complexes.

A further aspect of this invention resides in a method of assessing *in vitro* or *ex vivo* the level of transcription infidelity in a subject, preferably, a human subject, comprising placing a sample from the subject in contact with a polypeptide as defined above and determining the presence of immune cells expressing a TCR specific for such a polypeptide.

Another embodiment of this invention is directed to a method of determining the efficacy of a treatment of a cancer, the method comprising (i) determining the level of at least one polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity or the level of TIAB or corresponding TCR-bearing cells, in a sample from the subject and (ii) comparing said level to the level in a sample from said subject taken prior to or at an earlier stage of the treatment. Preferably, polypeptide(s)

is(are) selected from SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof, and more preferably from the 45 PAP of table 5.

A further aspect of this invention is directed to a method of determining whether an individual is making a polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, and particularly comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof, said method comprising contacting a sample obtained from said individual with an agent indicative of the presence of said polypeptide and determining whether said agent binds to said sample.

In a first embodiment, the sample obtained from the subject is placed in contact with a polypeptide which binds to antibodies specific for a polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, and particularly, comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof.

In another embodiment, the sample obtained from the subject is placed in contact with a polypeptide which binds immune cell comprising a TCR specific for a polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, and particularly, comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof.

In another embodiment, the sample obtained from the subject is placed in contact with an antibody or portion thereof which is specific for a polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, and particularly, comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof.

In another embodiment, the sample obtained from the subject is placed in contact with immune cells comprising TCR specific for a polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, and particularly,

comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334 or an epitope-containing fragment thereof.

This invention further relates to a method of monitoring the progression or the extension of a cancer in a subject, said method comprising (i) contacting a sample obtained from said subject with at least one polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, (ii) determining the level of TIAB or corresponding TCR-bearing cells in said sample and (iii) comparing said level to reference level, said reference level being a mean or median value from subjects not having a cancer or control value from the subject before cancer onset. Preferably the polypeptide comprises the sequence of PAP 12, for which titers of antibodies are significantly increased in non operable patients versus operable ones and thus provide indication related to disease extension.

The presence (or increase) in TIAB or corresponding TCR-bearing immune cells in a sample is indicative of the presence, predisposition or stage of progression of a cancer disease. Therefore, the invention allows the design of appropriate therapeutic intervention, which is more effective and customized. Also, this determination at the pre-symptomatic level allows a preventive regimen to be applied.

The diagnostic methods of the present invention can be performed *in vitro*, *ex vivo* or *in vivo*, preferably *in vitro* or *ex vivo*. The sample may be any biological sample derived from a subject, which contains antibodies or immune cells, as appropriate. Examples of such samples include body fluids, tissues, cell samples, organs, biopsies, etc. Most preferred samples are blood, plasma, serum, saliva, seminal fluid, and the like. The sample may be treated prior to performing the method, in order to render or improve availability of antibodies for testing. Treatments may include, for instance one or more of the following: cell lysis (e.g., mechanical, physical, chemical, etc.), centrifugation, extraction, column chromatography, and the like.

In a preferred embodiment, the test is performed on serum or plasma.

Furthermore, in a most preferred embodiment, the sample is treated as disclosed in EP08 305 293.6 prior to testing. Indeed, the applicant has shown that optimal testing conditions are different when the samples have been kept fresh frozen at -20°C or at -80°C. More preferably, the sample to be tested is subjected to a chemical or physical treatment suitable to unveil antibody binding site, change the conformation of the binding site, or unmask the antibody binding site. More preferably, the treatment comprises heating the sample at a temperature of at least 36° C for a period of time sufficient to activate the antibody. Preferred temperatures are comprised between 36° C and 70° C, preferably between 36° C and 60° C.

Determination of the presence, absence, or relative abundance of a TIAB or specific immune cell in a sample can be performed by a variety of techniques known per se in the art. Such techniques include, without limitation, methods for detecting an immune complex such as, without limitation, ELISA, radio-immunoassays (RIA), fluoro-immunoassays, microarray, microchip, dot-blot, western blot, EIA, IEMA, IRMA or IFMA (see also Immunoassays, a practical approach, Edited by JP Gosling, Oxford University Press). In a particular embodiment, the method comprises contacting the sample and polypeptide(s) under conditions allowing formation of an immune complex and revealing said formation using a second labelled reagent.

In a typical embodiment, the method comprises comparing the measured level of TIAB or immune cells to a reference level, wherein a difference is indicative of a dysfunction in the subject, e.g., a cancer. A change is typically a 10%, 20%, 30%, 40%, 50% or more variation as compared to the reference value. More particularly, the change in the level as compared to the reference value is an increase, which is indicative of the presence of a cancer.

The reference value may be a mean or median value determined from individuals not having a cancer or disease, a reference level obtained from a control patient, a reference level obtained from the subject before cancer onset or with a control polypeptide.

In a preferred embodiment, a change (e.g., an increase) in the level of TIAB or immune cells in said sample as compared to the reference level is indicative of the presence, risk or stage of development of a cancer.

Contacting may be performed in any suitable device, such as a plate, microtitration dish, test tube, wells, glass, column, and so forth. In specific embodiments, the contacting is performed on a substrate coated with the polypeptide. The substrate may be a solid or semi-solid substrate such as any suitable support comprising glass, plastic, nylon, paper, metal, polymers and the like. The substrate may be of various forms and sizes, such as a slide, a membrane, a bead, a column, a gel, etc. The contacting may be made under any condition suitable for a detectable antibody-antigen complex to be formed between the polypeptide and antibodies of the sample.

In a specific embodiment, the method comprises contacting a sample from the subject with (a support coated with) a plurality of polypeptides as described above, and determining the presence of immune complexes.

In a particular embodiment, the method comprises contacting the sample with a plurality of sets of beads, each set of beads being coated with a distinct polypeptide as defined above.

In an other particular embodiment, the method comprises contacting the sample with a slide or membrane on which several polypeptides as defined above are arrayed.

In an other particular embodiment, the method comprises contacting the sample with a multi-wells titration plate, wherein at least part of the wells are coated with distinct polypeptides as defined above.

The invention may be used for determining the presence, risk or stage of any cancer in a subject. This includes solid tumors, such as, without limitation, colon, lung, breast, ovarian, uterus, liver, or head and neck cancers, as well as melanoma, brain tumors, etc. The invention may also be used for liquid tumors, such as leukemia. The invention may be used in a first screening, to detect a cancer, even at early stages thereof, in a subject having a risk of developing such a disease. In a second screen, the invention may be used to more precisely identify the type of cancer, depending on the

polypeptides used for detection. In this respect, as disclosed in Figure 3, polypeptides comprising the sequence of PAP 1, 2, 3, 4 or 7 (Table 5), or an epitope-containing fragment thereof, allow the identification of patients with various types of cancers.

In a particular embodiment, the invention is used to determine the presence, risk or stage of a lung cancer and the polypeptide comprises a sequence selected from PAP 1, 2, 4, 6, 7, 24, 25, 28, 29, 44 and 48 (Table 5) (Fig 4) or an epitope-containing fragment thereof.

Drug screening

The invention also allows the design (or screening) of novel drugs by assessing the ability of a candidate molecule to modulate TIAB levels or corresponding immune cells.

A particular object of this invention resides in a method of selecting, characterizing, screening or optimizing a biologically active compound, said method comprising determining whether a test compound modulates TIAB levels. Modulation of TIAB levels can be assessed with respect to a particular protein, or with respect to a pre-defined set of proteins, or globally.

A further embodiment of the present invention resides in a method of selecting, characterizing, screening or optimizing a biologically active compound, said method comprising placing *in vitro* a test compound in contact with a gene and determining the ability of said test compound to modulate the production, from said gene, of RNA molecules containing transcription infidelity gaps and insertions.

A further embodiment of the present invention resides in a method of selecting, characterizing, screening or optimizing a biologically active compound, said method comprising placing *in vitro* a test compound in contact with an immune cell expressing a TCR receptor specific for a polypeptide as defined above, and determining the ability of said test compound to modulate the activity or growth of said cell.

The above screening assays may be performed in any suitable device, such as plates, tubes, dishes, flasks, etc. Typically, the assay is performed in multi-well microtiter dishes. Using the present invention, several test compounds can be assayed in parallel. Furthermore, the test compound may be of various origin, nature and composition. It may be any organic or inorganic substance, such as a lipid, peptide, polypeptide, nucleic acid, small molecule, in isolated or in mixture with other substances. The compounds may be all or part of a combinatorial library of compounds, for instance.

Further aspects and advantages of the present invention will be disclosed in the following experimental section, which should be considered as illustrative an not limiting the scope of this application.

EXAMPLES

A- MATERIALS AND METHODS

Plasma samples

Blood samples were drawn from normal human subjects (n=26) attending local university hospital for the purpose of biological testing not related to cancer (Nancy University Hospital). Clinical records were reviewed by a trained physician who ascertained that these subjects were free from acute disease, not suspected of active cancer, allergic or autoimmune conditions. This group includes patients with cancer risk factors, *e.g.*, smoking and obesity, one of the controls had uterus cancer surgically removed 10 years prior to blood sampling and one was pregnant at time of blood sampling. Patients with chronic obstructive pulmonary disease (n=12) were either recruited in the same department (n=6) or recruited in the nuclear medicine department of the same hospital (n=6). All patients with COPD were free from exacerbation episodes at the time of blood sampling. Patients with various forms of solid tumors (n=46) were sampled at the time of PET-CT cancer extension evaluation before treatment and staging was completed by analysis of pathology samples (Nancy University Hospital). Patients with active NSCLC (n= 49) were recruited in Strasbourg

University Hospital and were part of a lung cancer longitudinal study. Blood samples were drawn at the time of staging. All patients attending these medical research facilities agreed to have their samples anonymously tested for research purposes by signing consent forms. Collection and analysis of these samples were declared to the French ministry of Health and to the Ministry of Research in accordance with French laws.

Bioinformatic procedures

The analysis was performed as previously described, but with the following modifications [34]. Each EST was retrieved and assigned to either cancer or normal set of sequences using the tissue source available in database. Each sequence was then aligned once using MegaBlast 2.2.16 [38] against human RNA RefSeq from NCBI. The single best alignment score was retained. Each EST that did not align on more than 70% of its length was not taken into account. Positions with single base sequence variations were taken into account only if 10 bases upstream and the 10 downstream were a perfect match to RefSeq. The first and last 50 bases at each alignment extremity were deleted. Gaps and insertions were located on the last nucleotide of any n-uplet if need be.

Biochemical analysis

N-terminal biotinylated peptides with aa sequence defined by *in silico* translation of RefSeq taking into account the identified K gap and peptides corresponding to canonical sequence of albumin and MRPL12 gene were purchased from different manufacturers. Samples were diluted 100-fold and IgG detection was performed on ImmunoCAP100 (Phadia, Uppsala, Sweden) using commercially available reagents and following manufacturer instructions. Samples were analyzed in duplicate with a few exceptions in Figure 3 due to sample shortage. The order or testing of cancer and control as well as that of PAP was random. In absence of internal standards, results are expressed as fluorescence units (FU).

Statistical method

Testing for statistical significance of EST base composition and estimation of false positives due to multiple testing was performed as previously described [34]. Non parametric ranks comparison Wilcoxon test was used to test difference in IgG titers.

B - RESULTS

Identification of protein domains that result from TI gaps

To analyze EST heterogeneity on a genome-wide scale, we retrieved from noncurated dbEST [39] all sequences released between January 2000 and July 2007. These sequences were separated according to their normal or cancer origin, as indicated in the database. Each EST was then aligned once against all human RefSeq RNA sequences from NCBI (July 2007) [40]. We first tested for statistical significance of differences occurring at any given RefSeq position between normal and cancer matrices, and then estimated false positives due to multiple testing [41]. Positions with statistically significant sequence differences are referred to as K if the variation is in excess in cancer and conversely N when in excess in normal.

The most important observation to be drawn from the results of Table 1 is that K gaps occurred even more commonly than K base substitutions. The bioinformatic constraints defining these gaps are stringent: a given EST position with a single base gap is taken into account only if it is flanked upstream and downstream by 10 bases that are perfect matches to RefSeq. For the 2191 K gaps located within ORF, normal and cancer gapped ESTs percentages are represented on Figure 1 Strikingly, K insertions were 5 fold more common than N insertions and K gaps were ~13-fold more common than N gaps. Subtracting the estimation of statistical false positives increased the ratio to 38 because p-values of K gaps are much lower than those of N gaps. Unlike K, N gaps were few and obviously contained a large proportion of false positives (Table 1). We therefore focused further analysis on K gap positions located within the ORF.

Table 2 summarizes the entries of our analysis. It is clear that there was no obvious bias resulting from differences in the number of ESTs or transcripts represented

in cancer and normal sets. From Tables 1 and 2, one can estimate that in cancer cells mRNA sequence variations i.e. substitutions, insertions and gaps occurred at the rate of 1-2 per thousand bases. This largely exceeded the rate of somatic mutations:1-2 per megabase [31].

SBG (single base gap) in mRNA can be caused by either somatic or germinal mutation, or by RNAP omitting the reading of a single DNA base and proceeding with transcription. We also considered the hypothesis of a slipping forward or backward of the splicing machinery causing SBG to be located on the first or last exon bases. The latter mechanism was however found unlikely because 99.2 % of identified SBG were not within immediate exon-intron boundaries.

The composition of missing mRNA bases were in the following order: U (47%) > C (39%) > > G (10%) > A (4%). This distribution deviated strongly from being random (goodness-of-fit chi² test, two-tailed, $\alpha = 0.05$, $P=10^{-248}$). Also, 99%, 76%, 98% and 95% of U, C, G, and A gaps, respectively, occurred within repeats of one or more identical bases. Finally, for 97% of U gaps, G was found immediately downstream. Thus genomic DNA context is determining in part the possible occurrence of mRNA gap in cancer cells. Detailed analysis of the impact of DNA context on the occurrence of TI events will be reported elsewhere.

Aberrant mRNA detection

To verify these bioinformatics conclusions, we cloned from a lung cancer patient c-DNA library a plasmid that after qPCR and sequencing was shown with SBG occurring at the predicted position (Fig 2 A-E). We analyzed the same number of clones obtained from the same individual normal tissue and did not find any sequence variation (data not shown). Direct sequencing of genomic DNA obtained from cancer, adjacent and normal tissue of the same patient unambiguously demonstrated the lack of either somatic or germinal mutation at this position. We can not exclude that the identified mRNA gap was artificially created during the cloning process. However it is unlikely that such event would precisely coincides with the position predicted by bioinformatics.

It was thus reasonable to assume that in cancer cells a small but detectable proportion of mRNA were not faithful copy of genomic DNA.

Materials and methods

Plasmid preparation:

The pBAD plasmid (Invitrogen) was used in order to have an inductible promoter upstream of the cloned sequence. The sequence of alpha peptide amplified from the pBS-SK+ plasmid was cloned out of phase of the ATG sequence present in the CCATGG cloning site of the pBAD plasmid to produce the pBAD-Alpha plasmid. In absence of a cloned sequence, no alpha peptide is produced and the *E. coli* colony is white colored.

Insert preparation and cloning:

cDNA from cancerous lung and adjacent normal tissue obtained from the same individual (Biochain Inc.) were amplified by PCR using oligonucleotides complementary to the CFL1 gene and the high fidelity Phusion polymerase (Finnzyme) following manufacturer recommendations. cDNA were then purified on Nucleospin Extract II columns (Macherey Nagel), visualized on agarose gel and digested with the NcoI and NheI restriction enzymes (Biolabs). The products were then ligated in the pBAD-alpha plasmid digested with the same enzymes and dephosphorylated. E. coli TOP10 (Invitrogen) cells were transformed with the ligation mix and spread on LB ampicillin (100mg/L) arabinose (0.5%) X-Gal (80µg/mL) plates.

Colonies screening:

When a CFL1 sequence with no gap is cloned, the alpha peptide is not in phase with the ATG: the colony is white colored. If a CFL1 sequence with a gap is cloned, the alpha peptide is produced, the inactive β-galactosidase (present on the genome of the bacteria) is complemented and the *E. coli* colony became blue (Figure 2B).

Blue colonies had grown in LB medium supplemented with ampicillin (100 mg/L) and 1µL of culture was screened with a Real-time PCR using CFL1 specific oligonucleotides and Syber-green I (Sigma) (Figure 2C). The first oligonucleotide (green on the Figure) is specific of both sequences and shows that the number of plasmid copies is not different between the 2 samples. The second oligonucleotide is

specific of the CFL1 Reference sequence (Refseq) and shows a difference between Ct when the sequence is not identical to the Refseq (red on the Figure).

Plasmid DNA of clones that show a difference between Ct were extracted and sequenced using an oligonucleotide present on the plasmid (GATC biotech) (Figure 2C).

Sequences were aligned to the CFL1 Reference sequence.

TIAB detection

Figure 1 shows the percentage of deviations recorded for each K position in both the cancer (Y axis) and the normal set (X axis). K gaps were distributed on a limited number (532 or 1.4%) of transcripts (Fig 1 insert). These mRNA have lost their canonical reading frame and would be translated into aberrant possibly immunogenic proteins. To test this hypothesis, we selected a panel of 15 K gaps representative of the 2206 ORF gap positions (Table 5, PAP 1 to 15) and that were distributed on 8 different human chromosomes. These 15 positions resulted from 8, 5 and 2 omissions of U, C and G, respectively. We verified that AA sequences predicted to result from translation of mRNA with single base gap 1) encode AA sequence longer than 12, 2) did not match with any known human protein on more than 7 consecutive AA (Swiss-Prot [42]), 3) had no AA sequence homology with one another. We also established that selected K gaps had not been identified as cancer somatic mutations by either Sanger Institute Catalogue Of Somatic Mutations (http://www.sanger.ac.uk/cosmic) [43] nor by 2 recent large scale in depth cancer cell genome sequencing efforts that included 11 out of 15 genes involved in the current screening [31,32]. K gaps did not correspond to biologically validated or putative SNP. Finally, and according to the most recent update of the dbSNP database (September 21, 2007), there was no SNP introducing a frameshift identical to that caused by a single gap upstream of the defined position [44].

Blood samples are drawn from human subjects divided into two or more groups. All samples are residual sera. At least one group includes patients with active cancers. Clinical data relevant to all groups, including controls and active cancers are collected and ascertained by a trained physician. Data on controls may include cancer risk factors. Data on cancer patients may include staging and response to treatment. The groups are

designed to evaluate a panel of TIABs and their specificity and sensitivity for a particular diagnostic indication such as early cancer detection, identification of cancer type, prediction of disease severity and progression, or response to treatment.

Synthetic N-terminal biotinylated peptides corresponding to these predicted aberrant peptides (PAP) were produced and coated individually onto streptavidin Elia wells (Phadia, Uppsala Sweden). Sera from 46 cancer patients (Study I) and 26 control subjects (Table 4) were incubated with either blank (non peptide coated wells) or peptide coated wells. IgG bound to the wells after washing were revealed with commercial secondary antihuman IgG invariable domain antibodies generating fluorescence. In the first analysis, the intensity of fluorescence measured with any given PAP in a given subject was subtracted from that measured in the same subject using non peptide coated streptavidin well (blank). The results showed in cancer patients versus controls statistically significant increase in IgG directed against PAP 1, 2, 4, 7 (Wilcoxon test $P< 2E^{-8}$; $P< 2E^{-3}$; $P< 5E^{-2}$; $P< 3E^{-2}$ respectively) (Fig 3A). There were no statistically significant differences in the level of any of the IgG detected in young (<50 years) versus older (>50 years) normal subjects (Wilcoxon tests = NS) and no significant differences due to gender of controls (Wilcoxon tests = NS). We next tested whether detection of IgG directed against PAP allowed discrimination between cancer and control subjects. The test was considered positive (Fig 3 light and dark red) when the difference in fluorescence intensity between PAP coated wells and blank wells was higher than that of the highest value measured in the control group. Thus, specificity was arbitrarily set at 100 %. Under these conditions, all but one PAP detected at least one cancer patient; 6 out of 15 PAP identified IgG levels in excess of control in more than 10 % of patients. Considered together, 35 out of 46 patients (76 %) with 7 forms of the most common solid tumors had IgG levels above threshold defined by the control group. Figure 3 shows that colon, lung and head and neck cancer patients had a more diversified panel of positive signals and were positive for 11, 10 and 8 PAP respectively. Breast cancer patient IgG bound to only 6 PAP. This diversity further decreased in patients with cancer of ovary, skin and uterus (Fig 3).

Thus with this first panel of 15 PAP, coverage and sensitivity was optimal for lung cancer. No sensitive early stage lung cancer test exists thus diagnostic at this stage is rare, and 5 year survival only 14 % (46).

TIAB detection for early stage lung cancer diagnostic

An important implication of this invention is that early stage lung cancer diagnostic might become possible based on simple blood testing. Analysis of data required no sophisticated statistical method, thus the risk of over fitting is minimal [10]. Also, our test was not a systematic search of biomarkers, but rather hypothesis driven based on bioinformatic predictions, thus the risk of bias due to multiple testing was low [10]. However, because of the clinical implications of such finding, we sought for replication in an independent study.

Synthetic N-terminal biotinylated peptides with AA sequence selected from the 45 PAP of table 5 constituting a panel of TIAB baits are purchased from different manufacturers and coated individually onto Reacti-Bind streptavidin coated plates (Pierce Biotechnology, Rockford, Illinois). Samples are diluted 100 fold and analyzed in duplicate. Serum IgG bound to peptides are revealed with commercial secondary antihuman IgG invariable domain antibodies conjugated with enzyme, particularly phosphatase. Reaction with a fluorescence substrate is performed using commercially available reagents. Fluorescence reading is performed on FLUOstar Galaxy microplate reader (BMG Labtech, Offenburg, Germany) following manufacturer instructions.

A TIAB of the panel is selected for a particular diagnostic indication if a threshold can be established to separate at least two groups of human subjects designed for this indication. For the selected TIAB the absolute fluorescence intensity in one group is higher than the threshold and the fluorescence measured in the same way in the other group is lower than the threshold.

Selection of PAP to monitor disease progression or extension

We set out to select PAP from a panel in a group of patients representative of various stages of disease progression or extension following the method of the previous example (Selection of PAP/TIABs for a particular diagnostics indication).
We next set out to test the efficacy of this panel of PAP in a first group of 49 patients representative of the various stages of non small cell lung cancer (NSCLC) (Fig 4).

Blood samples were obtained from 25 healthy controls, 12 subjects with non cancer lung disease and, at the time of diagnosis, from 49 patients with different stages of NSCLC (Table 4, study II). These NSCLC patients were representative of the current status of this disease in France. In absence of reliable early stage testing procedures most of NSCLC were diagnosed with advanced diseases and only 20 % were at early stage. The result of fluorescence intensity (FI) recorded for the 37 first PAPs of Table 5 for each patient and control are shown as Figure 4A. Statistical significance of difference between groups was determined by Wilcoxon test, the results are indicated on each panel. It can be seen that the FI significantly increased in lung cancer patients compared to controls for 33 out of 37 PAP. P value of Wilcoxon tests ranged from 10^{-15} to 10^{-10} for 10 most discriminating PAP (fig 4B). PAP1 alone allows to perfectly discriminate controls and NSCLC (specificity = 100% and sensitivity = 100%).

Thus the hypothesis that by–products of translation of aberrant mRNA with SBG contributed to modulate humoral immune response to NSCLC appeared valid. We verified this by testing that IgG binding to PAP was specific of their AA sequence. We thus measured the levels of IgG directed against albumin peptide, peptide corresponding to canonical reading of genome on gene 1 and 3 of the most discriminating PAP to those of IgG directed toward their corresponding canonical peptides (CP). These CPs are encoded by the same genes and segment encoding PAP (7, 24, 28), but their AA sequences were those derived from a canonical reading of the human genome i.e. without frame shift. The data show that in lung cancer patients the titers of Ig directed against CPs were much lower than those directed against PAP (fig 5) and that CPs did not discriminate between cases and controls (Wilcoxon NS).

TIAB detection in mice

To extend the TIAB concept to other mammals and verify that TIAB detection is caused by cancer, we sought to transpose the observation to a mouse model. We first selected 5 PAP that effectively discriminated patients with NSCLC from controls. As shown in fig 6 A, three of these are derived from genes highly conserved at the genomic level between mice and human. Most importantly the potentially affected bases were identical in both species this is also the case for the 4 bases upstream and the 2 downstream. We have previously shown the importance of this short DNA context allowing the occurrence of TI event [34]. The 2 other selected PAP were from genes not conserved between mice and human. The gene that can lead to PAP 9 is present in human but not murine genome. In mice, the occurrence of SBG at predicted position of the gene leading to translation of PAP 2 introduced a stop codon after encoding 7 AA. An additional negative control corresponding to CP of PAP 7 was also included. Immuno-competent (C57Bl6) mice (n=12) were inoculated subcutaneously with mice Lewis Lung Cancer (LLC1) [45,46]. Ig G binding to 5 PAP and one CP were measured before LLC1 transplantation and at weekly interval for up to 21 days. At this time, average tumor sizes were 3.15 +/- 0.4 cm^3. As shown in Figure 6B, Ig G against PAP 7, PAP 48 and PAP 62 increased significantly 2 weeks after tumor implantation. P values of paired t-Test were $1*10^{-4}$, $3*10^{-4}$ $9*10^{-4}$ for PAP 7, PAP 48 and PAP 62 respectively. We did not observe significant increase of the level of IgG directed against CP7.

Materials and methods

Cell culture

The murine Lewis Lung carcinoma cell line (LLC1) was obtained from American Type Culture Collection (ATCC). The cells were cultured in 75 cm^2 flask containing RPMI 1640 medium (Invitrogen, France) supplemented with 10% FBS, streptomycin (0.1 mg/ml) and penicillin (100 units/ml) and maintained at 37°C in humidified atmosphere containing 5% CO2 in air.

Tumor transplantation

LLC1 tumor cells ($5*10^5$ cells in a 0.1 ml final volume of RPMI 1640) were injected subcutaneously (s.c.) in the right hindquarters area of 7 weeks C57bl/6 female

mice (Janvier, France). 21 days after s.c. injection of LLC1 cells, the tumor volumes were measured by measuring bisecting diameters of each tumor and calculating using the formula $V= a^2 * b * 0.5236$ with "a" as the larger diameter and "b" as the smaller diameter. Before s.c. injection of the tumor cells, a 100 µl of sample blood was taken under isoflurane anaesthesia as the T0. Once a week, a sample of 100 µl blood was taken in EDTA tube under isoflurane anaesthesia as the T 7, 14 and 21 days.

Clinical validation

To validate the fact that a combination of PAP of the invention can lead to robust lung cancer diagnosis, we conducted large scale retrospective case control study that included 161 control subjects that were healthy blood donor with age ranging from 18 to 65 years old and that did not excluded smokers. The patients were 140 individuals with early stage non small cell lung cancer. Blood from these patients was collected at time of diagnosis. All patients in this group matched surgical intervention criteria and were thus early stage for the large majority. All patients in this group underwent surgery thus postoperative staging and pathological classification was obtained for all patients. It must be emphasized that patients in this group did not receive pre-operative chemotherapy or radiotherapy. The clinical characteristics of controls and patients are shown as table 4 study III. Patients and controls were tested for TIAB directed against 6 PAP measured under specific experimental conditions. Statistical analysis of the diagnostic value of this combination of markers was determined using support vector machine. SVM was retained after analyzing the performance of alternative classification methods. SVM defines a 6 dimensions hyperplane and provides a measure of individual distance of controls and patients to this hyperplane. The data are therefore presented as the relative distance to the hyperplane for each subject. It can be seen that the 2 populations of cancer patients and controls are well separated (Fig 7A). The overall test performance was 86 % sensitivity and 97 % specificity. Only 5 controls are on the wrong side of hyperplane. And 19 patients with lung cancer are on the wrong side of hyperplane. After iterative cross validation sensitivity and specificity were 82% and 95% respectively. Sensitivity is not different between younger or elder patients (Fig 7B). It can be seen from examination of the data that sensitivity of current test is lower

for patients with adenocarcinomas that are diagnosed with 76 % sensitivity in contrast test performance are > 90 % sensitivity for the other classes of non small cell carcinoma. These difference in test performance were statistically significant (Fig 7C).

The benefit of surgery for lung cancer is well established. This benefit does not translate only in term of increased life expectancy but into definitive cure that is ascertained by the analysis of number of diagnosis and number of death. In France these numbers are ≈33000 new diagnosis and ≈28000 deaths. Thus it can be firmly ascertained that 5000 patients are cured from lung cancer i.e 15 %. All lung cancer survivors undergo surgical procedure but not all patients with lung cancer undergoing surgery are cured. Currently there is no procedure able to distinguish individuals that will benefit from alone surgery from those that will not. We thus sought to evaluate the performance of this test based on 6 PAP with respect to prediction of severity at the time of diagnosis. To achieve this we subdivided the studied population into 2 groups. In the first group are patients for whom we had documented evidence of disease free survival longer than 36 months. In the second group, patients were either deceased or with recurrence of the disease occurring within 36 months post surgery. We did not include patients for whom follow up was shorter than 36 months simply because the follow up time of these individuals was to short to ascertain outcome. We then compared the distance to the hyperplane for these 2 groups of patients that were of similar size. It is clear from the data of Figure 7D that patients that strongly benefited from surgery and were disease free at 3 years had a distance to hyperplane significantly (P= 0.005) longer than those that benefited less from surgery. This finding bears 2 immediate applications. First, it is likely that this test will identify patients with early stage lung cancer and that will most favourably benefit from surgery. Second appropriate alternative therapeutic intervention should be set in place for patients diagnosed with positive test but with a low distance to hyperplane. Such alternative measure may be conventional radio or chemotherapy. However our current interpretation of the data is that patients with low immunological response to the presence of a lung cancer are less likely to benefit from surgery alone. In this perspective it might be useful to pharmacologically boost their immunological response prior or shortly after surgery. We have therefore exemplified here the importance of PAP discovery as guide for future innovative therapeutic strategy.

We next sought to develop test that are lung cancer specific by testing various combinations of PAP under specific experimental condition. A novel combination of PAP achieved 100 % specificity and 90 % sensitivity for lung cancer versus controls. Most importantly when applied to patients with breast cancer only 3 out of 20 patients with breast cancer showed positive test. Thus in the comparison of lung versus breast cancer, 10 and 15% of patients are misclassified respectively. We considered that this rate of error will lead to unnecessary and costly downstream diagnosis procedure. The follow up diagnosis for lung cancer is chest CT scan while that of breast cancer is mammography or echography. Pet scan is useful for evaluation of disease extention. We thus developed specific tests able to more efficiently distinguish lung from breast cancer. Four combinations of PAP were found to achieve this objective. These combinations of markers are exemplified in Figure 8B. In all 4 cases lung cancer patients are distinguished from breast cancer patients with one to three patients in overlap. The clinical significance of these combinations of markers are presented because their predictive value with respect to the severity of the disease requires further evaluation. We indeed predict that, similar to what has been exemplified for lung, a specific combination of PAP will reveal in large scale study predictive of clinical outcome for breast cancer. Before this can be achieved a specific breast cancer test is needed. We currently have identified a combination of 3 markers that under specific conditions identifies breast cancer from control with 60 % sensitivity and 95 % specificity (fig 8C). Our current view is that it will be possible to identify a combination of PAP that will indicate the presence of most common cancers at early stage from a simple blood test. Secondary combinations of PAP will provide accurate indication of the precise localisation of the disease thereby allowing the selection of adequate secondary diagnostic procedure e.g CT scan, mammography, ultrasonography, fibroscopy, endoscopy, biopsies. A third line of PAP testing will provide prognostic for each individual response to surgical treatment and therefore indication as to the need of additional therapeutic measures. A fourth line of PAP will provide tools for monitoring of disease recurrence and/ or its favourable response to treatment.

C – DISCUSSION

We have identified a novel and predictable source of human cancer cell protein heterogeneity that triggers weak but diversified production of IgG. Accurate detection of these low titers specific IgG creates promising opportunities for early stage cancer diagnostic and can provide information regarding disease extension. This discovery stems from convergence of bioinformatic predictions with immunological detection of specific IgG directed toward aberrant peptides. Proteins containing PAP sequence can not be translated from normal human mRNA, but solely from mRNA that have lost their canonical genomic information due to single base gap.

Our data support the conclusion that predicted mRNAs with single gap are present in cancer cells and at least partially translated. Occurrence of identified EST gaps is too common to be generated by cancer somatic mutations [31,32]. Current estimates of cancer somatic mutation rate lead to a prediction of 12 deletions out of which 4 would be single gaps. Instead, we observed 2206 statistically significant events. Also, none of the selected gaps corresponded to either putative or biologically validated SNPs [44]. Thus, it is unlikely that mRNA gaps arose at the genomic level and must thus occur downstream, *i.e.*, during or shortly after transcription.

It has been established that pre and mature mRNA bases can be modified by enzymes, but no known human mRNA editing enzymes have been shown to remove a single base from single stranded human RNA [47]. It has been shown that Trypanosoma mitochondrial mRNA editosome is capable of U specific deletion [48]. However, no homologs of Trypanosoma and Leishmania editosome proteins were found in the human genome. Further, this editing mechanism is U specific and cannot explain the observed 53% of human cancer non-U gaps. We have considered the possibility that slipping forward or backwards of splicing machinery could introduce single base gaps that would affect either exon's last or first base [49]. None of the tested gaps were located on such positions. Moreover, the latter mechanism was found unlikely because 99.2 % of more than 2000 identified SBG were not within immediate exon-intron boundaries. Finally, it is clear that a short DNA context exerts a strong influence on the occurrence of cancer EST gaps similar to what was demonstrated for EST base substitutions [34,36,37]. We therefore currently hypothesize that skipping the incorporation of a single base by Pol II, *i.e.*, TI is causing the occurrence of gapped mRNA.

We have further exemplified that mRNA with simple gap occurring at predicted position occur in cancer cell in absence of somatic or germinal mutation. Thus, the bioinformatic concept of transcription infidelity is biologically validated in human.

The second most important consequence of the finding reported here is that a canonical reading of the human genome is insufficient to explain cancer cell heterogeneity. We therefore propose that transcription infidelity increases in cancer cells. The evidence supporting this proposal are as follows. First, we have previously shown by DHPLC that cancer cells mRNAs are more heterogeneous than those isolated from normal cells[34]. Increased sequence variations in cancer versus normal mRNA is confirmed by independent studies relying on SAGE experiments. Second, analysis of all available human mRNA derived sequences showed statistically significant increase in base substitutions, insertions and gaps (SBG) in cancer relative to normal libraries. The occurrence of these events is 10^3 more common than that of cancer somatic mutations. If present at the DNA level this rate of mutations would most likely be lethal. It is thus reasonable to assume that these variations occurred during or shortly after transcription and affect only pre and mature mRNA i.e. transient molecules. Third, there are currently no known molecular mechanisms other than TI that can either remove or add single base from mRNA and then reassemble the sequence. Thus, direct observation of predicted SBG occurring in human lung cancer cells in absence of mutation at the DNA level indicated that RNAP can skip the reading of a single DNA base and nevertheless proceed. Finally, studies from other groups showed that TI occurs *in vivo* even in absence of cancer. Specifically, in Brattleboro rat GA deletion occurring within GAGAG sequence reverts vasopressin transcript to normal thereby suppressing diabetes insipidus. Transcription frame shift affecting repetitive A sequence of β amyloid and ubiquitin B yield proteins that are detected by immunological staining of Alzheimer disease plaque.

We currently favor the hypothesis that increased cancer mRNA heterogeneity is a consequence rather than a cause of carcinogenesis. Indeed, we are detecting specific IgG directed against our current PAP panel in sera of children that developed anaplastic large cell lymphoma and that carry anaplastic lymphoma kinase (ALK) translocation (G Delsol and B Bihain, unpublished results) [50]. Rodent studies have demonstrated that

translocation causing constitutive expression of this kinase is a primary oncogenic event that alone is sufficient to cause transformation [51]. Thus, detection of positive signals that reflect the production of abnormal mRNA encoding functionally non related genes -that are not part of this specific translocation- suggests that the phenomenon occurs as a consequence of this oncogenic lesion. However, it is possible that TI contributes to accelerate carcinogenesis. Indeed, several genes involved in the regulation of transcription, translation and DNA repair - not included in the current study because putative gene function was not part of the PAP selection process- are identified through bioinformatics with K gap. It is thus possible that we are confronted with an autocatalytic process that increases in diversity and intensity as the severity of the diseases progresses.

Bioinformatics indicated that the occurrence of SBG in mRNA is a common feature of cancers. Nevertheless, differences in IgG profiles were also found in lymphoma patients (N= 27). PAP 1 and 2 that are commonly positive in NSCLC were negative in both follicular and anaplastic lymphoma patients. This contrasted with PAP 4 and 7 that were commonly positive in anaplastic large cell lymphoma but not in follicular lymphoma. Therefore, with the diversity (> than 2000 candidates) of the available panel of PAP we propose to design tumor specific PAP panels. We have exemplified this concept by demonstrating the capacity of PAP to effectively separate patients with lung cancer from those with breast cancer.

Our conclusion that mRNA with single base gap are translated at least partially into aberrant proteins suggests that in cancer cells the nonsense-mediated mRNA decay might be defective [52]. Considering current proteomic efforts, it is surprising that such highly diversified panel of aberrant proteins has remained thus far undetected. The explanation is 2 fold. 1) Protein identification by mass spectrometry relies on matching observed with predicted spectra defined by known or putative AA sequences [53]. The AA sequences of aberrant proteins resulting from mRNA gaps are not in the current protein databases (Swiss-Prot/TrEMBL [42]) and thus can not be identified by MS/MS analysis. 2) Proteasome rapidly degrades aberrant proteins yielding potentially aberrant immunogenic peptides [54].

The notion that TI increases in cancer leads to question the current strategy of cancer biomarkers discovery and to propose novel methods. Systematic cancer proteomic approaches led to conflicting results, divergences were attributed to variations in pre-analytical conditions. This might very well be the case, but an alternate explanation must now be considered. If one accepts that cancer cell protein heterogeneity largely exceeds current estimates, it becomes possible that sample sizes were insufficient to thoroughly probe a highly diversified repertoire of protein variants. Another limitation of current proteomic is that, as previously mentioned, mass spectrometry data are currently interpreted with a canonical reading of the human genome. Thus, proteins with aberrant AA sequences may have escaped proper identification. It is therefore likely that not only methodological but conceptual changes will be needed before cancer proteomic succeeds. By considering transcription infidelity according to the present invention, more reliable and relevant biomarkers can be identified.

We have shown in mouse cancer cells 3 aberrant proteins encoded by highly conserved but functionally unrelated genes. The most abundant aberrant protein in LLC1 was that derived from Poly(A)binding protein cytoplasmic 1 (PABPC1) (PAP 62). PABPC1 normally binds to mRNA poly A and modulates the nonsense-mediated decay (NMD) pathway that degrades mRNA with premature stop. Tethering of PABPC1 downstream of premature termination codon abolish NMD. The second most abundant aberrant protein in LLC1 was encoded by vimentin gene (VIM) (PAP 48). Vimentin is a type III intermediate filament protein that forms both homo and hereopolymeric structures contributing to support cellular membranes, to keep the nucleus and organelle in defined places as well as to associate with microtubule. The third most abundant aberrant protein was that encoded by the IK gene (PAP 7). IK normal function is that of a cytokine inhibiting interferon gamma induced expression of class II major histocompatibility complex. IK is also identified as chondrosarcoma associated protein 2. The consequences of the presence in cancer cells of these variants are currently unknown. However, the possibility of strong interferences with cancer cell biology must not be excluded and their contribution to cancer cell metabolic, morphological changes as well as mRNA heterogeneity will require further

investigation. At this stage we have been able to establish that most of the PAP modulated humoral immune response to NSCL cancer in human and LLC1 in mice. We predict that production by cancer cells of these aberrant proteins might significantly alter cell function through dominant negative or positive effect. Thus, these three highly conserved genes might provide novel therapeutic targets.

Analysis of mice lung cancer model established a causal relationship between the presence of LLC1 and the detection of anti-PAP IgG. Thus anti-PAP IgGs appeared as part of a normal and timely immune response triggered by cancer. Interestingly in mice, the anti-PAP IgG levels were much higher (100 fold) than those measured in humans with lung cancer. The facts that the relative size of the lung tumor were also much greater in mice and that LLC1 were implanted ectopically in subcutaneous tissue provided possible explanations for these differences.

The present invention therefore describes a novel mechanism through which cancer modulates humoral immune response. At this stage we propose that a novel mechanism TI contributes to dramatic increase in the heterogeneity of cancer cell mRNA, part of these aberrant messages are translated into aberrant protein some of which accumulated in cancer cells and most of which modulated cancer humoral immune response. The present invention thus provides products and methods allowing to correctly differentiate patients with cancer from patients without active cancer. It is thus possible to elaborate systematic biochemical screening of at risk individuals, perform all body imaging on patients with positive tests, and increase the proportion of subjects diagnosed at early stage.

Table 1. Results of statistical testing

	K	LBE	N	LBE	K / N	(K - LBE) / (N - LBE)
Gaps	2,761	11	216	144	12.78	38.19
Gaps within ORF	2,191		162		13.52	
Substitutions	1,894	92	928	186	2.04	2.43

LBE refers to location based estimator of the false positive rate.

Table 2. Results of bioinformatics analysis

	Normal	Cancer
ESTs retrieved from NCBI	3,949,323	3,043,498
Number of transcripts with EST match	34,974	34,788
Number of transcripts with EST match	33,111	
Nucleotides analyzed	88,372,747	
Positions defined by > 70 ESTs	2,829,135	
Positions matching statistical constraints		
Substitutions	5,784	
Gaps	3,790	

Results of analysis drawn after retrieval of all available human ESTs release to noncurated public database from January 2000 to July 2007. The table also shows the number of positions matching first (effective >70) and second statistical test criteria.

Table 3. TI peptide and nucleic acid sequences

Table 3a: Nucleic and amino acid sequences of the 2206 gapTI peptides. SEQ ID NOs 1-2206 as referred to in this document represent the peptide sequences depicted in column 6 of Table 3a.

Table 3b: Nucleic and amino acid sequences of the 1128 insertion TI peptides. SEQ ID NOs 2207 – 3334 as referred to in this document represent the peptide sequences depicted in column 6 of Table 3b.

Table 4. Clinical data of control individuals and cancer subjects.

	Number	Age (Years)	Female	Male	Stage*
Healthy Control	**26**	**55 ± 18**	**13**	**13**	-
Chronic Obstructive Pulmonary Disease	**12**	**55 ± 11**	**2**	**10**	-
All Cancer STUDY I	**46**	**61 ± 11**	**26**	**20**	
Colon	9	65 ± 13	1	8	T+N0M0 → T+N+M+
Lung (7 NSCLC + 2 SCLC)	9	67 ± 7	2	7	T+N0M0 → T+N+M+
Breast	9	60 ± 11	9	0	T+N0M0 → T+N+M+
Ovarian	4	58 ± 6	4	0	T+N0M0 → T+N+M+
Uterus	5	50 ± 8	5	0	T+N0M0 → T+N+M0
Head & Neck	7	60 ± 11	3	4	T+N0M0 → T+N+M+
Melanoma	3	58 ± 14	2	1	T+N+M0 → T+N+M+
Lung Cancer STUDY II (NSCLC)	**49**	**67 ± 13**	**10**	**39**	
	10	66 ± 14	3	7	N0M0
	25	68 ± 14	4	21	N+M0
	14	67 ± 10	3	11	N+M1

- International Union Against Cancer (UICC): TNM Classification of malignant tumours. 4th ed. Hermanek P, Sobin LH, eds. Berlin, Heidelberg, New York: Springer Verlag; 1987. Revised 1992.

Study III

	n	%	Age (y)	min	max
Controls	161		42 ± 14	18	65
NSCLC	140		61 ± 11	38	86
T1-2-3-4N0M0	78	56	61	38	86
T+N+M0	43	31	69	44	86
T+N+M+	19	13	61	46	77
ADK	67	48	60	45	82
Squamous	40	33	65	48	86
Other	33	19	58	38	82

Study IV

	n	%	Age (y)	min	max
Controls	**20**		**47 ± 13**	**20**	**62**
NSCLC	**20**		**61 ± 10**	**48**	**83**
T1N0M0	**9**	**45**	**60 ± 7**	**49**	**70**
T2N0M0	**11**	**55**	**62 ± 12**	**48**	**83**
Breast	**20**		**56 ± 10**	**36**	**68**
Grade level 1	**4**	**20**	**60 ± 6**	**53**	**67**
Grade level 2	**9**	**45**	**56 ± 10**	**40**	**68**
Grade level 3	**7**	**35**	**54 ± 11**	**36**	**68**

5

Table 5. Characteristics of 45 PAP polypeptides

Gene	Accession Number	positio n	peptide length	%devN	%devC	peptide	PAP number	SEQ ID NO	PAP coordonates on SEQ ID
MRPL12	NM_002949.2	687	30	3,85	17,74	WRRWAAPWFWSSLQLGGLVFRGPGPRARSR	1	1	1-30
DECR1	NM_001359.1	257	35	9,30	41,38	NLNSFHLFKKRCYHLIVFKEKWHSLLGEVLALVKE	2	2	1-35
ALDOA	NM_184041.1	602	25	0,42	4,26	GWMGCLSAVPSTRRTELTSPSGVVC	3	3	1-25
COX4I1	NM_001861.2	549	18	2,43	9,05	RKALTKSGWPSRPRGCWT	4	4	1-18
TPI1	NM_000365.4	149	13	1,59	8,59	LPLLPISTSPGRS	5	5	1-13
ENO1	NM_001428.2	548	17	1,13	8,23	SLTWLATLKSSCQSRRS	6	6	1-17
IK	NM_006083.3	614	19	0,00	15,38	LCFKRYELRLPAKRKRKRN	7	7	1-19
LYZ	NM_000239.1	78	12	8,58	41,03	ARSLKGVSWPEL	8	8	1-12
PRR4	NM_007244.2	215	30	0,00	30,14	VIVVTKMMVLSRDHQNQEAITAILPHLLFK	9	1770	2-31
CCNB1	NM_031966.2	539	27	0,93	22,73	LILPLQAQWKHLDVPLQKKTCVRLSLM	10	10	1-27
CRABP2	NM_001878.2	537	37	2,00	18,62	STSESEWPQVEPRPKPTTGHAHRPASLPPPSHPLLLG	11	11	1-37
HSPA8	NM_153201.1	165	20	0,09	2,23	PMIRETEPLQAMSPLRTLNG	12	12	1-20
LCP1	NM_002298.2	227	16	1,31	18,89	LPKLILMAMDTSASMS	13	13	1-16
PSMD13	NM_002817.2	223	21	1,60	12,57	LPKEMVSLSFMKTLSVNLNTG	14	14	1-21
FH	NM_000143.2	141	16	0,00	21,13	FGLRTRLEWQAKIPSG	15	15	1-16
GPI	NM_000175.2	1255	16	5,56	33,04	PMASMLFTSSSTKAPR	20	44	1-16
ACO2	NM_001098.2	623	12	5,17	20,21	MLWMSWLGSPGS	21	532	10-21
NDUFB5	NM_002492.2	194	13	4,84	15,08	LSSDLLDSMTGVF	22	1083	1-13
NDUFS3	NM_004551.1	195	23	4,41	17,95	LESMWLKSCPSMSNKFRCPASMS	23	1380	1-23
NDUFAB1	NM_005003.2	366	19	2,07	13,08	WTKWRLSWPWKTNLGLKFL	24	1413	1-19
ECH1	NM_001398.2	871	30	1,59	14,61	RFPARAPWRCRAPRSTCCIPATIRWPRAST	25	581	1-30
NPM1	NM_002520.5	589	11	8,33	27,59	LEVVARFHRKK	26	1075	1-11
ECHS1	NM_004092.2	149	22	5,13	12,41	SPRVLTLSTSSQKKEGRITPWG	27	1330	1-22
CFL1	NM_005507.2	447	30	0,48	4,31	LSRCCQIRTAAMPSMMQPMRPRRARRRIWC	28	1457	1-30
MRPL3	NM_007208.2	275	30	1,76	12,38	LLEVFMERVVHGGMSIFLKKMSHSLSSWSL	29	1769	1-30
CYCS	NM_018947.4	351	12	8,78	35,77	WRIPRSTSLEQK	35	1972	1-12
B2M	NM_004048.2	193	15	2,58	10,85	AMCLGFIHPTLKLTY	37	1323	1-15
ILF2	NM_004515.2	360	19	1,92	15,81	HLKCKLKKFDRWDPIKRGQ	38	1373	1-19
BCAP31	NM_005745.6	194	27	1,54	8,72	LLCCFSAFPSFLLKDGRRFSSPGWWSC	39	1499	1-27
PHB	NM_002634.2	703	12	11,69	20,41	SSLLRATPRQLS	40	1098	1-12
PGK1	NM_000291.2	567	11	0,33	5,71	LALLTEPTAPW	41	57	1-11
UQCRC1	NM_003365.2	323	30	1,44	10,55	WSIWLSREQRIGLAVPWRRRWRAWGPILMP	44	1270	1-30
NDUFV1	NM_007103.2	657	29	2,47	11,94	LWCAGLGPTSVERRQRSSSPLRASRASPA	46	1757	1-29
ATF4	NM_001675.2	1387	11	5,50	17,22	SCPPLQIIPLV	47	864	1-11
VIM	NM_003380.2	652	30	0,00	2,49	TWPRTSCASGRNCRRRCFRERKPKTPCNLS	48	1276	2-31
PTTG1	NM_004219.2	236	11	1,82	19,77	WELSTELQKSL	49	1336	1-11
PABPC1	NM_002568.3	1293	19	1,16	4,17	VELRKRWNGRRNLSANLNR	62	1079	1-19
CDKN1A	NM_000389.2	378	30	0	6,09	LAPHLLCCRGQQRKTMWTCHCLVPLCLAQG	70	80	1-30
PRDX6	NM_004905.2	322	30	0,31	3,66	VKSPQKSYLFPSSMIGIGSLPSCWACWIQQ	66	1396	1-30
RPL13A	NM_012423.2	355	17	0,63	4,88	TTRKSGWWFLLPSRSCV	68	1810	1-17
APEX1	NM_001641.2	419	30	0,81	8,72	QRKMTKRQQERAQPCMRTPQIRKPHPVANL	74	847	1-30
MLF2	NM_005439.1	343	31	2,27	12,64	PGLPAAGCSRLELSPPLGCWECRVVSWTCLG	82	1451	1-31
TUBB	NM_178014.2	370	30	0,40	9,76	LARSLDQTTLYLVSLGQVTTGPKATTQRAP	94	2119	1-30
LAPTM4A	NM_014713.3	758	21	2,01	7,03	MLCTLPLKHLLSTFCQPMKWP	69	1864	1-21
CCT8	NM_006585.2	536	22	0,21	5,70	WYVVLQKTFEILMKSHLYFVPP	86	1699	1-22

"PAP number" is the number of the PAP as referred to in the text.

"SEQ ID NO" is the identifier in the Sequence Listing.

"PAP coordonates on SEQ ID" designates the position of the amino acid residues of the PAP polypeptide in the quoted SEQ ID.

Table 6. Five negative controls

symbol	gene name	sequence
	ALB	QHKDDNPNLPRLVRP
CP on gene 1	MRPL12	IQQLVQDIASLTLLEISDLNELLKKTLKIQ
CP 7	IK	ALLQKVRAEIASKEKEEEE
CP 28	CFL1	FVKMLPDKDCRYALYDATYETKESKKEDLV
CP 24	NDUFAB1	LDQVEIIMAMEDEFGFEIP

REFERENCES

1. Weinberg, R. in *The biology of cancer* 655-724 (Garland Science, Taylor and Francis Group, LLC, 2007).

2. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat Med* **10**, 789-99 (2004).

3. ASCO. American Society of Clinical Oncology policy statement update: genetic testing for cancer susceptibility. *J Clin Oncol* **21**, 2397-406 (2003).

4. Fackenthal, J. D. & Olopade, O. I. Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nat Rev Cancer* **7**, 937-48 (2007).

5. Guillem, J. G. et al. ASCO/SSO review of current role of risk-reducing surgery in common hereditary cancer syndromes. *J Clin Oncol* **24**, 4642-60 (2006).

6. van de Vijver, M. J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009 (2002).

7. Anderson, N. L. et al. The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol Cell Proteomics* **3**, 311-26 (2004).

8. Wulfkuhle, J. D., Liotta, L. A. & Petricoin, E. F. Proteomic applications for the early detection of cancer. *Nat Rev Cancer* **3**, 267-75 (2003).

9. Ishikawa, N. et al. ADAM8 as a novel serological and histochemical marker for lung cancer. *Clin Cancer Res* **10**, 8363-70 (2004).

10. Ransohoff, D. F. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* **4**, 309-14 (2004).

11. Ransohoff, D. F. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* **5**, 142-9 (2005).

12. Stroun, M. et al. Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* **46**, 318-22 (1989).

13. Boddy, J. L., Gal, S., Malone, P. R., Harris, A. L. & Wainscoat, J. S. Prospective study of quantitation of plasma DNA levels in the diagnosis of malignant versus benign prostate disease. *Clin Cancer Res* **11**, 1394-9 (2005).

14. Boddy, J. L. et al. The role of cell-free DNA size distribution in the management of prostate cancer. *Oncol Res* **16**, 35-41 (2006).

15. Lund, A. H. & van Lohuizen, M. Epigenetics and cancer. *Genes Dev* **18**, 2315-35 (2004).

16. Ducasse, M. & Brown, M. A. Epigenetic aberrations and cancer. *Mol Cancer* **5**, 60 (2006).

17. Goessl, C. et al. Fluorescent methylation-specific polymerase chain reaction for DNA-based detection of prostate cancer in bodily fluids. *Cancer Res* **60**, 5941-5 (2000).

18. Jeronimo, C. et al. Quantitative GSTP1 hypermethylation in bodily fluids of patients with prostate cancer. *Urology* **60**, 1131-5 (2002).

19. Reibenwein, J. et al. Promoter hypermethylation of GSTP1, AR, and 14-3-3sigma in serum of prostate cancer patients and its clinical relevance. *Prostate* **67**, 427-32 (2007).

20. Wang, Y. et al. Identification of epigenetic aberrant promoter methylation of RASSF1A in serum DNA and its clinicopathological significance in lung cancer. *Lung Cancer* **56**, 289-94 (2007).

21. Diehl, F. et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc Natl Acad Sci U S A* **102**, 16368-73 (2005).

22. Korshunova, Y. et al. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Res* (2007).

23. Bentley, D. R. Whole-genome re-sequencing. *Curr Opin Genet Dev* **16**, 545-52 (2006).

24. Meyer, M., Stenzel, U., Myles, S., Prufer, K. & Hofreiter, M. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* **35**, e97 (2007).

25. Tan, E. M. Autoantibodies as reporters identifying aberrant cellular mechanisms in tumorigenesis. *J Clin Invest* **108**, 1411-5 (2001).

26. Finn, O. J. Immune response as a biomarker for cancer detection and a lot more. *N Engl J Med* **353**, 1288-90 (2005).

27. Zinkernagel, R. M. What is missing in immunology to understand immunity? *Nat Immunol* **1**, 181-5 (2000).

28. Wang, X. et al. Autoantibody signatures in prostate cancer. *N Engl J Med* **353**, 1224-35 (2005).

29. Somers, V. A. et al. A panel of candidate tumor antigens in colorectal cancer revealed by the serological selection of a phage displayed cDNA expression library. *J Immunol* **169**, 2772-80 (2002).

30. Hardouin, J., Lasserre, J. P., Sylvius, L., Joubert-Caron, R. & Caron, M. Cancer immunomics: from serological proteome analysis to multiple affinity protein profiling. *Ann N Y Acad Sci* **1107**, 223-30 (2007).

31. Sjoblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-74 (2006).

32. Wood, L. D. et al. The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-13 (2007).

33. Nelkin B, P. D., Robinson S, Small D, Vogelstein B. (ed. Owens, A., Coffey, DS, Baylin, SB) 441-460 (Academic Press, New York, 1982).

34. Brulliard, M. et al. Nonrandom variations in human cancer ESTs indicate that mRNA heterogeneity increases during carcinogenesis. *Proc Natl Acad Sci U S A* **104**, 7522-7 (2007).

35. Armache, K. J., Kettenberger, H. & Cramer, P. The dynamic machinery of mRNA elongation. *Curr Opin Struct Biol* **15**, 197-203 (2005).

36. Kashkina, E. et al. Template misalignment in multisubunit RNA polymerases and transcription fidelity. *Mol Cell* **24**, 257-66 (2006).

37. Pomerantz, R. T., Temiakov, D., Anikin, M., Vassylyev, D. G. & McAllister, W. T. A mechanism of nucleotide misincorporation during transcription due to template-strand misalignment. *Mol Cell* **24**, 245-55 (2006).

38. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203-14 (2000).

39. Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. dbEST--database for "expressed sequence tags". *Nat Genet* **4**, 332-3 (1993).

40. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-5 (2007).

41. Dalmasso, C., Broet, P. Procédures d'estimation du false discovery rate basées sur la distribution des degrés de signification. *Journal de la Société Française de Statistiques* **146** (2005).

42. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res* **27**, 49-54 (1999).

43. Bamford, S. et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* **91**, 355-8 (2004).

44. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).

45. Sharma, S. et al. T cell-derived IL-10 promotes lung cancer growth by suppressing both T cell and APC function. *J Immunol* **163**, 5020-8 (1999).

46. Bertram, J. S. & Janik, P. Establishment of a cloned line of Lewis Lung Carcinoma cells adapted to cell culture. *Cancer Lett* **11**, 63-73 (1980).

47. Gott, J. M. & Emeson, R. B. Functions and mechanisms of RNA editing. *Annu Rev Genet* **34**, 499-531 (2000).

48. Rogers, K., Gao, G. & Simpson, L. U-specific 3' - 5' exoribonucleases involved in U-deletion RNA editing in trypanosomatid mitochondria. *J Biol Chem* (2007).

49. Lodish, H., Berk, A., Zipursky, L., Matsudaira, P., Baltimore, D., Darnell, J. in *Molecular cell biology* (ed. Freeman) 404-452 (2000).

50. Lamant, L. et al. Gene-expression profiling of systemic anaplastic large-cell lymphoma reveals differences based on ALK status and two distinct morphologic ALK+ subtypes. *Blood* **109**, 2156-64 (2007).

51. Chiarle, R. et al. NPM-ALK transgenic mice spontaneously develop T-cell lymphomas and plasma cell tumors. *Blood* **101**, 1919-27 (2003).

52. Wormington, M. Zero tolerance for nonsense: nonsense-mediated mRNA decay uses multiple degradation pathways. *Mol Cell* **12**, 536-8 (2003).

53. Biemann, K. Sequencing of peptides by tandem mass spectrometry and high-energy collision-induced dissociation. *Methods Enzymol* **193**, 455-79 (1990).

54. Glickman, M. H. & Ciechanover, A. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol Rev* **82**, 373-428 (2002).

5

CLAIMS

1. A method for detecting the presence, risk or stage of development of a cancer in a subject, the method comprising contacting in vitro a sample from the subject with a polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, wherein the formation of a complex between said polypeptide and an antibody (TIAB) or TCR-bearing cell present in said sample is an indication of the presence, risk or stage of development of a cancer.

2. The method of claim 1, wherein the level of said TIAB or immune cells in said sample is compared to a reference value, and wherein a variation as compared to the reference value is indicative of the presence, risk or stage of development of a cancer.

3. The method of claim 2, wherein said variation as compared to the reference value is 10%, 20%, 30%, 40%, 50% or more.

4. The method of any one of claims 1 to 3, wherein the polypeptide comprises a sequence selected from SEQ ID NOs: 1 to 3334 or an epitope-containing fragment thereof.

5. The method of claim 4, wherein the polypeptide comprises the sequence of any one of the 45 PAP polypeptides of table 5.

6. The method of claim 4 or 5, wherein an increase in the level of TIAB or immune cells in said sample as compared to the reference level is indicative of the presence, risk or stage of development of a cancer.

7. The method of any one of claims 1 to 6, wherein the cancer is a solid tumor or a liquid tumor, preferably colon cancer, lung cancer, breast cancer, ovarian cancer, uterus cancer, head and neck cancer or melanoma.

8. The method of any one of claims 1 to 7, wherein the cancer is a lung cancer and the polypeptide comprises a sequence selected from SEQ ID NO 1, 2, 4, 6, 7, 1413, 581, 1457, 1769, 1270, 1276, 1396, 80, 1810, 1864, 847, 2119 or 1451 or

an epitope-containing fragment thereof, particularly from PAP 1, 2, 4, 6, 7, 24, 25, 28, 29, 44, 48, 66, 70, 68, 69, 74, 94, 82 as represented in Table 5.

9. The method of any one of claims 1 to 8, wherein the contacting step comprises contacting the sample simultaneously with several polypeptides comprising the sequence of a distinct aberrant protein domain created by transcription infidelity, preferably from 2 to 10.

10. A method of assessing the physiological status of a subject, the method comprising a step of measuring the presence or level of antibodies specific for aberrant protein domains created by transcription infidelity (TIAB) or of TCR-bearing immune cells that bind to such domains in a sample from the subject, wherein a modified level of said TIAB or immune cells in said sample as compared to a reference value is an indication of a physiological disorder.

11. A method of determining the efficacy of a treatment of a cancer, the method comprising (i) determining the level of at least one polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity or the level of TIAB or corresponding TCR-bearing cells, in a sample from the subject and (ii) comparing said level to the level in a sample from said subject taken prior to or at an earlier stage of the treatment.

12. A method of monitoring the progression or the extension of a cancer in a subject, said method comprising (i) contacting a sample obtained from said subject with at least one polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity, (ii) determining the level of TIAB or corresponding TCR-bearing cells in said sample and (iii) comparing said level to reference value, a variation is indicative of the progression or the extension of the cancer

13. The method of any one of claims 11 to 12, wherein the polypeptide comprises a sequence selected from SEQ ID NOs: 1 to 3334 or an epitope-containing fragment thereof.

14. The method of claim 12, wherein the polypeptide comprises the sequence of any one of the 45 PAP polypeptides of table 5.

15. The method of claim 12 or 13, wherein the contacting step comprises contacting the sample simultaneously with several of said polypeptides, preferably from 2 to 10.

16. A method of determining whether an individual is making a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334, comprising contacting a sample obtained from said individual with an agent indicative of the presence of said polypeptide and determining whether said agent binds to said sample.

17. The method of Claim 16, wherein said agent is a polypeptide which binds to an antibody which specifically binds to a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334.

18. The method of Claim 16, wherein said agent is a polypeptide which binds to an immune cell comprising a TCR specific for a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334.

19. The method of Claim 16, wherein said agent is an antibody or portion thereof which specifically binds to a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334.

20. The method of Claim 16, wherein said agent is an immune cell comprising a TCR specific for a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs: 1-3334.

21. The method of any one of claims 1 to 18, wherein the polypeptide is immobilized on a support.

22. A method of selecting, characterizing, screening or optimizing a biologically active compound, said method comprising placing *in vitro* a test compound in contact with a gene and determining the ability of said test compound to

modulate the production, from said gene, of RNA molecules containing transcription infidelity gaps.

23. A method of producing a peptide specific for transcription infidelity, the method comprising:

a) identifying a protein domain resulting from a transcription infidelity gap;

b) synthesizing a peptide comprising the sequence of said protein domain of a); and

c) optionally verifying, in a biological sample from a mammalian subject, that the peptide binds an antibody.

24. A polypeptide comprising a sequence selected from SEQ ID NOs: 1 to 3334 or an epitope-containing fragment thereof.

25. An isolated polynucleotide encoding a polypeptide of claim 24.

26. An isolated polynucleotide of claim 25, wherein said polynucleotide is labelled

27. An isolated nucleic acid comprising a first nucleotide sequence encoding a polypeptide selected from the group consisting of SEQ ID NOs: 1-3334 or a sequence complementary thereto and a second nucleotide sequence of 100 or less nucleotides in length, wherein said second nucleotide sequence is adjacent to said first nucleotide sequence in a naturally occurring nucleic acid.

28. A cloning or expression vector comprising a polynucleotide of claim 25 or 27.

29. A cell transformed or transfected with a vector of claim 28.

30. An isolated antibody or portion of an antibody which specifically binds to a polypeptide selected from the group consisting of SEQ ID NOs: 1-3334.

31. An isolated cell which specifically binds to a polypeptide selected from the group consisting of SEQ ID NOs: 1-3334.

32. The isolated cell of Claim 31, wherein said cell is an immune cell comprising a TCR specific for said polypeptide selected from the group consisting of SEQ ID NOs: 1-3334.

33. A solid support comprising at least one polypeptide comprising a sequence selected from SEQ ID NOs 1 to 3334 or an epitope-containing fragment thereof.

34. A solid support comprising at least one antibody or portion of an antibody which specifically binds to a polypeptide selected from the group consisting of SEQ ID NOs: 1-3334.

35. A device or product comprising, immobilized on a support, at least one polypeptide comprising the sequence of an aberrant protein domain created by transcription infidelity gap.

36. The device of claim 35, which comprises one or several polypeptides of claim 23.

37. A kit comprising at least a device or product of claim 35 or 36 and a reagent to perform an immune reaction.

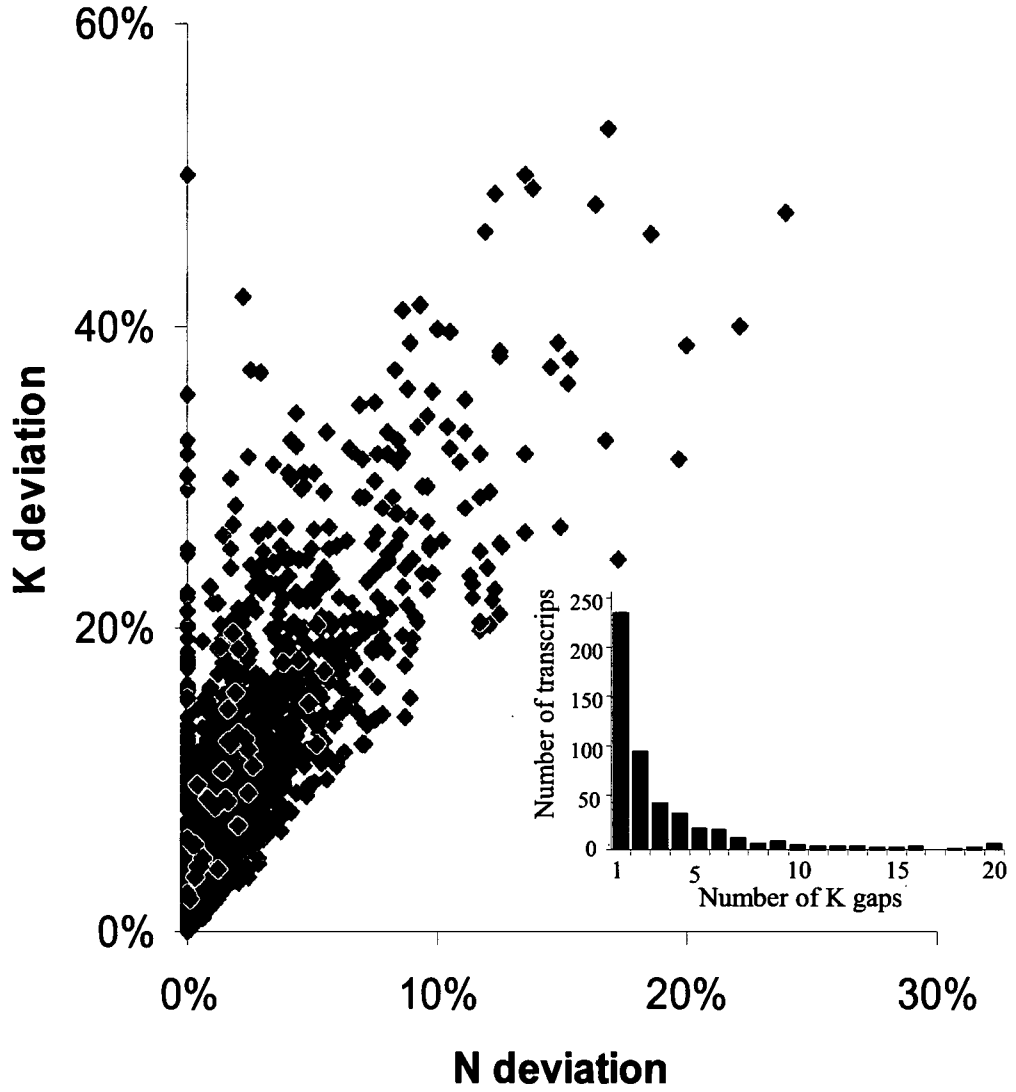

60%
40%
20%
0%
K deviation
0%
10%
20%
30%
N deviation
Number of transcrips
250
200
150
100
50
0
1
5
10
15
20
Number of K gaps

Fig. 1

CFL1 : NM_005507.2 position 447

GACGACCCCTACGCCACCTTTGTCAAGATGCTGCCAGATA

	T	deletion
Cancerous EST	1167	57
Normal EST	622	3

Fig. 2A

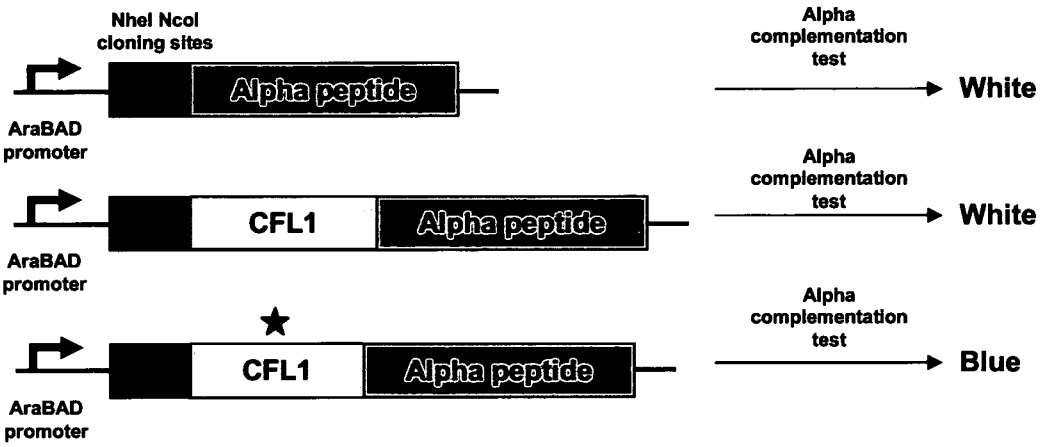


Fig. 2B

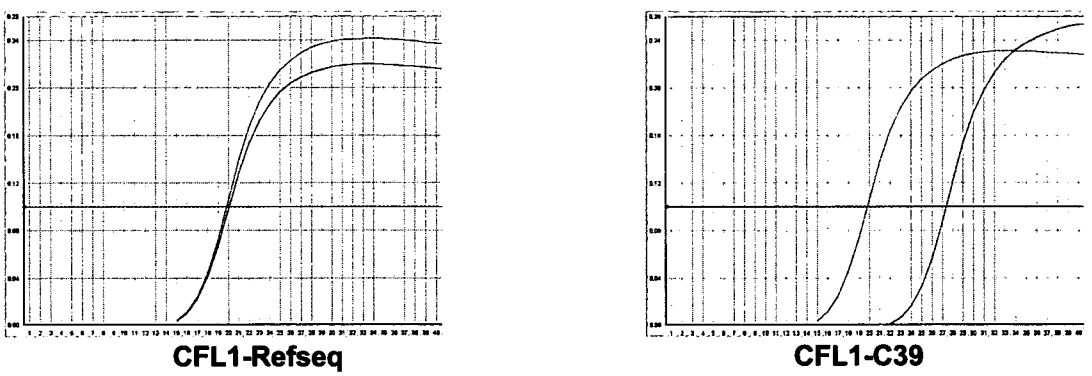


Fig. 2C

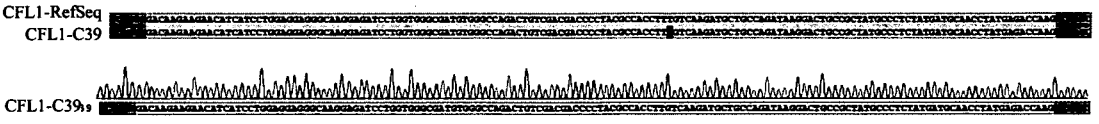

CFL1-RefSeq
CFL1-C39
CFL1-C39₉₉

Fig. 2D

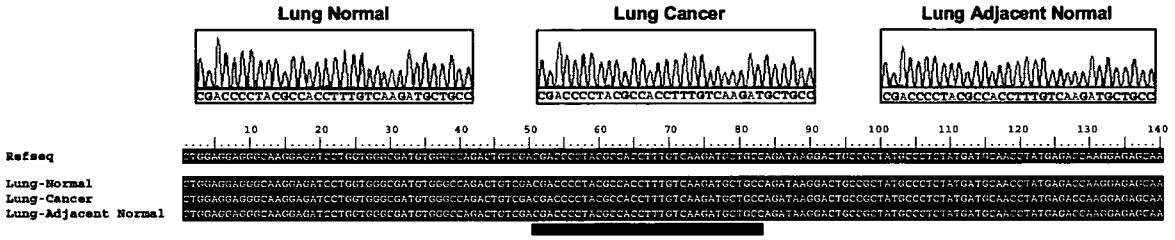

Lung Normal
Lung Cancer
Lung Adjacent Normal
CGACCCCTACGCCACCTTTGTCAAGATGCTGCC
10
20
30
40
50
60
70
80
90
100
110
120
130
140
Refseq
Lung-Normal
Lung-Cancer
Lung-Adjacent Normal

Fig. 2E

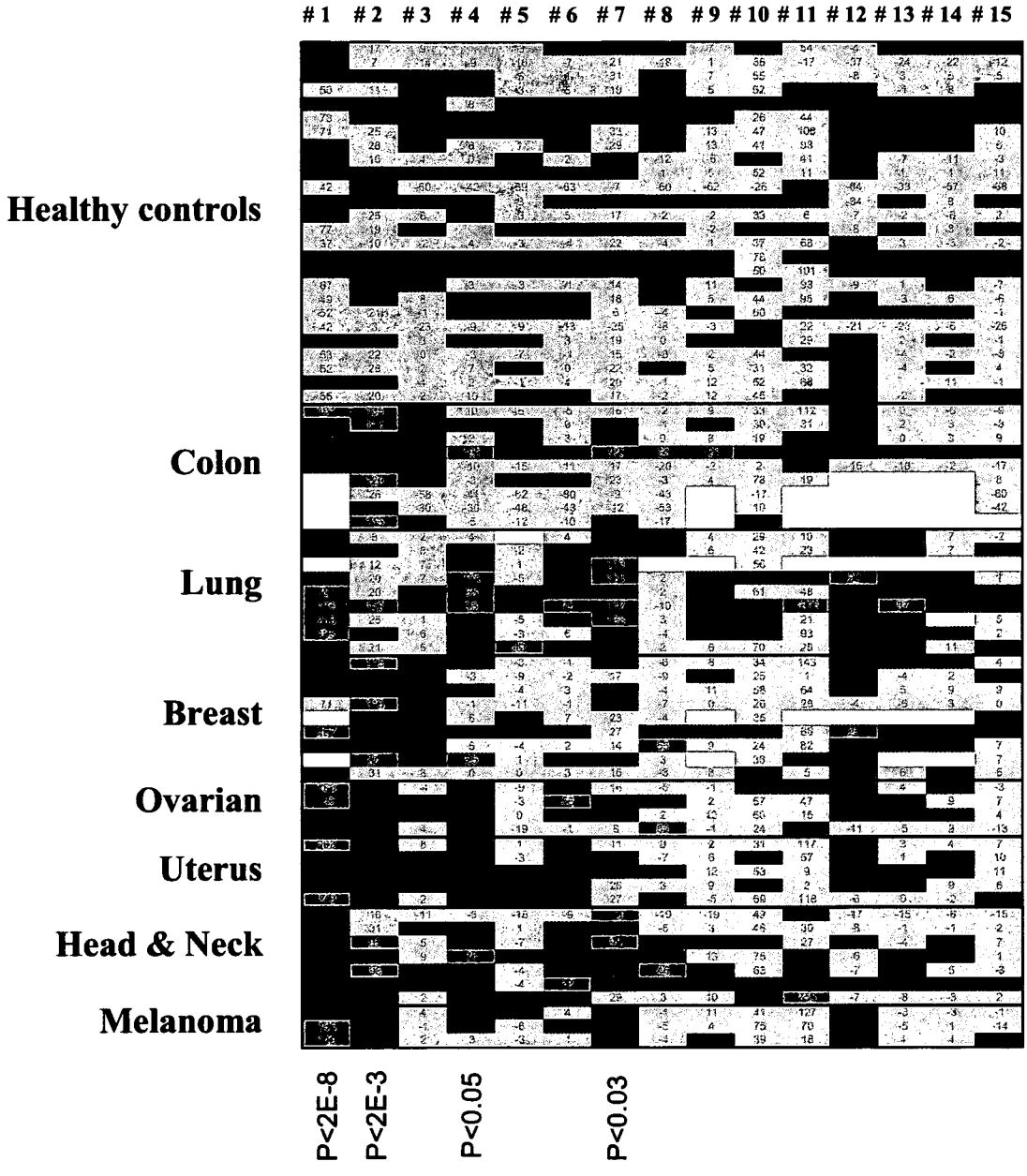

#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15
Healthy controls
Colon
Lung
Breast
Ovarian
Uterus
Head & Neck
Melanoma
P<2E-8
P<2E-3
P<0.05
P<0.03

Fig. 3A

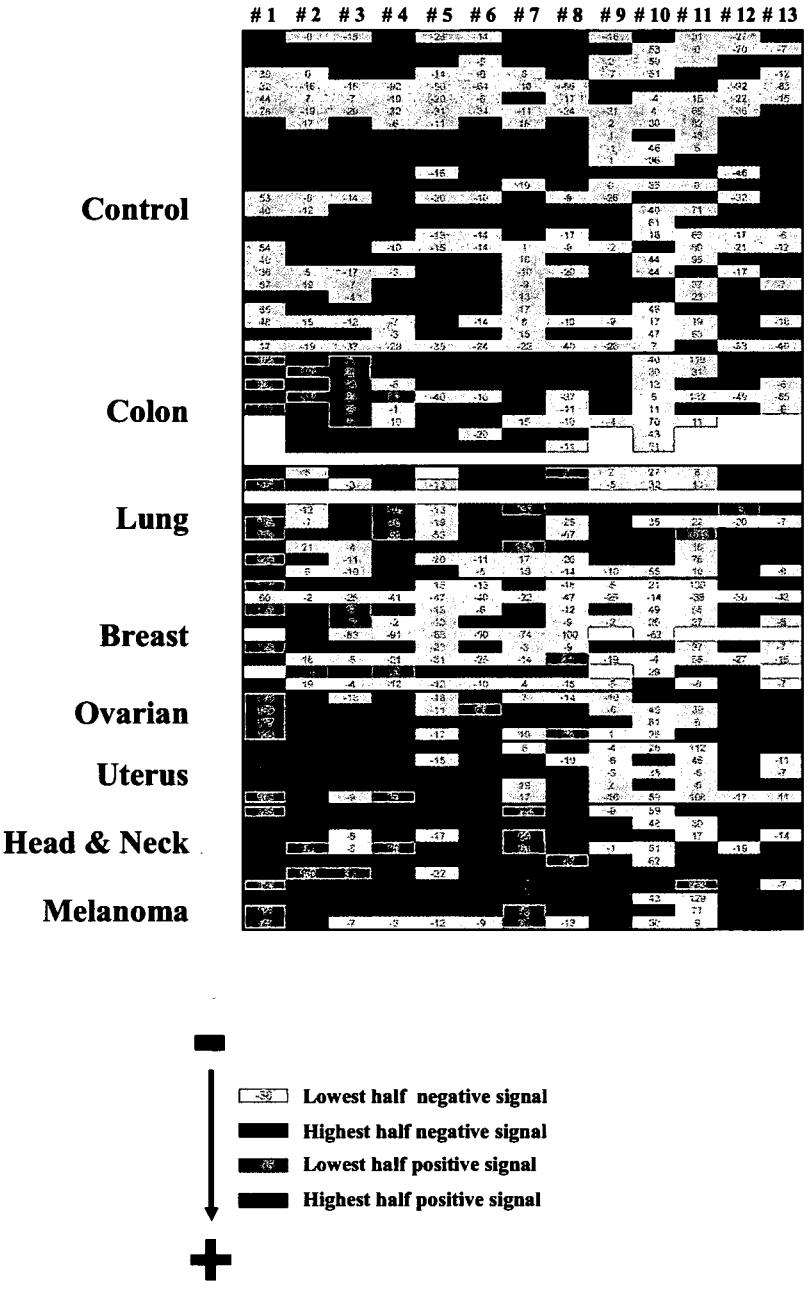
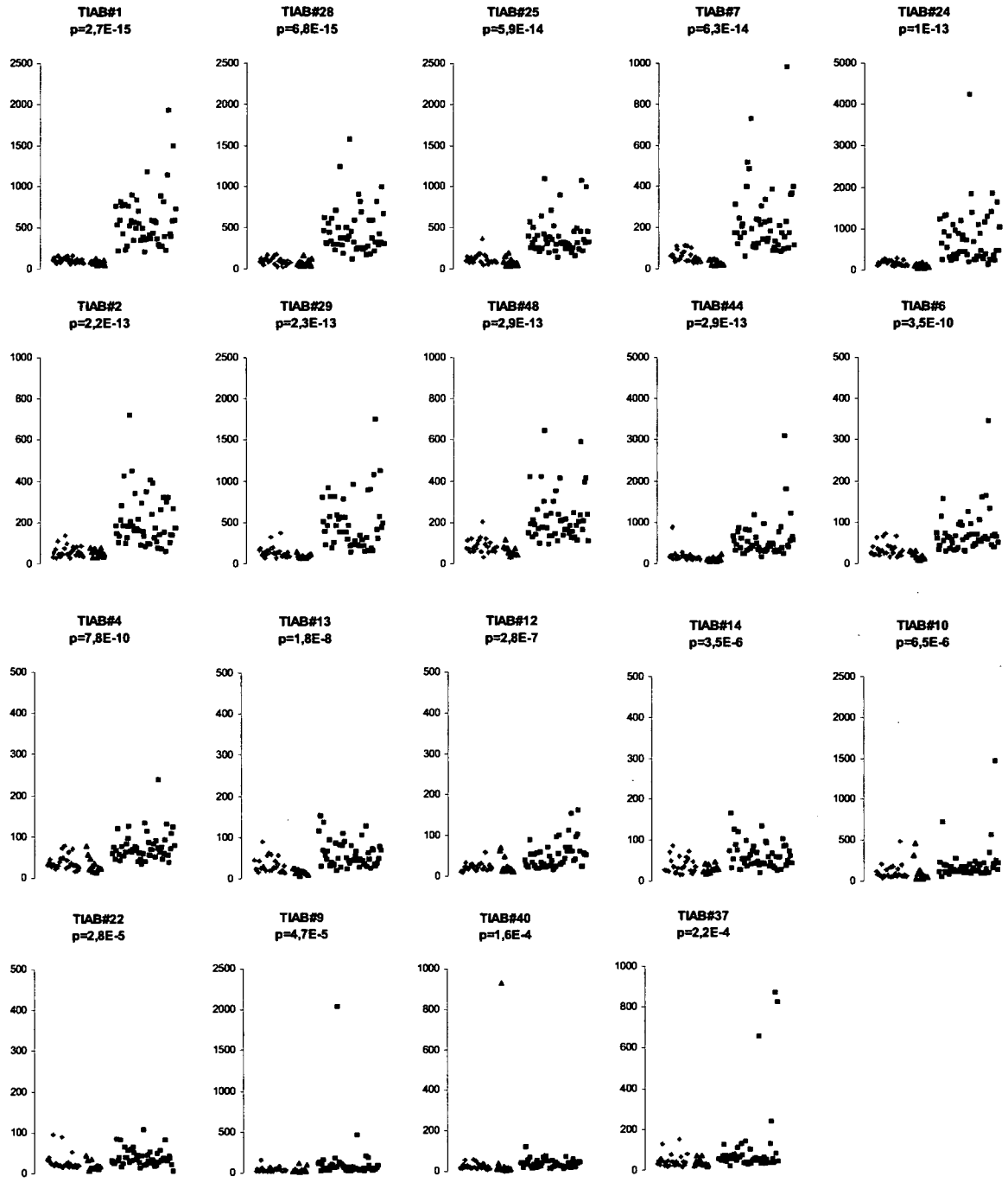

#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13
Control
Colon
Lung
Breast
Ovarian
Uterus
Head & Neck
Melanoma
Lowest half negative signal
Highest half negative signal
Lowest half positive signal
Highest half positive signal

Fig 3B


TIAB#1
p=2,7E-15
TIAB#28
p=6,8E-15
TIAB#25
p=5,9E-14
TIAB#7
p=6,3E-14
TIAB#24
p=1E-13
TIAB#2
p=2,2E-13
TIAB#29
p=2,3E-13
TIAB#48
p=2,9E-13
TIAB#44
p=2,9E-13
TIAB#6
p=3,5E-10
TIAB#4
p=7,8E-10
TIAB#13
p=1,8E-8
TIAB#12
p=2,8E-7
TIAB#14
p=3,5E-6
TIAB#10
p=6,5E-6
TIAB#22
p=2,8E-5
TIAB#9
p=4,7E-5
TIAB#40
p=1,6E-4
TIAB#37
p=2,2E-4

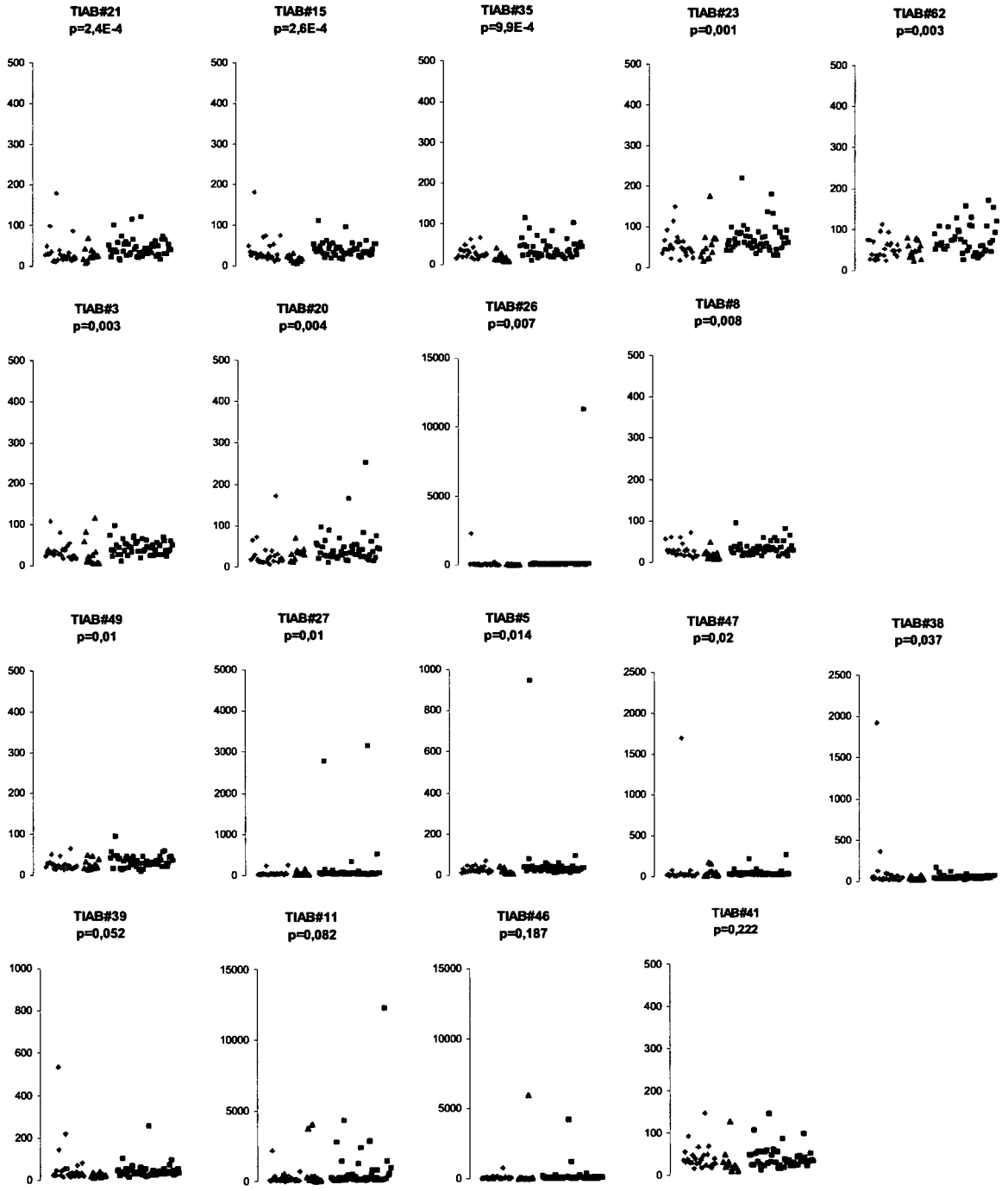

TIAB#21
p=2,4E-4
TIAB#15
p=2,6E-4
TIAB#35
p=9,9E-4
TIAB#23
p=0,001
TIAB#62
p=0,003
TIAB#3
p=0,003
TIAB#20
p=0,004
TIAB#26
p=0,007
TIAB#8
p=0,008
TIAB#49
p=0,01
TIAB#27
p=0,01
TIAB#5
p=0,014
TIAB#47
p=0,02
TIAB#38
p=0,037
TIAB#39
p=0,052
TIAB#11
p=0,082
TIAB#46
p=0,187
TIAB#41
p=0,222

Fig. 4A

Variable	p-value Wilcoxon
TIAB_1	2,7E-15
TIAB_28	6,8E-15
TIAB_25	5,9E-14
TIAB_7	6,3E-14
TIAB_24	1,0E-13
TIAB_2	2,2E-13
TIAB_29	2,3E-13
TIAB_48	2,9E-13
TIAB_44	2,9E-13
TIAB_6	3,5E-10
TIAB_4	7,8E-10
TIAB_13	1,8E-08
TIAB_12	2,8E-07
TIAB_14	3,5E-06
TIAB_10	6,5E-06
TIAB_22	2,8E-05
TIAB_9	4,7E-05
TIAB_40	1,6E-04
TIAB_37	2,2E-04
TIAB_21	2,4E-04
TIAB_15	2,6E-04
TIAB_35	9,9E-04
TIAB_23	0,001
TIAB_62	0,003
TIAB_3	0,003
TIAB_20	0,004
TIAB_26	0,007
TIAB_8	0,008
TIAB_49	0,010
TIAB_27	0,010
TIAB_5	0,014
TIAB_47	0,020
TIAB_38	0,037
TIAB_39	0,052
TIAB_11	0,082
TIAB_46	0,187
TIAB_41	0,222

Fig. 4B

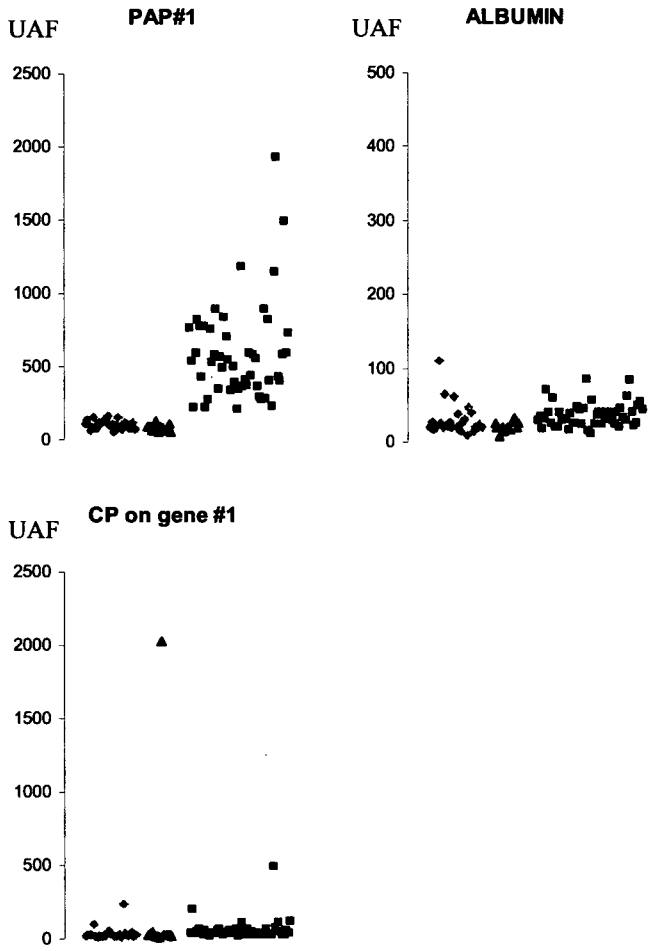

UAF
PAP#1
2500
2000
1500
1000
500
0
UAF
ALBUMIN
500
400
300
200
100
0
UAF
CP on gene #1
2500
2000
1500
1000
500
0

Fig. 5A

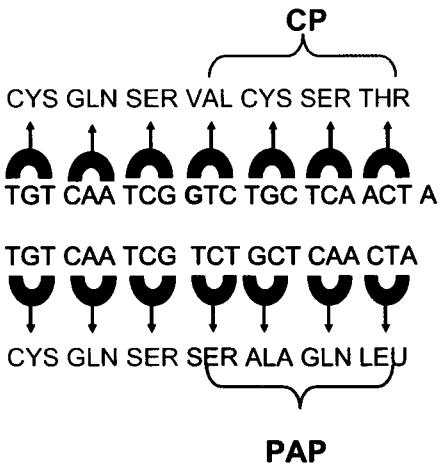

CP
CYS GLN SER VAL CYS SER THR
TGT CAA TCG GTC TGC TCA ACT A
TGT CAA TCG TCT GCT CAA CTA
CYS GLN SER SER ALA GLN LEU
PAP

Fig. 5B

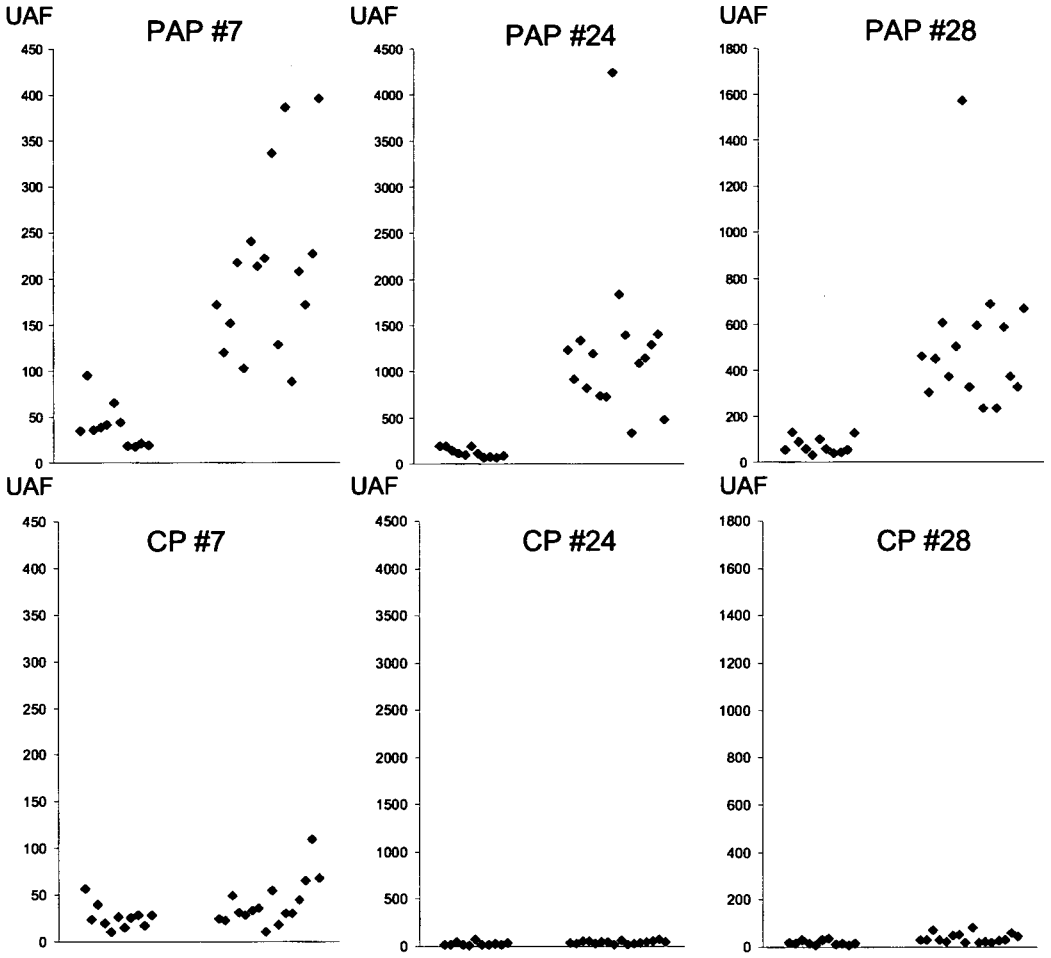


Fig. 5C

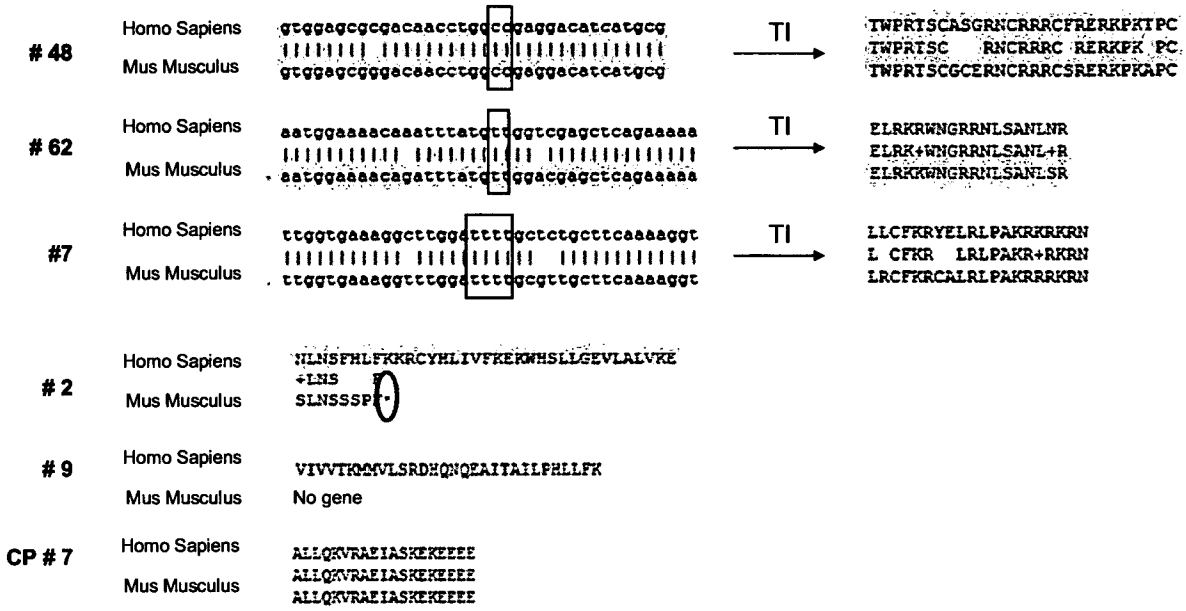

#48
Homo Sapiens
Mus Musculus
TI
#62
Homo Sapiens
Mus Musculus
TI
#7
Homo Sapiens
Mus Musculus
TI
#2
Homo Sapiens
Mus Musculus
#9
Homo Sapiens
Mus Musculus
No gene
CP #7
Homo Sapiens
Mus Musculus

Fig. 6A

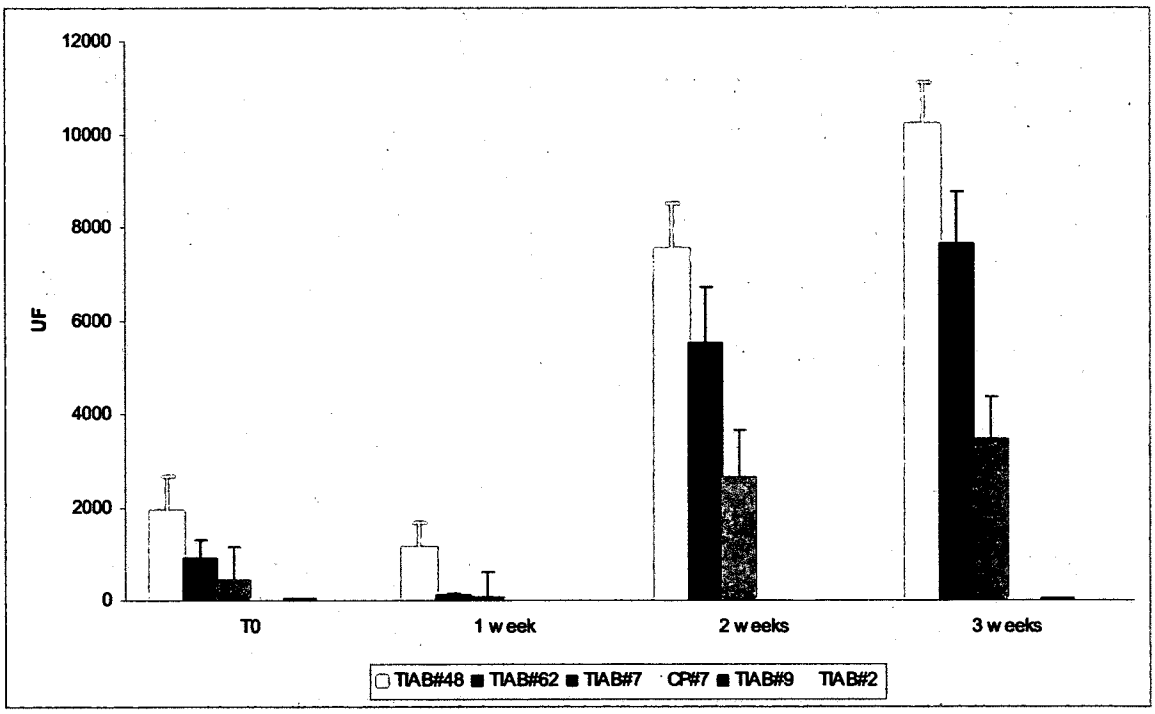

12000
10000
8000
6000
4000
2000
0
UF
T0
1 week
2 weeks
3 weeks
TIAB#48 TIAB#62 TIAB#7 CP#7 TIAB#9 TIAB#2

Fig. 6B

PAP COMBINATION : IK (PAP 7), PRDX6 (PAP 66), CDKN1A (PAP 70), MRPL3 (PAP 29), RPL13A (PAP 68), VIM (PAP 48)

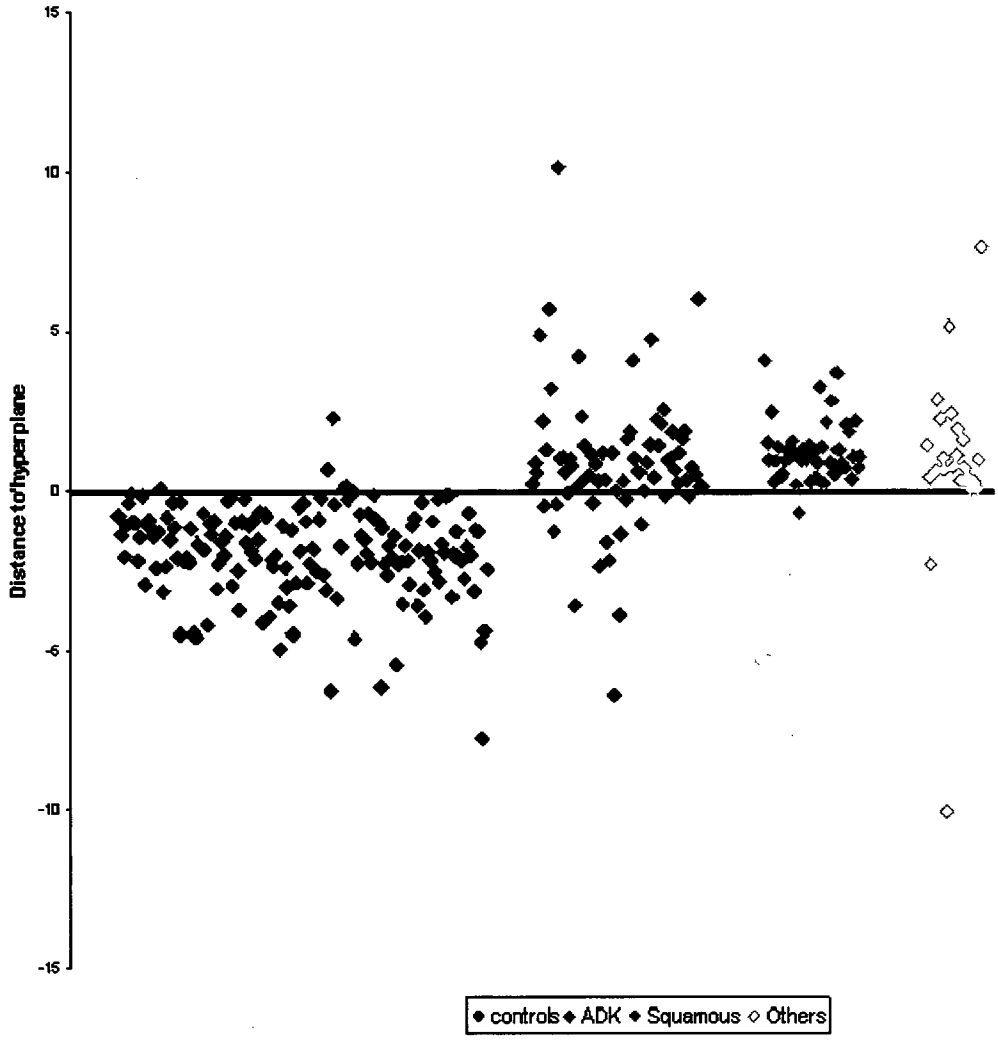


Fig. 7A

Age (y)	n	Positive test	%	
35 – 50	17	15	88	NS
50 – 65	68	59	87	
> 65	55	47	85	

Fig. 7B

Histopathology	n	Positive test	%
ADK	67	51	76
Squamous	40	39	98
Other	33	31	93

X^2 P<0.001

Fig. 7C

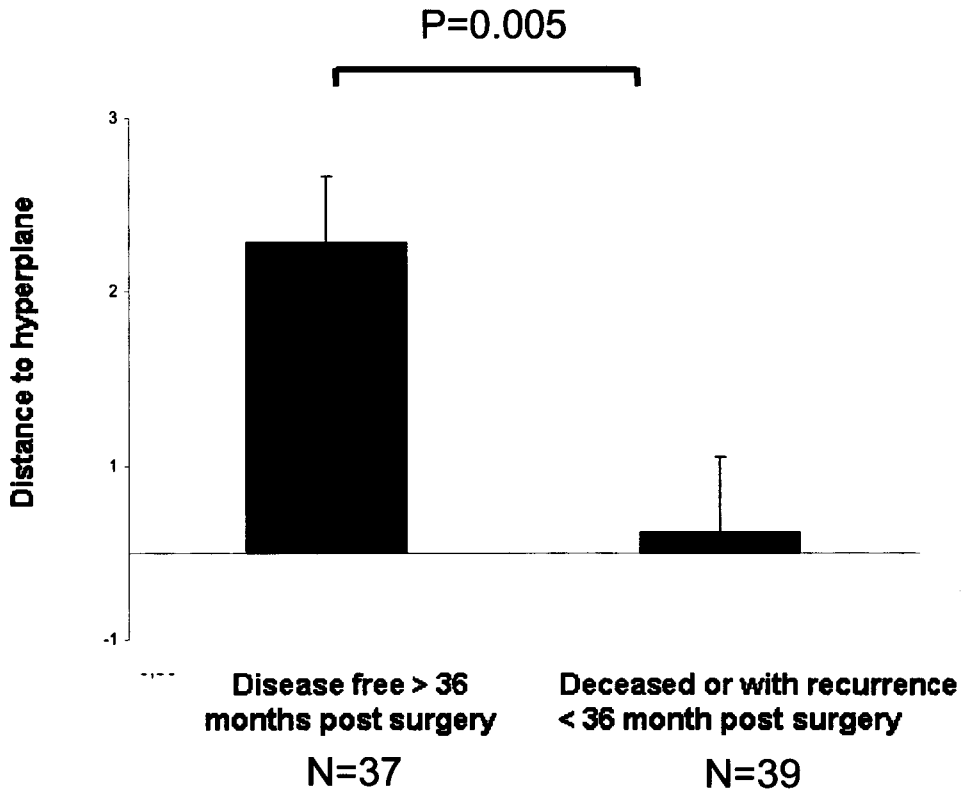


Fig. 7D

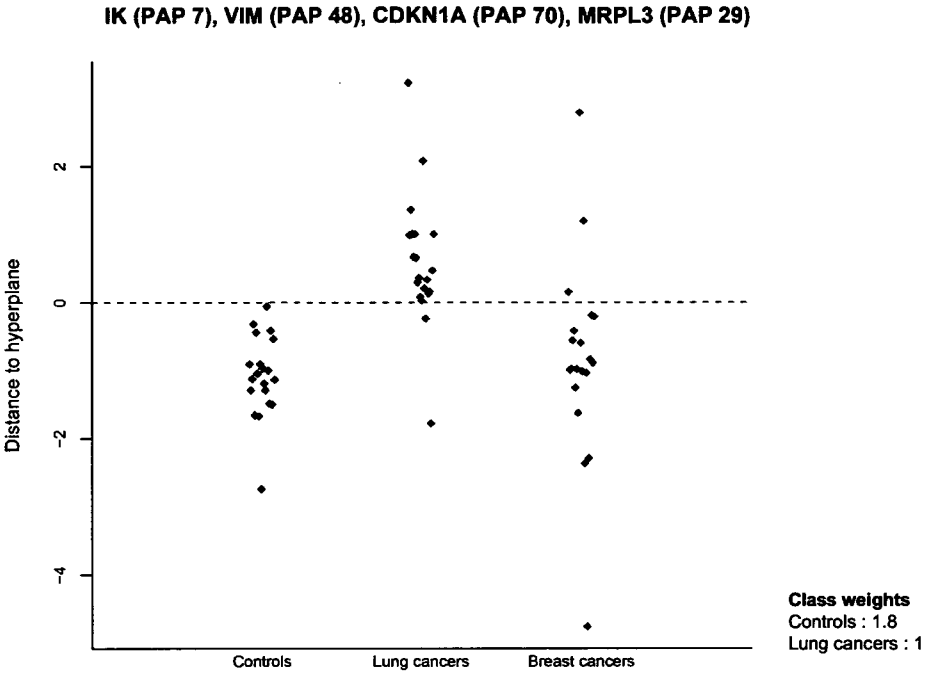
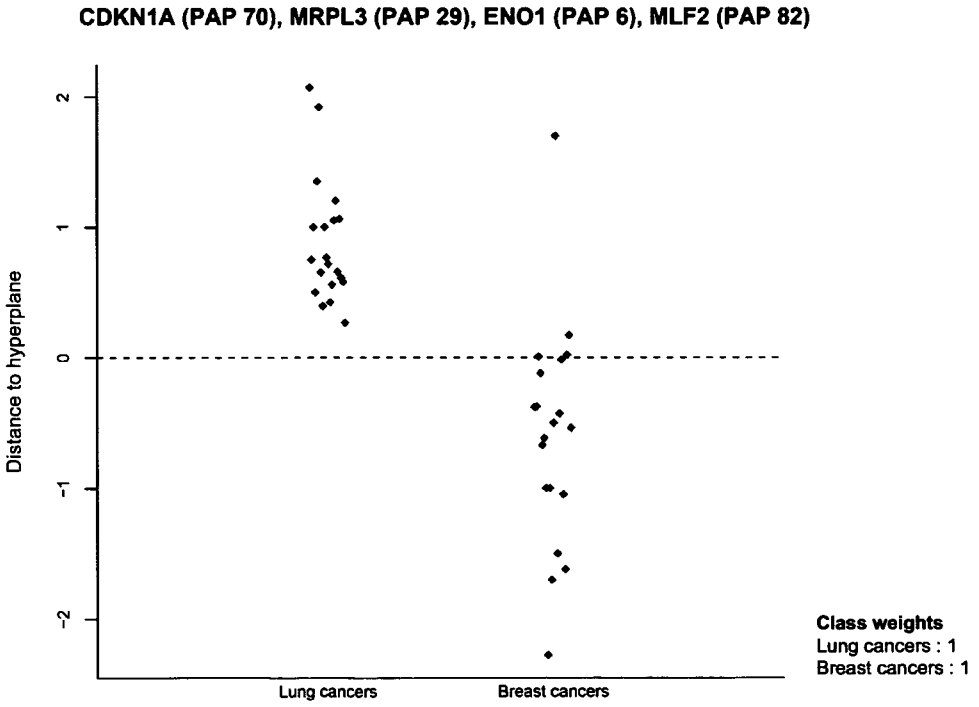
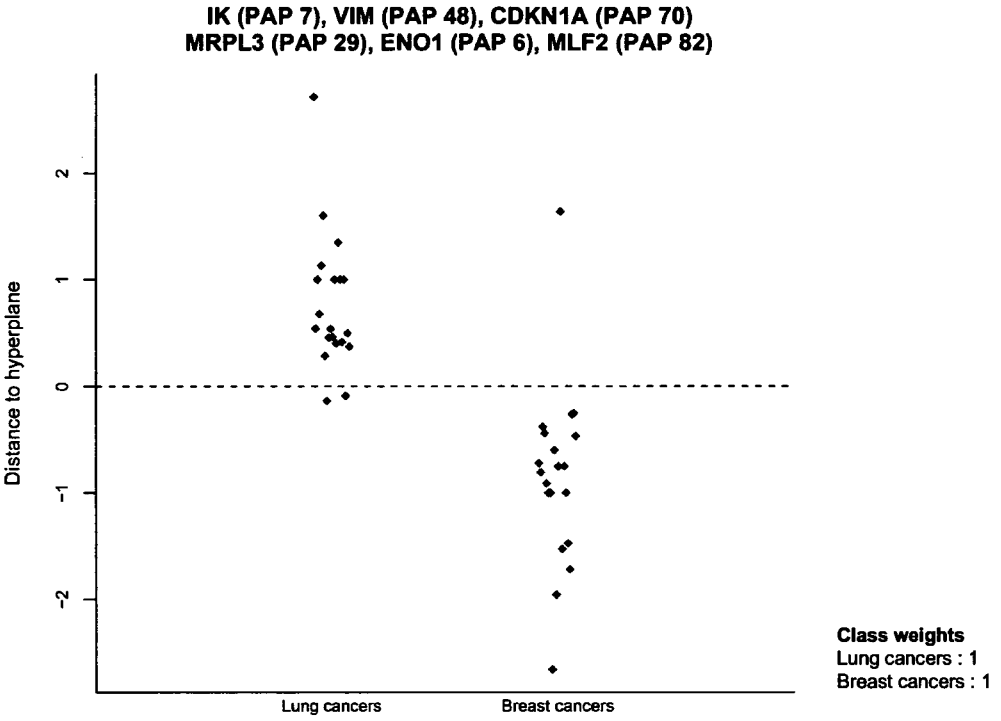
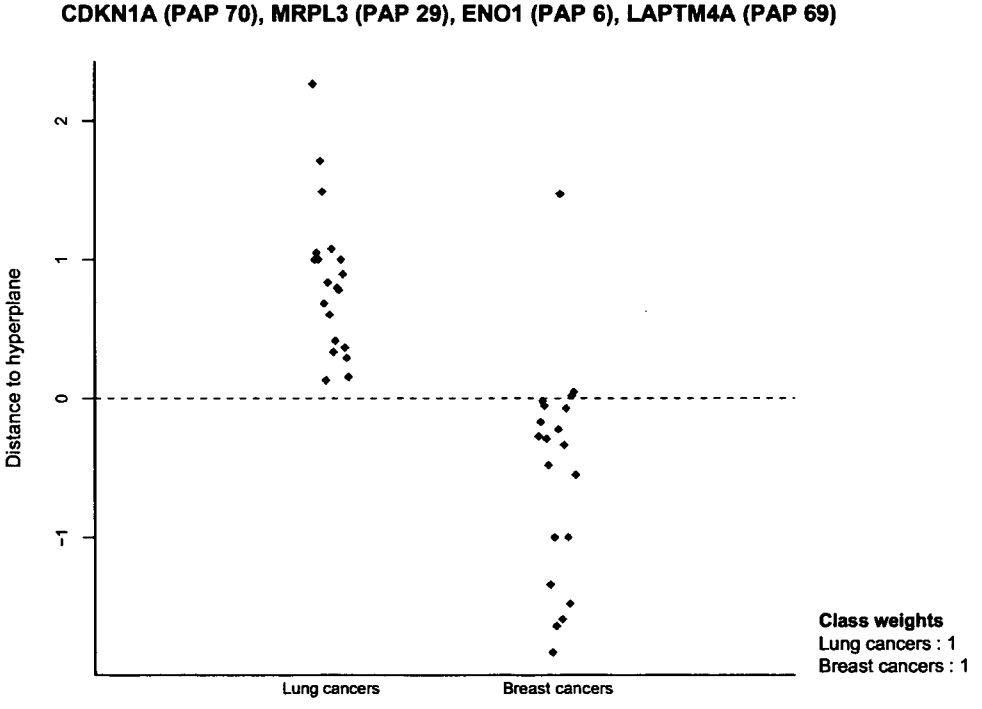

IK (PAP 7), VIM (PAP 48), CDKN1A (PAP 70), MRPL3 (PAP 29)
Distance to hyperplane
2
0
-2
-4
Controls
Lung cancers
Breast cancers
Class weights
Controls : 1.8
Lung cancers : 1

Fig. 8A


CDKN1A (PAP 70), MRPL3 (PAP 29), ENO1 (PAP 6), MLF2 (PAP 82)
Distance to hyperplane
2
1
0
-1
-2
Lung cancers
Breast cancers
Class weights
Lung cancers : 1
Breast cancers : 1


IK (PAP 7), VIM (PAP 48), CDKN1A (PAP 70)
MRPL3 (PAP 29), ENO1 (PAP 6), MLF2 (PAP 82)
Distance to hyperplane
2
1
0
-1
-2
Lung cancers
Breast cancers
Class weights
Lung cancers : 1
Breast cancers : 1


CDKN1A (PAP 70), MRPL3 (PAP 29), ENO1 (PAP 6), LAPTM4A (PAP 69)
Distance to hyperplane
2
1
0
-1
Lung cancers
Breast cancers
Class weights
Lung cancers : 1
Breast cancers : 1

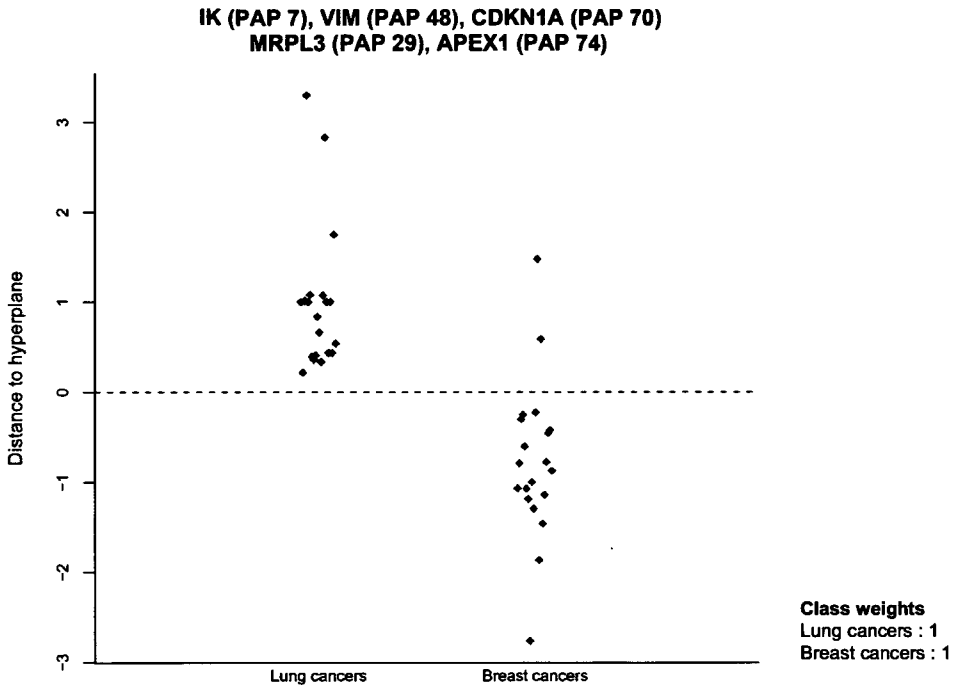

IK (PAP 7), VIM (PAP 48), CDKN1A (PAP 70)
MRPL3 (PAP 29), APEX1 (PAP 74)
Distance to hyperplane
3
2
1
0
-1
-2
-3
Lung cancers
Breast cancers
Class weights
Lung cancers : 1
Breast cancers : 1

Fig. 8B

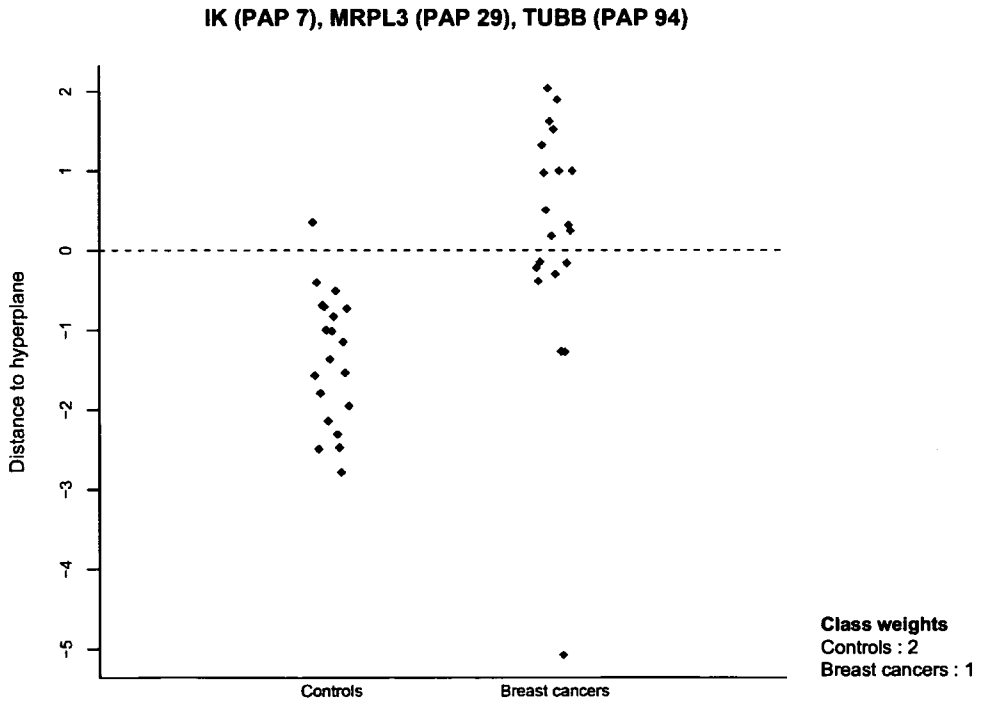

IK (PAP 7), MRPL3 (PAP 29), TUBB (PAP 94)
Distance to hyperplane
2
1
0
-1
-2
-3
-4
-5
Controls
Breast cancers
Class weights
Controls : 2
Breast cancers : 1

Fig. 8C