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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 January 2008 (24.01.2008)

PCT

(10) International Publication Number
WO 2008/009751 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/EP2007/057541

(22) International Filing Date: 20 July 2007 (20.07.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
06291176.3 20 July 2006 (20.07.2006) 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) Inventor; and
(75) Inventor/Applicant *(for US only)*: **BIHAIN, Bernard** [FR/FR]; 26 rue de Metz, 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, 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, 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, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, 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*
— *with sequence listing part of description published separately in electronic form and available upon request from the International Bureau*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2008/009751 A2

(54) **Title:** TRANSCRIPTION INFIDELITY, DETECTION AND USES THEREOF

(57) **Abstract:** The present invention relates to the identification of a novel mechanism of transcription infidelity in cells. The invention provides compositions and methods to detect the level of transcript ion infidelity in a sample, as well as the use thereof, e.g., for therapeutic, diagnostic, pharmacogenomic or drug design. As will be disclosed, the invention is particularly suited for detecting, monitoring or treating proliferative cell disorders, for the design and/or screening of drugs, for patient or disease profiling, prediction of disease severity and evaluation of drug efficacy.

Transcription Infidelity, Detection And Uses Thereof

The present invention relates to the identification of a novel mechanism of transcription infidelity in cells. The invention provides compositions and methods to detect the level of transcription infidelity in a sample, as well as the use thereof, e.g., for therapeutic, diagnostic, pharmacogenomic or drug design. As will be disclosed, the invention is particularly suited for detecting, monitoring or treating disorders (such as for example, proliferative cell disorders), for the design and/or screening of drugs, for patient or disease profiling, prediction of disease severity and evaluation of drug efficacy.

Introduction to the invention

Nucleic acid DNA and RNA are macro-molecules that serve to store genomic information in the cell nucleus (DNA) and to transfer genomic information into the cytoplasm after a process called transcription that generates messenger RNA in order to produce protein. Both are linear polymers composed of monomers called nucleotides. DNA and RNA each represent combinations of four types of nucleotides. All nucleotides have a common structure: a phosphate group linked by a phosphodiester bond to pentose that in turn is linked to an organic base: adenine, cytosine, thymine and guanine (A, C, T, G) for DNA and adenine, cytosine, uracil and guanine (A, C, U, G) for RNA. In RNA, the pentose is ribose and in DNA, it is deoxyribose. The DNA molecule is organized as a double strand of ordered nucleotides arranged as a double helix. RNA molecules are single strand polynucleotides that mainly yield 4 types of molecules: 1) messenger RNA (mRNA), which are single strand nucleotides that are transcribed from DNA and translated into proteins in the cytoplasm, 2) transfer RNA (tRNA), which adopt a well defined 3 dimensional structure and bind specific amino acids (AA) that are transferred to polypeptide chains to form protein with specific AA sequences in a process called translation, 3) ribosomal RNA (rRNA), which are larger molecules that together with specific proteins constitute the ribosome, a structure that allows the assembly of AA into protein following the information provided by the sequence specified by codons (one codon = 3 nucleotides) present in a given order on the mRNA and 4) short regulating RNA referred to as non-coding RNA (ncRNA).

The sequence of mRNA is then transformed into a different language, i.e., the AA language by a process referred to as translation.

Here, we demonstrate for the first time that the fidelity of transcription, i.e., the transfer of DNA information into mRNA, is dramatically reduced in pathological cells, particularly in cancer cells. This lack of transcription fidelity in cancer cells is a general phenomenon that affects all cancers and most of the genes tested in the reported examples as well as the majority of transcript present in available databases and that has several immediate and important implications. In particular, the discovery of transcription infidelity allows the improvement of currently used proteomic and transcriptomic approaches for the investigation of pathological conditions. The discovery of transcription infidelity also allows a better classification of diseases with respect to their severity, and improves the prediction of therapeutic effect and design novel methods to screen the efficacy of drugs on the basis of their ability to correct a lack or modification of transcription fidelity. The present invention has thus major implications and utilities in diagnostic and in the development of drugs, as well as in the treatment, detection and monitoring of patients and drug efficacy.

In order to understand the impact of the discovery described below, it is important to recognize that it is currently believed that absolute fidelity in the transcription of DNA to RNA is needed for normal cell function. However, a currently important but unresolved issue persists. Indeed, sequencing and annotation of the human genome led to the notion that 30-40 thousand genes are encoded. This estimate remains far below the current number of proteins that can be identified by proteomic methods: up to 300 thousand proteins have been identified. To reconcile these differences, it is proposed that a single gene can produce several different mRNA encoding different proteins through a process called alternative splicing. Alternative splicing removes different parts of pre-messenger RNA (pmRNA), thereby leading to removal of different elements corresponding to specific introns, and produces different mature mRNA. Analysis of RefSeq database indicate it contains 11259 transcripts corresponding to 6946 genes. Thus, alternative splicing can increase transcripts heterogeneity by 76 %. The mechanism that we describe here is different from alternative splicing and induced

a far greater protein heterogeneity. Indeed, we demonstrate the occurrence of non-random modifications of mRNA sequence that result in changes in encoded AA, introduction of premature stop codons that yield shorter protein isoforms, alterations of natural stop codons implying introduction of novel coding sequences that yield elongated protein isoforms. Introduction of gap and insertion thereby modify mRNA reading frame and thereby create unsuspected protein sequences. The data described in this invention show that this phenomenon of transcription infidelity is present in normal cells but dramatically enhanced in pathological cells, such as cancer-derived cells. It is thus proposed that transcription infidelity (TI) contributes to diversification of the information transferred from DNA to RNA. Our results further establish that transcription infidelity does not occur randomly but follows specific rules in which the context surrounding b0, bases affected by TI event, is important.

Another mechanism could be envisioned to explain that RNA heterogeneity does not occur at the genomic but at the RNA level: RNA editing. However, RNA editing cannot explain two types of TI event, introduction of gap and insertion. RNA editing is characterized by a post-transcriptional base changes in mRNA and tRNA in eukaryotes. The vast majority of the known substitution editing events consists of C to U or A to I (read as G) conversions (Gott, J. M. & Emeson, R. B. (2000) *Annu Rev Genet* **34,** 499-531 - Maas, S. & Rich, A. (2000) *Bioessays* **22,** 790-802. - Niswender, C. M. (1998) *Cell Mol Life Sci* **54,** 946-64). Different works have shown that these conversions arise via deamination mechanisms, catalyzed by adenosine and cytidine deaminases. One other substitution editing event is the ''U to C'' conversion. While the theory of microscopic reversibility dictates that the cytidine deaminase reaction could go backward to generate this conversion, it has also been proposed that a CTP synthetase-type activity could be responsible. It has been shown in several examples that RNA editing can be specific to cancer cells versus normal cells. Moreover, the rate of this phenomenon can be affected in cancer conditions.

In the cancer set, we observed at the cDNA level *5.7% C → T, 9.2% T → C and 4.7% A → G* changes. Thus, mRNA editing cannot account for more than 20% of single base substitutions described here. Indeed, the most common base substitutions: *A → C (24.6%)* and *T → G (16.8%)* represent base family changes that are currently not explained by known human enzymatic RNA editing processes.

Furthermore, transcription infidelity is expected to cause base deletions and/or insertions through a mechanism (or mechanisms) different from mRNA editing.

Moreover, a recent study shows that number of dbSNP records (Single Nucleotide Polymorphism database) are in fact editing sites (Eisenberg, E. et *al.* (2005) *Nucleic Acids Res.***33** (14), 4612-7). In our approach all SNPs from dbSNP are not considered, therefore known SNPs or false SNPs corresponding to editing sites are excluded.

The mechanism to explain the observed cancer mRNA heterogeneity is therefore transcription infidelity.

Summary of the Present Invention

The present invention relates to compositions and methods of using, detecting or altering transcription infidelity in cells, particularly in mammalian cells. The invention is particularly suited for therapeutic and diagnostic purposes, e.g., to detect and/or treat diseases caused by or associated with an increased or reduced level of transcription infidelity.

A particular object of this invention relates to a method of detecting the presence or stage of a disease in a subject, the method comprising assessing (*in vitro* or *ex vivo*) the presence or rate of transcription infidelity in a sample from said subject, said presence or rate being an indication of the presence or stage of a disease in said subject.

A further object of this invention resides in a method of screening, identifying or optimising a drug, the method comprising a step of assessing whether said drug alters (e.g., reduces or increases) the rate of transcription infidelity of a gene.

A further object of this invention is a method of treating a subject in need thereof, the method comprising administering to said subject an effective amount of a compound that alters (e.g., reduces or increases) the rate of transcription infidelity of a mammalian gene.

A further object of this invention relates to the use of a compound that alters (e.g., reduces or increases) the rate of transcription infidelity of a mammalian gene for the

manufacture of a pharmaceutical composition for use in a method of treatment of a human or animal, particularly for treating diseases such as proliferative cell disorders including without limitation, cancers, immune diseases, inflammatory diseases or aging.

A further object of this invention resides in a method of assessing the efficacy of a drug or candidate drug, the method comprising a step of assessing whether said drug alters (e.g., reduces or increases) the rate of transcription infidelity of a mammalian gene, such an alteration being an indication of drug efficacy.

The invention further relates to methods and products (such as probes, primers, antibodies or derivatives thereof), for detecting or measuring (the level of) transcription infidelity in a sample, as well as to corresponding kits.

The invention also relates to methods of identifying transcription infidelity sequences in proteins or nucleic acids, as well as to their use e.g., as markers, immunogens and/or to generate specific ligands thereof.

In this respect, the invention relates to a method of identifying and/or producing biomarkers, the method comprising identifying, in a sample from a subject, the presence of transcription infidelity site(s) in a target protein or nucleic acid and, optionally, determining the sequence of said transcription infidelity site(s). In a particular and preferred embodiment, the target protein is a cell surface protein or a secreted protein, particularly a cell surface protein or a plasma protein.

A further object of the invention relates to a method of identifying and/or producing a ligand specific for a trait or pathological condition, the method comprising identifying, in a sample from a subject having said trait or pathological condition, the presence of at least one transcription infidelity site in one or several target proteins or nucleic acid optionally determining the sequence of said at least one transcription infidelity site, and producing a ligand that specifically binds to said at least one protein (or domain) or nucleic acid created by transcription infidelity.

The invention is particularly suited for identifying biomarkers of cell proliferative disorders, such as cancers, immune diseases, inflammatory diseases or aging. It is particularly useful for producing ligands that are specific for such disorders in mammalian subjects, in particular ligands that can detect the presence or severity of a cell proliferative disorder in a subject.

A further object of this invention resides in a ligand that specifically binds a protein (or domain) or nucleic acid created by transcription infidelity. The ligand may be an antibody (or any fragment (such as Fab, Fab', CDR, etc.) or derivative thereof, as described later), that specifically binds a protein (or domain) created by transcription infidelity. The ligand may also be a nucleic acid that specifically binds a nucleic acid created by transcription infidelity (e.g., probe, primer, RNAi (interference RNA), etc.).

A further aspect of this invention relates to a peptide comprising a domain of a protein created by transcription infidelity, particularly of a mammalian protein, more preferably of a human protein. The peptide is typically a synthetic peptide, i.e., a peptide that has been prepared artificially, e.g., by chemical synthesis, amino acid synthesis and/or extension, protein digestion, peptide assembling, recombinant expression, etc. The peptide typically comprises the sequence of a C-terminal fragment of said protein. The peptide preferably comprises less than 100, 80, 75, 70, 65, 60, 50, 45, 40, 35, 30, 25 or even 20 amino acids (although, in other embodiments, the peptide length can be higher). The protein may be a cell surface protein (e.g., a receptor etc.), a secreted protein (e.g., a plasma protein etc.), or an intracellular protein.

A further aspect of this invention resides in the use of a peptide created by transcription infidelity as defined above, as an immunogen. The invention also relates to a vaccine composition comprising a peptide comprising a part of a protein created by transcription infidelity, as defined above, and optionally a suitable carrier, excipient and/or adjuvant.

A further aspect of this invention relates to a method of producing an antibody, the method comprising immunizing a non-human mammal with a peptide created by transcription infidelity, as defined above, and recovering antibodies that bind said

peptide, or corresponding antibody-producing cells. Optionally, derivatives of the antibody may be produced.

In a particular embodiment, the invention relates to a method of producing an antibody, the method comprising (i) identifying a domain of a protein created by transcription infidelity and (ii) producing an antibody that specifically binds said domain. Optionally, derivatives of the antibody may be produced.

A further aspect of this invention resides in an antibody that specifically binds a part of a protein created by transcription infidelity, or a derivative of such an antibody having substantially the same antigen specificity. The antibody may be polyclonal or monoclonal. The term derivative includes any fragment (such as Fab, Fab', CDR, etc.) or other derivatives such as single chain antibodies, bi-functional antibodies, humanized antibodies, human antibodies, chimeric antibodies, etc.

A further aspect of this invention relates to an antibody or derivative thereof, as defined above, that is conjugated to a molecule. The molecule may be a drug, a label, a toxic molecule, a radioisotope, etc.

The invention also relates to the use of an antibody or derivative thereof as defined above, for detecting or quantifying (e.g., *in vitro*) transcription infidelity of a gene.

The invention also relates to the use of a conjugated antibody or derivative thereof as defined above, as a medicament or diagnostic reagent.

The invention also relates to a device or product comprising, immobilized on a support, a reagent that specifically binds a protein or nucleic acid created by transcription infidelity. The reagent may be e.g., a probe or an antibody or derivative thereof.

A further aspect of this invention relates to a method of causing or inducing or stimulating transcription infidelity in a cell or tissue or organism. Such a method may comprise, typically, the introduction of a transcription infidelity site into normal genomic sequences, e.g., by means of a gene therapy vector. Such a modification can

result in a destruction of a cell or tissue. In particular, it is possible to create an open reading frame in any gene sequence that results in the production of cell toxic proteins or compounds. It is also possible to cause expression, in a diseased (or target) cell, of a specific biomarker that can be targeted using toxic or therapeutic molecules. Using this approach, it is therefore possible to cause cell death or therapy when transcription infidelity occurs or exceeds a certain rate.

In this regard, a further object of this invention is a method of treating a subject in need thereof, the method comprising introducing into cells of said subject a nucleic acid that contains transcription infidelity site(s), whereby expression of said nucleic acid results in the treatment of said subject.

A further object of this invention is a method of causing or stimulating or controlling transcription infidelity in a subject in need thereof, the method comprising introducing into cells of said subject a nucleic acid that contains transcription infidelity site(s).

A further object of this invention relates to the use of a nucleic acid that contains transcription infidelity site(s) for the manufacture of a pharmaceutical composition for use in a method of treatment of a human or animal. The gene may be any gene, including a mammalian gene or a gene from a pathogenic agent, such as a viral gene, a bacterial gene, etc. In this regard, the invention relates to a method of treating a disease caused by a pathogen, the method comprising causing or stimulating transcription infidelity of a gene encoded by said pathogen.

The invention also relates to methods of producing recombinant polypeptides in vitro with reduced transcription infidelity. Such methods allow a reduction in microheterogeneity in recombinant polypeptides. The method comprises using host cells comprising a recombinant nucleic acid with adapted codon usage, to reduce the occurrence of transcription infidelity. The method may also include the use of any compound or treatment which reduces the occurrence of transcription infidelity. The method may be used in prokaryotic or eukaryotic hosts cells, e.g., in E. coli or CHO strains.

The invention may be used in any mammalian subject, particularly any human subject, to detect, monitor or treat a variety of pathological conditions associated with an increased or reduced transcription infidelity rate, such as e.g., cell proliferative disorders (e.g., cancers), immune diseases (e.g., auto-immune diseases (multiple sclerosis, Amyotrophic Lateral Sclerosis (ALS)), graft rejection), aging, inflammatory diseases, diabetes, etc., and/or to produce, design or screen therapeutically active drugs. The invention may also be used to detect, monitor, modulate or target transcription infidelity in any other tissues or cell types, such as prokaryotic cells, lower eukaryotes (e.g., yeasts), insect cells, plant cells, fungi, etc.

Legend to the Figures

Fig 1: Principle of cDNA library construction and sequencing.

Fig 2: (a-q) mRNA reference sequences used for the analysis. In order to avoid distorting the blast, the polyA tail of mRNA reference sequences were systematically removed. Fig 2r provides a list of tested genes.

Fig 3: Typical MegaBLAST output files.

Fig 4: Percentage of nucleotide sequence deviation at any position for each studied genes.

Fig 5: TPT1, VIM = number of ESTs and proportion test analysis.

Fig 6: Variations in ESTs before and after sequential application of electronic filters. Clip 400 and cell lines removal effects.

Fig 7: DNA context: a) Effect of pmRNA base composition on base b0 heterogeneity. b) Composition of replacement base for each substituted base. c) Repartition of affected and replacement bases within statistically significant C>N tests. d) Effect of pmRNA base composition on substitution base corresponding to repetition of b-1 or b+1

Fig 8: Virtual variant mRNA and protein for VIM.

Fig 9: Coding impact

Fig 10: Proteins of interest = amino acid sequences when the stop codon is statistically affected.

Fig 11: DHPLC = principle and limits of detection.

Fig 12: DHPLC = primers and expected PCR products.

Fig 13: List of 60 studied proteins

Fig 14: Antibody production.

Fig 15: DHPLC Results

Fig 16: Sanger sensibility

Fig 17: Detection of APOAII and APOCII PSP in cancerous-patient plasma

Fig 18: Potential PSP on plasma proteins

Fig 19: Evaluation of mRNA sequence substitutions predictions

Detailed Description of the Invention

Definitions

Transcription Infidelity

The term transcription infidelity designates a novel mechanism by which several distinct RNA molecules are produced in a cell from a single gene sequence. This newly identified mechanism potentially affects any gene, is non-random, and follows particular rules, as will be disclosed below. As shown in the examples, transcription infidelity can introduce substitutions, deletions and insertions in RNA molecules, thereby creating a diversity of proteins from a single gene. Transcription infidelity can also affect non coding RNA sequences, thereby modulating their functions. Measuring, modulating or targeting transcription infidelity therefore represents a novel approach for detecting or treating disorders, as well as for drug development.

Transcription Infidelity Site

Within the context of the present invention, the term Transcription Infidelity Site designates a sequence and/or position affected by transcription infidelity. This can be a nucleic acid or amino acid domain that contains at least one modification generated as a result of transcription infidelity. Such a modification can result e.g., from a switch in the coding frame (insertion and/or deletion), from the introduction or suppression of stop codons, from the introduction of a new start codon, from nucleotide(s) substitution(s) (implying or not change of AA), etc. A transcription infidelity site may comprise one or several sequence variations resulting from transcription infidelity. A transcription infidelity site can comprise from e.g., 1 modified nucleotide or amino acid residue to

e.g., 150 modified nucleotides or amino acid residues, or even more. The transcription infidelity site typically differs from the sequence resulting from faithful transcription by at least one nucleotide or amino acid modification (e.g., substitution, deletion, insertion, inversion, etc.). A protein or a domain created by transcription infidelity (TI-protein) typically comprises from 1 to 50 amino acids (or even more), with at least one modified amino acid residue. A nucleic acid transcription infidelity sequence typically comprises from 1 to 150 nucleotides (or even more), with at least one modified nucleotide residue.

Transcription Infidelity Rules and Identification

Based on specific techniques of detection and/or particular rules of transcription infidelity as defined in the present application, it is now possible to predict and identify, for any gene or protein, Transcription Infidelity Sites. These sites can also be obtained by aligning sequences available for given RNA and identifying base substitutions. They can ultimately be validated by various techniques, e.g. by using ligands specific for a predicted TI-protein.

A particular method for identifying a transcription infidelity site comprises:

- providing the sequence of a given gene, RNA or cDNA molecule, or a portion thereof, and
- identifying, within said sequence, the presence of a nucleotide change (e.g., substitution, deletion, insertion, inversion) resulting from transcription infidelity, following the transcription infidelity rules, e.g., as discussed in (iii) below.

A further particular method for identifying a transcription infidelity site comprises:

- providing the sequence of a given protein or a portion thereof, and
- identifying, within said sequence, the presence of an amino acid change (e.g., substitution, deletion, insertion, inversion) resulting from transcription infidelity, following the transcription infidelity rules, e.g., as discussed in (iii) below.

The presence of a change resulting from transcription infidelity in a molecule can be identified in a three step process, comprising:

(i) identification of transcription infidelity sites with a learning machine relying on quadratic discrimination on Multiple Correspondence Factorial Analysis (MCFA) factors

(ii) identification of the category of the transcription infidelity sites (b-1 or b+1 or else) with the same method, and

(iii) predicting the replacement base using the following transcription infidelity rules :

. For any base present as singleton, i.e., any base that is not preceded or followed by itself, then the substituted base is most likely identical to the base preceding the one that is substituted (in case of b-1 category) or identical to the base following the one that is substituted (in case of b+1 category). For example CAT will become CCT (b-1 rule); ATG will become AGG (b+1 rule);

. When substitutions occur within three consecutive A, substitution of second A goes preferentially to C;

. When substitutions occur within three consecutive T, substitution of second T goes preferentially to C, then A, then G;

. Stretches of C and G are rarely substituted on the second, but if any, substitution of second C goes preferentially to A and substitution of second G goes preferentially to C;

. For other cases, replacement base is preferentially C.

Further methods to identify transcription infidelity sites within the sequence of any target protein or nucleic acid comprise e.g., comparison of existing expressed sequence tag (ESTs), direct amplification of identified sequences and sequencing of amplification products with specific sequencing method (Sanger, chemical, Pyro-sequencing). Such an approach yields variant sequences that involve AA changes, modifications of protein length, etc. Oligonucleotide selectively matching mRNA or cDNA bearing transcription infidelity sites can then be designed, allowing selective amplification of nucleotide with transcription infidelity sites from a pool of sequences. Antibodies targeting TI-protein can also be produced.

Alternatively, it is possible to identify a transcription infidelity site by means of sequence analysis or bioinformatics rules. The resulting sequence may be cloned in any transfection vector. It is also possible to design a construct comprising such transcription infidelity sequence in frame with a reporter gene, that will be transcribed and translated only if transcription infidelity occurs. The reporter gene may be an enzyme or fluorescent protein or any gene that will make the target cell sensitive, or resistant to exposure to a toxin.

A detailed method that allows the identification of transcription infidelity sites (or substitutions) includes the TDG method as disclosed below:

A technique described by Pan and Weissman and Liu et *al.(* PNAS, 2002, 99(14), 9346-9351 and Anal Biochem. 2006 Sep 1;356(1):117-24) can be adapted for detecting transcription infidelity site(s) in RNA. This technique is based on the use of an enzyme, Thymine DNA glycosylase (TDG) that is capable of separately enriching mismatch-containing DNA duplexes from complex mixtures. These enzymes specifically recognize nucleotide mismatches and generate abasic sites (the bond between the deoxyribose and one of the bases of DNA is cleaved). Then, TDG can reversibly bind these abasic sites and thus can be used for affinity purification of mismatch-containing DNA fragments. In a typical experiment, RNA are extracted from 2 cell types (normal and cancerous) and reverse transcribed to make double-stranded cDNA. After heat denaturation and slow renaturation of each sample, cDNA with mismatche(s) are generated and can be separated from perfect duplexes. These cDNA can then be analysed either by direct sequencing or compared by DHPLC (see later).

Further techniques that may be adapted to identify transcription infidelity sites or mutations are disclosed, e.g., in US6,329,147 ; US4,979,330 ; WO02/077286 or US6,120,992.

Typically, the methods of identifying transcription infidelity sites further comprise a step of validating the sequence of the transcription infidelity site by producing a molecule comprising the identified sequence, generating a ligand that specifically binds

to said molecule and verifying, in a biological sample, the presence of an antigen specifically recognized by said ligand.

A typical transcription infidelity sequence of a nucleic acid is a sequence of between 1 and 150 nucleotides in length comprising at least one base substitution, deletion or insertion caused by transcription infidelity, preferably following a rule as disclosed in Example 5. The transcription infidelity sequence could be higher.

A typical domain of a TI-protein is an amino acid sequence of between 1 and 50 amino acids in length comprising at least one amino acid substitution (or more changes due to modification in the coding frame resulting from base deletion or insertion) caused by transcription infidelity, preferably following a rule as disclosed in Example 5. The domain of a TI-protein could be higher. Examples of domains of TI-proteins are provided in the experimental section, e.g. in Figures. 13 and 14.

Detecting or Measuring Transcription Infidelity

This invention is based on the unexpected discovery of a high rate of natural sequence abnormalities occurring during or shortly after transcription of DNA to RNA. The process may be present in normal cells but greatly increases in pathological cells, e.g., cancer cells. The discovery that DNA transcription into RNA is introducing sequence variations following rules that are different from those defined by one to one base complementarities allows the rational design of novel reagents (e.g., probes, primers, antibodies, aptamers etc.) that are specific for nucleic acid or protein created by transcription infidelity. Such reagents can therefore allow to detect or measure transcription infidelity, and to discriminate between normal or diseased conditions, as well as to target molecules to cells exhibiting transcription infidelity.

Furthermore, the invention also allows the rational design of reagents (e.g., probes, primers, antibodies, aptamers) that are predictive of disease (e.g., cancer) severity. Indeed, a greater rate of transcription infidelity is expected to correlate with the severity of the disease and this rate is increasing progressively as more and more gene products are affected. The direct measure of transcription (in)fidelity in diseased cell(s) using the

methods described herein or any other technology allows to detect cells at different stages of disease progression in any given tissue. Accurately measuring these variations in gene expression (transcripts and proteins) also improves the capacity to evaluate drug efficacy with respect to disease severity. Currently, various microarray techniques allow measurement of changes in gene expression. However, the reliability and, most importantly, the reproducibility of these data are currently not sufficient. We can now postulate that a great deal of variability of these transcriptomic experiments is caused in part by transcription infidelity as discovered by the inventors. Thus, the discovery of this common phenomenon allows the design of gene expression reagents that either are minimally affected by transcription infidelity or that directly reflect transcription infidelity occurring in specific sequences following the rules described in the present application.

It is also possible and even likely that transcription infidelity is modulated by the transcription rate. We speculate that increase expression of a given gene in pathological condition will increase TI and thereby increase protein heterogeneity.

Transcription infidelity can be detected or measured using a number of techniques known per se in the art, which can be adapted to the present invention. In particular, transcription infidelity can be measured using reagents specific for nucleic acid or protein created by transcription infidelity, such as specific probes, primers or antibodies, by electrophoresis, gel migration analysis, spectrometry, etc., that allow the detection between various forms of a protein or nucleic acid. The rate of transcription infidelity can be determined by assessing the number of genes in a cell that are subject to transcription infidelity, and/or the number of transcription infidelity sites generated for a given gene. Such a rate can be determined by comparing the level of nucleic acid or protein created by transcription infidelity in a test sample to that obtained in a reference sample.

Detection of Transcription Infidelity at the level of nucleic acids

As discussed above, transcription infidelity introduces sequence variation(s) in RNA molecules. Detecting or measuring transcription infidelity can thus be accomplished by detecting the presence or (absolute or relative) amount of such sequence variation(s) in RNAs encoded by one or several genes, or in a whole cell or tissue or sample. The

detection of transcription infidelity in nucleic acids can be performed by various techniques such as hybridization, amplification, heteroduplex formation, etc.

Virtually all technologies used for identification of nucleotides present in diverse tissue types rely on a process called hybridization. Hybridization is critical for technologies such as microarrays used to identify polymorphisms and search for variation in gene expression. This technique is applied to measure the level of gene expression. Hybridization of oligonucleotide probes is also used to subtract sequences of gene abundantly expressed in order to allow better study of low abundance messages – this procedure is called subtractive hybridization. Hybridization is also the first step of gene amplification reaction commonly used in PCR experiments either under the direct form or after application of reverse transcriptase to convert message sequence from RNA type into DNA type sequence.

The basic mechanism underlying hybridization is that single strand nucleotide(s) sequences can bind to one another provided that their respective sequences are complementary. Both the length and the specific ordering of each of the 4 bases define the efficacy of hybridization and specificity of the sequence. This implies that each specific base binds in a non-covalent manner to its complementary base: A to T and C to G and vice versa. This non-covalent binding in a sequence match is conditioned by the ordering of the base. Thus, in general practice, a defined sequence of bases binds to a single complementary sequence. This specific binding reaction provides the basis for hybridization that allows identification of any complementary sequence in any given mixture of nucleotides. The efficacy of hybridization is determined by the number of bases that, in the appropriate order, match to one another (some degree of mismatches are tolerated) and determined by the conditions of the experiment such as the stringency of the binding buffer, the relative content of CG versus TA – CG and TA are linked by 3 and 2 hydrogen bonds, respectively - and by the melting temperature, i.e. the temperature that is needed to allow dissociation of 50% of double-stranded DNA.

Detecting transcription infidelity using hybridization typically comprises placing a sample in contact with a nucleic acid probe that is specific for a transcription infidelity sequence, and detecting the presence (or amount) of hybrids formed, said presence or

amount being a direct indication of the presence or rate of transcription infidelity. In a preferred embodiment, the method uses a set of nucleic acid probes that are specific for distinct transcription infidelity sequences of one or several genes, respectively. Nucleic acid probes specific for a transcription infidelity sequence can be prepared for any gene using the sequence information and nucleotide substitution, insertion or deletion rules as disclosed in the present application (see e.g., Example 5).

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 sequence or a complement thereof, and which is suitable for detecting the presence (or amount thereof) in a sample containing said sequence or complement. Probes are preferably perfectly complementary to a transcription infidelity sequence however, some 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 sequence.

A specific embodiment of this invention is a nucleic acid probe that selectively hybridizes to a region of a nucleic acid molecule that does not contain a transcription infidelity sequence.

Selectivity, when used to denote nucleic acid hybridization, indicates that hybridization of the probe to the target sequence is distinct from hybridization of said probe to another sequence. In this regard, although perfect complementarity between the probe and the target sequence is not required, it shall be sufficiently high to allow selective hybridization.

Preferred probes of this invention comprise a sequence which is complementary to a transcription infidelity sequence in a RNA molecule (or corresponding cDNA molecule) that encodes a cell surface protein or a secreted protein. Specific examples of probes of

this invention have a sequence complementary to a transcription infidelity sequence or encoding a peptide sequence of any one of SEQ ID Nos : 1 to 32.

The sequence of the probes can be modified, e.g., chemically, in order e.g., to increase the stability of hybrids (e.g., intercalating groups or modified nucleotides, such as 2' alkoxyribonucleotides) or to label the probe. Typical examples of labels include, without limitation, radioactivity, fluorescence, luminescence, enzymatic labelling, and the like. The probe may be hybridized to the target nucleic acid in solution, suspension, or attached to a solid support, such as, without limitation, a bead, column, plate, a substrate (to produce nucleic acid arrays or chips), etc.

According to another embodiment (which may be used in combination with hybridization probes), transcription infidelity is detected or measured by selective amplification using specific primers. Detecting transcription infidelity by selective amplification typically comprises placing a sample in contact with a nucleic acid primer that specifically amplifies a transcription infidelity sequence, and detecting the presence (or amount) of amplification products formed, said presence or amount being a direct indication of the presence or rate of transcription infidelity. In a preferred embodiment, the method uses a set of nucleic acid primers that allow specific amplification of distinct transcription infidelity sequence of one or several genes, respectively. Amplification may be performed according to various techniques known per se in the art, such as, without limitation, by polymerase chain reaction (PCR), ligase chain reaction (LCR), transcription-mediated amplification (TMA), strand displacement amplification (SDA) and nucleic acid sequence based amplification (NASBA).

Nucleic acid primers specific for a transcription infidelity sequence can be prepared for any gene using the sequence information and nucleotide substitution, insertion or deletion rules as disclosed in the present application (see e.g., Example 5).

The term "primer" designates a nucleic acid or oligonucleotide having a polynucleotide sequence which is capable of selective hybridization with a transcription infidelity sequence or a complement thereof, or with a region of a nucleic acid that flanks a

transcription infidelity site, and which is suitable for amplifying all or a portion of said transcription infidelity site in a sample containing said sequence 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.

The term "flanks" indicates that the region should be located at a distance of the transcription infidelity site that is compatible with conventional polymerase activities, e.g., not above 250 bp, preferably not exceeding 200, 150, 100 or, further preferably, 50 bp upstream from said site.

A further aspect of this invention also includes at least one pair of nucleic acid primers, wherein said pair of primers comprises a sense and a reverse primer, and wherein said sense and reverse primers allow selective amplification of a transcription infidelity sequence, or of the exactly complementary sequence. Specific examples of primers of this invention are disclosed e.g., in ***Fig 12***.

According to a further embodiment of this invention, transcription infidelity is measured by Denaturing High Performance Liquid Chromatography (DHPLC). The principle of this method is disclosed e.g., in ***Fig 11***. Basically, amplification products are denatured and re-annealed. During re-annealing, both homoduplexes and heteroduplexes are formed, as a result of the presence of transcription infidelity sequences. The mixture is then analyzed by DHPLC. Since heteroduplexes and homoduplexes DNA structures are different at the analysis temperature, they are eluted differently and their (relative) amount can be assessed.

As a verification of transcription infidelity sequences existing in cancer cells with greater frequency than in normal cells, predicted transcription infidelity of selected genes (ENO1, GAPDH, TMSB4X) were amplified by RT-PCR using the indicated oligonucleotides (see Example 6). Variations of homo- and heteroduplexes were verified (fig. 15).

Any other technique suitable for detecting nucleic acids can be used or adapted for use in the present invention, to detect, quantify or monitor transcription infidelity.

Detection of Transcription Infidelity at the level of proteins

Because transcription infidelity leads to changes in protein sequence, the presence or level of transcription infidelity can also be measured by detecting the presence or amount of TI-proteins.

Various techniques known per se in the art can be used and/or adapted to measure transcription infidelity in proteins. In particular, since transcription infidelity causes modifications of proteins leading to direct changes in their behavior, these changes can be detected by e.g., 2-dimensional gel electrophoresis and mass spectrometry or by surface enhanced laser desorption ionization. As a verification that TI protein are present in human plasma, we show the mass spectrometry profile of a peptide located after the canonical STOP of ApoAII. Mass spectrometry data shows that arginine is substituted to canonical STOP. Furthermore, since transcription infidelity causes the occurrence of longer and shorter (cellular and plasma) protein isoforms (as a result of premature stop codon due to base substitution, or a switch in open reading frame in case of deletions or insertions), such isoforms may be detected or quantified using specific ligands thereof and/or protein sequencing strategies. In this regard, we demonstrate that changes in protein length and AA primary sequence predominantly occur on the carboxy-terminal domain of the protein. Accordingly, sequencing strategies shall be directed mainly towards the C-terminal end of proteins, a previously unsuspected region (indeed, direct protein sequencing methods presently used can only be achieved starting at the NH_2 terminal AA).

Detecting transcription infidelity using specific ligands typically comprises placing a sample in contact with a ligand that is specific for a domain of a TI-protein, and detecting the presence (or amount) of complex formed, said presence or amount being a direct indication of the presence or rate of transcription infidelity. In a preferred embodiment, the method uses a set of ligands that are specific for distinct domains of one or several TI-proteins, respectively. Ligands specific for these domains can be

prepared for any protein using the sequence information and amino acid substitution rules as disclosed in the present application (see e.g., Example 5). The ligand may be used in soluble form, or coated on a surface or support.

In this respect, the invention also relates to any ligand that selectively binds a domain of a TI-protein. Different types of ligands may be contemplated, such as specific antibodies, synthetic molecules, aptamers, peptides, and the like.

In a specific embodiment, the ligand is an antibody, or a fragment or derivative thereof. Accordingly, a particular aspect of this invention resides in an antibody that specifically binds a domain of a TI-protein.

Within the context of this invention, an antibody designates a polyclonal antibody, a monoclonal antibody, as well as fragments or derivatives thereof having substantially the same antigen specificity. Fragments include e.g., Fab, Fab'2, CDR regions, etc. Derivatives include single-chain antibodies, humanized antibodies, human antibodies, poly-functional antibodies, etc.

Antibodies against domains of TI-proteins may be produced by procedures generally known in the art. For example, polyclonal antibodies may be produced by injecting the domain of the TI-protein (e.g., as a protein or peptide) alone or coupled to a suitable protein into a non-human animal. After an appropriate period, the animal is bled, sera recovered and purified by techniques known in the art (Paul, W.E. "Fundamental Immunology" Second Ed. Raven Press, NY, p. 176, 1989; Harlow et *al.* "Antibodies: A laboratory Manual", CSH Press, 1988 ; Ward et al (Nature 341 (1989) 544).
Peptides with the same sequence as TI-proteins may be produced by procedures generally known in the art. Such peptides can be coupled to a suitable support and used for the detection of auto-antibodies present in biological samples (for example body fluids).
Monoclonal antibodies against domains of TI-proteins may be prepared, for example, by the *Kohler-Millstein* (2) technique (Kohler-Millstein, Galfre, G., and Milstein, C, Methods Enz. 73 p. 1 (1981)) involving fusion of an immune B-lymphocyte to myeloma cells. For example, an immunogen as described above can be injected into a non-human

mammal as described. Subsequently, the spleen is removed and fusion with myeloma conducted according to a variety of methods. The resulting hybridoma cells can then be tested for the secretion of antibodies against domains of TI-proteins.

An antibody "selective" for a particular domain of TI-proteins designates an antibody whose binding to said domain (or an epitope-containing fragment thereof) can be reliably discriminated from non-specific binding (i.e., from binding to another antigen, particularly to the native protein not containing said domain). Antibodies selective for domains of TI-proteins allow the detection of the presence of proteins containing such domains in a sample.

Diagnostics

The present invention allows the performance of detection or diagnostic assays that can be used, among other things, to detect the presence, absence, predisposition, risk or severity of a disease from a sample derived from a subject. 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 diseases in a biological sample, preferably, a human biological sample, comprising placing said sample in contact with a ligand that specifically binds a transcription infidelity site and determining the formation of a complex.

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 ligand that specifically binds a transcription infidelity site expressed by cancer cells, and determining the formation of a complex.

Another embodiment of this invention is directed to a method of assessing the response or responsiveness of a subject to a treatment of a cancer, the method comprising

detecting the presence or rate of transcription infidelity in a sample from the subject at different times before and during the course of treatment.

This invention also relates to a method of determining the efficacy of a treatment of a cancer disease, the method comprising (i) providing a tissue sample from the subject during or after said treatment, (ii) determining the presence and/or rate of transcription infidelity in said sample and (iii) comparing said presence and/or rate to the amount of transcription infidelity in a reference sample from said subject taken prior to or at an earlier stage of the treatment.

The presence (or increase) in transcription infidelity 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 nucleic acids or polypeptides, as appropriate. Examples of such samples include body fluids, tissues, cell samples, organs, biopsies, etc. Most preferred samples are blood, plasma, serum, saliva, urine, seminal fluid, and the like. The sample may be treated prior to performing the method, in order to render or improve availability of nucleic acids or polypeptides 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.
A method of the present invention consist at determining the presence in human samples of antibodies directed against novel peptides produced by transcription infidelity and generating immune stimulation leading to antibody production. A second method of this invention is directed at detecting cells bearing immunological structure directed against transcription infidelity peptides.

Drug design and Therapy

As discussed above, the invention allows the design (or screening) of novel drugs by assessing the ability of a candidate molecule to modulate transcription infidelity. Such methods include binding assays and/or functional (activity) assays, and may be performed *in vitro* (e.g., in cell systems or in non cellular assays), in animals, etc.

A particular object of this invention resides in a method of selecting, characterizing, screening or optimizing a biologically active compound, said method comprising determining *in vitro* whether a test compound modulates transcription infidelity. Modulation of transcription infidelity can be assessed with respect to a particular gene or protein, or with respect to a pre-defined set of genes or 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 sites.

A further embodiment of this invention resides in a method of selecting, characterizing, screening and/or optimizing a biologically active compound, said method comprising contacting a test compound with a cell and determining, in said cell, whether the test compound modulates transcription infidelity.

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.

Compounds that modulate transcription infidelity of the present invention have many utilities, such as therapeutic utilities, to reduce or increase transcription infidelity in a cell. Such compounds may be used to treat (or prevent) diseases caused by or associated

with abnormal transcription infidelity, such as proliferative disorders (e.g., cancers), immune diseases, aging, inflammatory diseases, etc.

Other compounds of the present invention are compounds that target transcription infidelity, i.e., compounds that selectively bind transcription infidelity sites. Such compounds (e.g., antibodies, RNAi, etc.), may be used as diagnostic reagents, or as therapeutic agents, either alone or conjugated to a label.

Compounds that modulate or target transcription infidelity of the present invention can be administered by any suitable route, including orally, or by systemic delivery, intravenous, intra-arterial, intra-cerebral or intrathecal injections. The dosage can vary within wide limits and will have to be adjusted to the individual requirements in each particular case, depending upon several factors known to those of ordinary skill in the art. Any pharmaceutically acceptable dosage form known in the art may be used, such as any solution, suspension, powder, gel, etc., including isotonic solution, buffered and saline solutions, etc. The compounds can be administered alone, but are generally administered with a pharmaceutical carrier, with respect to standard pharmaceutical practice (such as described in Remington's Pharmaceutical Sciences, Mack Publishing). Antibodies may be administered according to methods and protocols known per se in the art, which are presently used in human trials and therapies.

Effect of transcription infidelity on normal cell function

It is generally accepted that proteins are responsible for most cell functions. However, it is also clear that RNA, known as non-coding RNA, are therefore also functional molecules by themselves. These transcripts, whose importance was previously underestimated, would be present in many organisms and regulating a lot of cellular functions. Among them, one finds rRNA, tRNA, tmRNA (transmessenger RNA), but also other non-coding RNA (snoRNA, snRNA etc.) intervening in post-transcriptional modifications of RNA, in splicing or another cellular functions. The functions of both RNA and protein can be affected by a lack of transcription fidelity described in this application. The sequence of DNA is transcribed into mRNA in the nucleus by the action of a protein called RNA polymerase II that recognizes the sequence of the DNA

molecule and synthesizes a single strand polymer of nucleotides that is complementary to the mother DNA template. The order and type of RNA base at any given position is determined by the order and type of base present on the DNA strand that serves as template. The 5' end of the transcribed RNA, which corresponds to the first base of mRNA, is linked to 7-methylguanylate attached to the initial nucleotide thereby constituting the 5' CAP that protects RNA from degradation. This modification occurs before transcription is completed. Processing at the 3' end of the primary transcript involves cleavage by an endonuclease to yield a free 3'-hydroxyl group to which a string of adenine residues is added by an enzyme called poly(A) polymerase. The resulting poly(A) tail contains between 100-250 nucleotides. The final step of the processing is splicing, which consists of the cleavage from the primary transcript of the element corresponding to intronic sequences followed by ligation of the exons. This process is controlled by numerous proteins and ncRNA, some of which bind to RNA. It is possible that this binding is one step that can be affected by lack of transcription fidelity. Indeed, all of these proteins are themselves encoded by specific RNA; the function of these proteins is therefore potentially affected by a lack of transcription fidelity. The highly complex process of RNA maturation leads to production of mRNA that will be exported from the nucleus in order for translation to occur. It is important to recognize that, at this stage, not all nucleotide triplets that are present on mature mRNA will be translated into AA. Mature messenger RNA contain non-coding regions referred to as the 5' and 3' untranslated regions that bear functional significance in determining the overall stability of mRNA and other roles in translation. In order to effectively convey the information contained in the DNA into the proper protein AA sequence, absolute fidelity of transcription is required. Indeed, any error occurring either during the transcription or during the maturation processes will potentially result in the introduction of changes in mRNA stability and ultimately in a variation in protein AA primary sequence. These variations in protein sequence might then exacerbate the phenomenon of transcription infidelity by alteration of the sequence of protein involved in initiation of transcription, transcription itself, 5' capping of RNA, 3' polyadenylation, RNA splicing and/or the export of RNA. Thus, the demonstration in this invention of transcription infidelity affecting a great number of genes isolated from cancer cells opens the possibility that the phenomenon might exacerbate itself thereby leading to progressive increase in disease severity. Variation in RNA sequences introduced by

infidelity of transcription may have immediate consequences on cell function. Indeed, introduction of bases into the RNA primary transcript that are not complementary to that of the DNA template has immediate potential consequences on the resulting protein AA primary sequence. Where the change affects the first 2 bases of the codon, it typically exerts a direct impact on the AA sequence. Variations affecting the third base of the codon will have a lower impact on protein AA sequence because the genetic code is degenerated. Thus, the changes of the base located on the third position of the codon may not directly influence the AA sequence of the protein. On the basis of the data described below, we believe that infidelity of transcription is a phenomenon that affects all bases irrespective of their position on the codon, and therefore impacts on protein AA primary sequence. The changes in protein can be either neutral or cause a modification of protein function, whether it be an increase or loss of activity. The phenomenon of transcription infidelity is predominantly observed after encoding at least the first 400-500 bases of mature mRNA is completed; the 5' end of the mRNA are relatively less affected. It is believed that shorter as well as longer fragments of protein will be present in cancer cells as well as in plasma from cancer patients. These shorter and longer isoforms can be directly deduced from transcription infidelity rules described below. This allows the production of rationally designed methods for identification of specific markers of cancer or of immunological disease severity. Because the transcription rate of most genes is controlled to adapt to cellular needs, it is proposed that transcription infidelity causes changes in gene expression directly due to either excessive or lack of protein function. Thus, the phenomenon of transcription infidelity exerts a profound effect on the cell capacity to perform its task. The identification of this defect as being a common feature of most genes isolated from all types of cancer cells therefore provides a rationale to seek new methods allowing quantitative measurement of the rate of transcription infidelity in any given cell. This screening assay allows testing of novel drugs capable of limiting transcription infidelity, thereby preventing progression of the disease and restoring normal cell function.

Transcription infidelity and Pathology

The present invention shows that transcription infidelity leads to important changes in protein sequence. The function of any given protein is determined by its 3-D structure

that is directly dependent on its AA sequence. Proteins with a variant AA, protein truncation or protein elongation might result in either a profound modification of protein activity or remain neutral. Examples of modified proteins that significantly inhibit the function of their normal homologs have been described. Transcription infidelity is a phenomenon that affects a great number of genes, hence yielding a large number of proteins. Because several of these proteins participate in maintaining stable DNA structure and participate in DNA repair, defective function of these proteins may result in defective DNA repair and result in a significantly decreased capacity of the cells to successfully repair any damaged DNA. Because transcription fidelity in normal cells is due to the activity of several protein complexes that control transcription, it is quite possible that initial small alterations of these proteins might progressively exacerbate the phenomenon. This will ultimately result in a greater rate of transcription infidelity that will consequently suppress cell differentiation status and lead to increasingly severe forms of the disease. The demonstration that transcription infidelity actually occurs in cancer cells and leads to wide diversification of proteins encoded by any given gene opens a novel area to screen for novel diagnostic probes and design novel therapeutic targets.

The present invention describes a novel phenomenon that contributes to diversification of the information present on DNA and that follows specific rules. This process of non-random transcription infidelity is greatly exacerbated in cancer cells but also occurs in normal cells following the same rules. This mechanism introduces novel bases in mRNA thereby causing profound changes in the message that will be translated at the ribosome level. The phenomenon is likely to be quite general because it is present in most tested genes. The immediate consequence of transcription infidelity is that a single gene can produce many more proteins than suspected. Thus, our discovery has the potential to explain in part the relative discrepancy between the limited number of genes present on the genome and the large number of proteins present in biological samples. We have shown that transcription infidelity is a non-random process and is governed by specific rules, some of which have already been described here. Some will be revealed by further characterization of the process using bioinformatics and biological methods. We have focused here on the implication of transcription infidelity in cancer. However, we believe that the process is general and that this can contribute to generate diversity

of proteins. This diversification of proteins might exert a significant influence on the capacity of the immune system to recognize a specific protein pattern. In this regard, it is believed that transcription infidelity plays a role in the pathogenesis of immune diseases. We have considered here the events of transcription infidelity that lead to the substitution of single base. We have focused on single base substitutions because this mechanism is unlikely to cause significant destabilization of mRNA structure. Thus, single base substitutions should not interfere with translation, but be recognized as a bona-fide novel message. We will see further that transcription infidelity is also expected to cause base deletions and/or insertions. The same method described above will detect alternate mechanisms of transcription infidelity.

The novel process described in this application has several immediate practical applications.

Optimization of proteomic experiments to identify novel disease-specific biomarkers and design novel antibodies that will be targeted towards specific disease proteins.

Based on the teaching of the present application, it is now possible to predict the changes that can occur in any protein sequence as a result of transcription infidelity, and therefore to predict changes in physico-chemical properties such as apparent molecular masses and isoelectric points. This can yield modified protein patterns in 2-D gels and mass spectrometry. Because transcription infidelity affects most genes, it is now possible to focus on a limited set of proteins and characterize their changes in order to identify disease (e.g. cancer) specific biomarkers. Transcription infidelity algorithms can speed up the biomarkers discovery process. Alternatively, it is possible to design specific antibodies directed toward domains of TI-proteins, i.e, the protein domains that contain modified AA, as well as additional new AA which extend the native protein generated by transcription infidelity. Antibodies directed against these domains will be specific for the protein variant and therefore targeted toward disease (e.g., cancer)-specific proteins present in body fluid, at the cell surface, or inside cancer cells. These antibodies will be used to detect proteins characteristic of a pathological condition in body fluids, to detect diseased cells circulating in plasma, to identify diseased cells in histological samples, to evaluate disease severity, and also to target medications toward

specific diseased cells. This can be achieved by gene transfer of transcription infidelity-prone sequence by means of gene therapy vector e.g. adenoviruses, liposomes, etc. Alternatively, we designed specific peptides having the same sequence as a domain that result from TI. They can be used to detect specific and natural auto-antibodies directed toward domains of TI-proteins present for example in body fluids.

Transcription infidelity occurs non-randomly and frequently affects mRNA stop codons in a great proportion of tested genes. Further, because one can estimate that up to 6% of any given population of mRNA contain these additional sequences, one can therefore detect in a given cancer cell a specific population of protein that exhibits a novel sequence in the carboxy-terminal end. Targeting of these sequences was thus far not conceivable because no process reported, except for rare genetic diseases affecting DNA, was capable of introducing in-frame reading of sequences that immediately follow canonical stop codons. The existence of these hidden sequences was not previously suspected. The discovery of transcription infidelity leading to single base substitution occurring on canonical stop codons reveals the existence of such coding sequences and further show that their presence is increased in cancer cells because of enhanced transcription infidelity in these cells. Thus, it is possible to identify and target these normally hidden sequences to design novel cancer-specific drugs.

TI concerning single base deletion or insertion occurring on the coding sequence leads to a switch in the coding frame. As a convergence, the sequences in the corresponding protein are modified (change in the AA sequence and in protein length). One can therefore detect in a given cancer cell a specific population of protein that exhibits a novel sequence. Thus, it is possible to identify and target these normally hidden sequences to design novel cancer-specific drugs.

Because of the existence of novel or modified protein sequences, and considering that these alterations are present on the protein present on the cell surface, it is proposed to use these sequences in order to vaccinate against disease-specific sequences. These vaccinations will be used to trigger immune responses in patients diagnosed with any given form of disease (e.g., of cancer) or to initiate preventive immune responses in patients with increased risk for developing specific diseases (e.g., cancer) because of

genetic predisposition or because of increased exposure to an environmental toxin or risk.

A particular object of this invention thus relates to a method of detecting the presence or stage of a disease in a subject, the method comprising assessing (*in vitro* or *ex vivo*) the presence or rate of transcription infidelity in a sample from said subject, said presence or rate being an indication of the presence or stage of a disease in said subject. The sample may be any tissue, cell, fluid, biopsy, etc., containing nucleic acids and/or proteins. The sample may be treated prior to the reaction, i.e., by dilution, concentration, lysis, etc. The method typically comprises assessing transcription infidelity of a number of distinct genes or proteins, such as plasma proteins, cell surface proteins, etc.

A further object of this invention is a method of treating a subject in need thereof, the method comprising administering to said subject an effective amount of a compound that alters (e.g., reduces or increases) the rate of transcription infidelity of a mammalian gene.

A further object of this invention relates to the use of a compound that alters (e.g., reduces or increases) the rate of transcription infidelity of a mammalian gene for the manufacture of a pharmaceutical composition for use in a method of treatment of a human or animal, particularly for treating a proliferative cell disorder, such as cancers and immune diseases.

A further object of this invention resides in a method of assessing the efficacy of a drug or candidate drug, the method comprising a step of assessing whether said drug alters the rate of transcription infidelity of a mammalian gene, such an alteration being an indication of drug efficacy.

The invention further relates to methods and products (such as probes, primers, antibodies or derivatives thereof), for detecting or measuring (the level of) transcription infidelity in a sample, as well as to corresponding kits.

The invention also relates to a method of identifying and/or producing biomarkers, the method comprising identifying, in a sample from a subject, the presence of transcription infidelity site(s) in a target protein, RNA or gene and, optionally, determining the sequence of said transcription infidelity site(s). In a particular and preferred embodiment, the target protein is a cell surface protein or a secreted protein, particularly a cell surface protein or a plasma protein.

A further object of the invention relates to a method of producing or identifying ligands specific for a trait or pathological condition, the method involving identifying, in a sample from a subject having said trait or pathological condition, the presence of transcription infidelity site(s) in one or several target proteins, RNA or gene, optionally determining the sequence of said transcription infidelity site(s), and producing (a) ligand(s) that specifically bind(s) to said transcription infidelity site(s).

The invention is particularly suited for identifying biomarkers of cell proliferative disorders, such as cancers, immune diseases, inflammation or aging. It is particularly useful for producing ligands that are specific for such disorders in mammalian subjects, in particular ligands that can detect the presence or severity of a cell proliferative disorder in a subject.

A further aspect of this invention relates to a (synthetic) peptide comprising a transcription infidelity site of a protein, particularly of a mammalian protein, more preferably of a human protein. The peptide typically comprises an internal fragment of protein, or the sequence of a C-terminal fragment of said protein. The peptide preferably comprises less than 100, 80, 75, 70, 65, 60, 50, 45, 40, 35, 30, 25 or even 20 amino acids. The protein may be a cell surface protein (e.g., a receptor, etc.), a secreted protein (e.g., a plasma protein), or an intracellular protein. Examples of such plasma proteins are provided e.g., in Fig. 13, and include apolipoproteins (e.g., AI, AII, CI, CII, CIII, D, E), complement components (e.g., C1s, C3, C7), C-reactive protein, serpin peptidase inhibitors, fibrinogen (e.g., FGA1, FGA2), plasminogen, transferrin, transthyretin, etc. Examples of cell surface receptors include e.g., cytokine receptors and hormone receptors, etc. In a specific embodiment, the invention relates to a synthetic peptide

comprising a transcription infidelity site of an abundant human plasma protein or cell surface receptor as listed above.

Specific examples of such peptides are disclosed in Figures 10, 14 and 18. More specifically, examples of synthetic peptides of the present invention comprise all or a fragment of the following amino acid sequences:

QMWQLFWIYHLSS (TPT1) (SEQ ID NO: 1)KLHTLSAAIYYQQE (VIM) (SEQ ID NO: 2)
DFLSNK (RPS6) (SEQ ID NO: 3)MYTVEFSVHKNN (RPL7A) (SEQ ID NO: 4)NGSLGDMSDLCT (RPS4X) (SEQ ID NO: 5)ASG (FTH1) (SEQ ID NO: 6)
EPSEPSDF (FTL) (SEQ ID NO: 7)
APSIFPTLPAKPGTKQPRSPVTALSLHMLLMVSSAPSCGLIQTVSSFTVYIFTL (TPI1) (SEQ ID NO:
8)ARHGRDEEVWHRKHSHHFVQAWAWVGGLVCWPRKCHMRSTLISSLDSLLPVIPHRTEAEWVV
VMFDRRH (AHSG) (SEQ ID NO: 9)
RSKAYSSVFLFRWCKANTLSKKHKFL (ALB) (SEQ ID NO: 10)
GLDSTRALENEMTV (APCS) (SEQ ID NO: 11)
GARRRPPSRCSE (APOA1) (SEQ ID NO: 12)
SVQTIVFQPQLASRTPTGQS (APOA2) (SEQ ID NO: 13)
IVFQPQLASRTPTGQS (APOA2) (SEQ ID NO: 14)QPDPPSVDKGRVPYSPDPPGSD (APOC2) (SEQ ID NO: 15)
DPPSVDKGRVPYSPD (APOCII) (SEQ ID NO: 16)
DLNTPSPPAYPSCELLGSCNLQGCPCRLLKRDSILSALLPHLMPGPPPGMLASQ (APOC3) (SEQ ID NO: 17)
PGSTGRLHPLHVTSASLSPTPPPPHKDKPINHDKGS (APOD) (SEQ ID NO: 18)
TPKPAAMRPHATPCLLPPRSLQRETLSPPQPSSWGGP (APOE) (SEQ ID NO: 19)
EARVGGNVGSQTQ (AZGP1) (SEQ ID NO: 20)
PSVLHTARGPRMPRPPLAPAGREPDHLPC (CHI3L1) (SEQ ID NO: 21)
DVDVAFAPTGASESSSPQDELQPPRESSARHQVTRPQPPGPQLRPASPRSGSCTLTLDSAAHGK
NRIAPACN (CLU) (SEQ ID NO: 22)
NVIPLKRKMNNTLN (HRG) (SEQ ID NO: 23)
TPAARLMWSSNMPYFAQKTAKDMTSSWLQPRFIFLFVVN (IGFBP3) (SEQ ID NO: 24)
GWGVFLLNPMAGGHAPTIISWEERQSWEIDGSHSSLLSLLCLWATLPTPLLSQ (INHA) (SEQ ID NO: 25)
TGPTHHSPSPSISTWCLVPVHSVNKKP (KLK11) (SEQ ID NO: 26)
WTPEPLLQPLSHPLPPAHPLGQQRL (PKM2) (SEQ ID NO: 27)
LDGRQSDALTHLEAGTWVGI (PLG) (SEQ ID NO: 28)
SLPSSSGALSKELGMQAGCLGLWAQPGPCAPSGHGMCGPVCLSLEGDSDSLCSSHMHRGPWTLQ
SGGSWAS (SERPINA3) (SEQ ID NO: 29)
NLRGRAATKVKMGTQMIHEFALVSLAQVVCANHVCLHSSVLPCVLNKK (TF) (SEQ ID NO: 30)
GPAPPRPAPAGPAPPRPAPAALPMGAVFKDTRAPSPPGAPLKMERGLRISVSLGACLGSPSLTF
PHSHSLSLPLCLLLPVCTIPLPGIKAQGTSGEHYCS (TGFB1) (SEQ ID NO: 31)
GTSPPVDLKDEGWDFM (TTR) (SEQ ID NO: 32)

A further aspect of this invention resides in the use of a peptide comprising a transcription infidelity site, as defined above, as an immunogen. The invention also

relates to a vaccine composition comprising a peptide comprising a transcription infidelity site, as defined above, and optionally a suitable carrier, excipient and/or adjuvant.

The invention also relates to a device or product comprising, immobilized on a support, a reagent that specifically binds a transcription infidelity site. The reagent may be e.g., a probe or an antibody or derivative thereof.

The invention may be used in any mammalian subject, particularly any human subject, to detect, monitor or treat a variety of pathological condition associated with an increased or reduced transcription infidelity rate, such as e.g., cell proliferative disorders (e.g., cancers), immune diseases (e.g., auto-immune diseases (multiple sclerosis, ALS), graft rejection), aging, inflammatory diseases, diabetes, etc., and/or to produce, design or screen therapeutically active drugs. The invention may also be used to detect, monitor, modulate or target transcription infidelity in any other cell type, such as prokaryotic cells, lower eukaryotes (e.g., yeasts), insect cells, plant cells, fungi, etc.

Producing and using agents that target transcription infidelity sites

Using the teaching of the present invention, it is possible to produce agents that can target transcription infidelity sites, e.g., that bind to proteins or nucleic acids containing a transcription infidelity site, or to cells or tissues containing or expressing such proteins or nucleic acids. The targeting agent may be an antibody (or a derivative thereof), a probe, a primer, an aptamer, etc., as disclosed above.

Such targeting agents may be used e.g., as diagnostic agents, to detect, monitor, etc. transcription infidelity in a sample, tissue, subject, etc. For that purpose, the agent may be coupled to a labeling moiety, such as a radiolabel, an enzyme, a fluorescent label, a luminescent dye, etc.

The targeting agents may also be used as a therapeutic molecule, e.g., to treat a cell or tissue or organism exhibiting abnormal transcription infidelity. The agent may be used

as such or, preferably, coupled to an active molecule, such as a toxic molecule, a drug, etc.

A particular embodiment of this invention thus resides in a method of producing an agent that targets transcription infidelity, the method comprising (i) identifying a transcription infidelity site of a protein or nucleic acid and (ii) producing an agent that selectively binds said site. The transcription infidelity site may be identified e.g., by alignment of sequences available for a given RNA molecule and identification of sequence variations, particularly at the 3' end. Transcription infidelity sites may also be identified by applying the transcription infidelity rules, as described in the present application (see e.g., Example 5) or new rules later described, to any gene sequence. The peptide or nucleic acid molecule comprising said identified site may then be produced using conventional methods. In a preferred embodiment, step (i) of the method comprises identifying a transcription infidelity site of a secreted or cell surface protein (e.g., a receptor, etc.), or of a nucleic acid. Indeed, secreted proteins and cell surface proteins can be easily targeted using a targeting agent that is contacted to a cell or administered to a subject. Examples of domains of such TI-proteins are disclosed in the examples.

Producing and using a vaccine that causes or stimulates or inhibits an immune response against a domain of a TI-protein

Using the teaching of the present invention, it is possible to produce a vaccine composition that causes or inhibits an immune response against TI-proteins, or against cells or tissues containing or expressing such TI-proteins or corresponding nucleic acids.

Typically, the vaccine composition comprises, as an immunogen, a molecule comprising the sequence of a transcription infidelity site. The vaccine composition may comprise any pharmaceutically acceptable carrier, excipient or adjuvant. The vaccine composition can be used to generate an immune response against a diseased cell or tissue expressing TI-proteins, which results in a destruction of said cell or tissue by the

immune system. Alternatively, the vaccine composition may be used to induce a tolerance against transcription infidelity sites involved in auto-immune diseases.

A particular embodiment of this invention thus resides in a method of producing an immunogen that causes, stimulates or inhibits an immune response against a domain of TI-proteins, or against a cell or tissue expressing such a protein, the method comprising (i) identifying a transcription infidelity site of a protein or nucleic acid and (ii) producing a peptide comprising such a site or a variant thereof. The transcription infidelity site may be identified e.g., by alignment of sequences available for a given RNA molecule and identification of sequence variations, particularly at the 3' end. Transcription infidelity site may also be identified by applying the transcription infidelity rules, as described in the present application (see e.g., Example 5) or new rules later described, to any gene sequence. The peptide molecule comprising said identified site may then be produced using conventional methods. In a preferred embodiment, step (i) of the method comprises identifying a transcription infidelity site of a secreted or cell surface protein (e.g., a receptor, etc.), or of a nucleic acid. Indeed, secreted proteins and cell surface proteins are more subject to a response by the immune system. Examples of domains of such TI-proteins are disclosed in the examples.

Reducing Transcription Infidelity in Recombinant Protein Production Systems

A particular object of this invention resides in a method of preventing or reducing transcription infidelity that can occur in recombinant production systems, particularly in therapeutic proteins production systems. Indeed, such a transcription infidelity may either decrease the yield of production of the system, or result in the production of mixtures on uncharacterized proteins which may exhibit various activity profiles. It is therefore recommended and highly valuable to be able to reduce the occurrence or rate of transcription infidelity in recombinant production systems, such as bacterial production systems (if transcription infidelity rules are the same for prokaryotic systems), as eukaryotic production systems (e.g., yeasts, mammalian cells, etc). Such a reduction may be accomplished e.g., by adapting the sequence of the coding nucleic acid molecule, and/or by inhibiting RNA molecules generated through transcription

infidelity. Alternatively, proteins containing transcription infidelity sites may be removed from the medium.

Reducing the sensibility to Transcription Infidelity for a gene

A particular object of this invention resides in a method of preventing or reducing TI that can occur in a gene. Indeed, such TI may affect either the expression or the activity of a protein, or result in the production of mixtures of uncharacterized proteins which may exhibit various activity profiles. It is therefore recommended and highly valuable to be able to reduce the occurrence or rate of TI for said gene. Such a reduction may be accomplished e.g., by modifying the sequence of the gene (gene therapy), and/or by inhibiting RNA molecules generated through transcription infidelity. Alternatively, proteins containing TI sites may be specifically degraded (e.g. specifically targeted for degradation).

Measuring transcription infidelity at the RNA level.

Transcriptomics is the science that measures variation in RNA (preferentially mRNA) levels occurring in a variety of pathological conditions, of which cancer is the most frequent. Transcriptomics relies on hybridization of cDNA with a specific set of oligonucleotides that identify predefined subsets of genes. Hybridization efficacy is dependent on specific sequences of any given RNA that will be reverse transcribed in cDNA. Introduction of unsuspected sequence variation in a given RNA will decrease the efficacy of hybridization and hence cause variation in the signal perceived by transcriptomic chips. The discovery that transcription infidelity causes alterations of RNA base sequence has two immediate consequences: first, it allows optimization of transcriptomic chips in order to minimize the consequence of base mismatch due to transcription infidelity and thereby improve the accuracy of the transcriptomic experiment. Indeed, the current limitation of transcriptomic experiments is their lack of reproducibility between studies. We currently believe that a major part of said variability is caused by transcription infidelity. Understanding the rules of transcription infidelity as described above now allows the design of microarray chips that specifically measure the rate of transcription infidelity in any given mixture of cDNA obtained from

normal or diseased cells. Monitoring of transcription infidelity rate at the RNA level allows high-throughput screening of the relative rate of transcription infidelity occurring in a given cell. This provides essential information to determine disease severity, define the therapeutic strategies and classify diseased cells according to their pharmacological sensitization profile. The same screen also allows testing for the efficacy of novel drugs, e.g., in cancer therapy.

Further applications of transcription infidelity

Transcription infidelity is a newly discovered natural mechanism that adds diversification to the well-established genetic code. We have observed that transcription infidelity occurs, albeit at low rates, even in normal tissues. We propose that transcription infidelity contributes to explain the relative low abundance of genes identified on the genome and the much greater diversity of proteins present in living cells of biological fluids and tissues.

We propose that transcription infidelity serves a specific function at the immunological level. We further propose that the immune system relies at least in part on a given rate of transcription infidelity affecting preferentially specific genes in order to identify self. Thus, abnormalities in the rate of transcription infidelity occurring for yet unknown reasons might contribute to trigger specific immune response and contribute to the pathogenesis of auto-immune and allergic diseases. Inter-individual differences in transcription infidelity might also be conditioned by polymorphisms of genes of the immune system, hence determining the rate of transcription infidelity occurring in circulating cells of the immune system, e.g., T or B lymphocytes, and might also contribute in the evaluation of the suitability of donor and acceptor of grafts and in determining the severity of graft versus host severity of diseases.

We also postulate that transcription infidelity occurs at greater rate during the normal aging process. This would create conditions of progressive reduction of performance affecting all enzymes generally and most specifically those proteins that are involved in cell replication, DNA repair and maintenance of normal rate of transcription fidelity. The concept described above might therefore redirect research onto the mechanisms of

aging. Biochemical, proteomic, transcriptomic and other methods that allow screening for mRNA and cDNA sequences and fidelity to that specified by DNA might therefore be used to quantify the rate of transcription infidelity occurring in a given individual. Testing of novel molecules reducing the rate of transcription infidelity will therefore be amenable to biological screening and lead to development of drugs reducing the aging rate.

Further aspects and advantages of the invention will be disclosed in the following experimental section, which illustrates the invention and does not limit the scope of the present application.

Experimental Section

Example 1. Principle of typical cDNA library construction and sequencing (see Fig 1)

The first step in preparing a complementary DNA (cDNA) library is to isolate the mature mRNA from the cell or tissue type of interest. Because of their poly(A) tail, it is straightforward to obtain a mixture of all cell mRNA by hybridization with complementary oligo dT linked covalently to a matrix. The bound mRNA is then eluted with a low salt buffer. The poly(A) tail of mRNA is then allowed to hybridize with oligo dT in the presence of a reverse transcriptase, an enzyme that synthesizes a complementary DNA strand from the mRNA template. This yields double strand nucleotides containing the original mRNA template and its complementary DNA sequence. Single strand DNA is next obtained by removing the RNA strand by alkali treatment or by the action of RNase H. A series of dG is then added to the 3' end of single strand DNA by the action of an enzyme called terminal transferase, a DNA polymerase that does not require a template but adds deoxyoligonucleotide to the free 3' end of each cDNA strand. The oligo dG is allowed to hybridize with oligo dC, which acts as a primer to synthesize, by the DNA polymerase, a DNA strand complementary to the original cDNA strand. These reactions produce a complete double strand DNA molecule corresponding to the mRNA molecules found in the original preparation. Each of these double strand DNA molecules are commonly referred to as cDNA, each containing an oligo dC - oligo dG double strand on one end and an oligo dT - oligo dA

double strand region on the other end. This DNA is then protected by methylation at restriction sites. Short restriction linkers are then ligated to both ends. These are double strand synthetic DNA segments that contain the recognition site for a particular restriction enzyme. The ligation is carried out by DNA ligase from bacteriophage T4 which can join "blunt ended" double strand DNA molecules. The resulting double strand blunt ended DNA with a restriction site at each extremity is then treated with restriction enzyme that creates a sticky end. The final step in construction of cDNA libraries is ligation of the restriction cleaved double strand with a specific plasmid that is transfected into a bacterium. Recombinant bacteria are then grown to produce a library of plasmids - in the presence of antibiotics corresponding to the specific antibiotic resistance of the plasmid. Each clone carries a cDNA derived from a single part of mRNA. Each of these clones is then isolated and sequenced using classical sequencing methods. A typical run of sequencing starts at the insertion site and yields 400 to 800 base pair sequences for each clone. This sequence serves as a template to start the second run of sequencing. This forward progression leads to progressive sequencing of the entire plasmid insert. The results of sequencing of numerous cDNA designated ESTs have been deposited in several public databases.

Example 2. Database annotation

EST databases contain sequence information that correspond to the cDNA sequence obtained from cDNA libraries and therefore correspond essentially to the sequence of individual mRNA present at any given time in the tissue that was used to produce these libraries. The quality of these sequences has been called into question for several reasons. First, as discussed above, the process of producing cDNA libraries initially relied heavily on the presence of a poly(A) tail at the 3' end of eukaryotic mRNA. Second, mRNA are quite fragile molecules that are easily digested by high abundance nucleases called RNases. Third, while building and sequencing these libraries, little attention was paid to the quality of the original material used and its storage. Because of this, EST sequences have been used to annotate genomic information i.e., to determine whether an identified and fully sequenced segment of genomic DNA encodes any specific mRNA. In this context, EST sequences were useful in order to identify coding genomic sequence. However, little attention has been paid to the information borne by

the EST sequence itself. Indeed, DNA genomic sequence is considered as much more reliable with strong technical arguments in support of this position. We speculate that diversity included in EST sequences might contain biologically, analytically or clinically relevant information. Indeed, EST databases were produced by a number of investigators that all used various methods: this led us to speculate that each methodological bias must contribute to a background noise level with a certain number of errors. However, if differences in errors were to exist due to the source of material used to generate the library, then the difference in error rate would be directly related to the underlying source.

In order to test this, we retrieved ***est_human*** database available at the NCBI ftp site. We selected these databases because these sequences were not annotated or cured by human or bioinformatic tools.

We used a library identification system in order to determine whether an EST was obtained from a cancerous tissue or a normal one. Each library has been labeled ***normal*** or ***cancerous.*** By matching the accession number of each EST with the identifier of the corresponding library, we classified 2.6 millions ESTs as those obtained from cancerous tissues and 2.8 millions ESTs as those obtained from normal tissues.

In order to obtain EST alignments, we used publicly available megaBLAST 2.2.13 software (*Basic Local Alignment Search Tool, Zheng Zhang, A greedy algorithm for nucleotide sequence alignement search, J Comput Biol, 2000)* and specified the following selection parameters:

-d est_human : database to search against (all human ESTs from dbEST),
-i sequences.fasta : FASTA file format containing the reference sequence corresponding to abundantly expressed genes (***Fig 2***) in,
-D 2 : traditional blast output,
-q -2 : Score of -2 for a nucleotide mismatch,
-r 1 : Score of 1 for single nucleotide match,
-b 100000 : maximal number of sequences for which the alignment is reported,
-p 90 : minimal percent identity between EST and the reference sequence,
-S 3 : takes into account ESTs that match reference sequence,
-W 16 : length of best perfect match to start with alignment extension

-e 10 : Expectation value (E-value), number of alignments expected to be found merely by chance for a given sequence and a given database, according to the stochastic model of Karlin and Altschul (1990),

-F F : none filtering the reference sequence.

-o NonFiltre_test_90.out : megaBLAST report outfile

-v 0 : maximal number of sequences to show one line description for,

-R T : (T for True) : report the log information at the end of the megaBLAST output,

-I T : show GI's in deflines [T/F]

The command line for the blast is :

megablast -d est_human -i C:\SeqRef\TPT1\sequences.fasta -o C:\blast\NonFiltre_test_90.out -R T -D 2 -I T -q -2 -r 1 -v 0 -b 100000 -p 90.00 -S 3 -W 16 -e 10.0 -F F

Figure 2(a-q) shows the 17 sequences used for the analysis ; in order to avoid distorting the blast, the polyA tail of mRNA reference sequences were systematically removed. Figure 2r provides a list of genes that were used to test the occurrence of sequence variations.

Example 3. Identification of sequence variations between ESTs from normal and cancer origins.

17 genes are selected on the basis of their large representation in the database. Each EST sequence was then aligned against its curated mRNA reference sequence (RefSeq) from NCBI using the MegaBLAST. This creates a matrix in which any given base is defined by the number of ESTs matching to this position. We then measured the proportion of ESTs deviating from RefSeq at any given position. Comparison of aligned EST sequences according to the tissue origin led us to identify sequence variations occurring at each base position in either the cancer or the normal group. Figure 4 provides a graphical representation of these variations occurring in the normal and cancer sets for the 17 selected genes. Visual examination of the data revealed that sequence variation occurred most frequently in the cancer set, and further that the phenomenon appeared most predominant in specific mRNA sites. The majority of the observations are located at least 400-500 bases after beginning of the mRNA sequence hence predominantly in the 3' part of the gene. The very high number of variations

could not be explained by SNPs. Putative SNPs (□) and biologically validated SNPs (○) are shown on the graph; these SNPs were obtained from the NCBI database (dbSNP, build 126, September 2006) (Sherry, S. T., et *al.* (2001) Nucleic Acids Res 29, 308-11). Both putative and biologically validated SNPs leading to EST variations (n = 442) were excluded from further analysis (therefore RNA editing sites (false SNPs) were also excluded).

The next step of this analysis was to test the statistical significance of the differences in sequence variation occurring between cancer and normal ESTs. For each position, we compared the proportion of the RefSeq base to that of the three other bases between normal and cancer groups using a two-sided proportion test. This test was systematically applied provided that the following conditions were met: $n > 70$ and $(n_{i.} * n_{.j})/n > 5$ $i=1,2$; $j=1,2$ (where n = the number of cancer and normal ESTs for a gene, $n_{1.}$ = the number of cancer ESTs, $n_{2.}$ = the number of normal ESTs, $n_{.1}$ = the number of ESTs having the RefSeq, $n_{.2}$ = the number of ESTs having a substitution). A statistical test is said to be positive at the threshold level of 5% whenever corresponding P-value is lower than 0.05; in this case, the null hypothesis (i.e both proportions are the same) is rejected.

Moreover the two following one-sided proportion tests were considered in order to precise in which set the variability was bigger. The first one allowed to conclude that variabilities were different in both groups when statistical test is positive, then it measured in this case whether variability was statistically greater in the cancer set. On the contrary the second test verified the hypothesis that variability was significantly higher in the normal set. Both one-sided tests were performed whenever the same conditions than two-sided test ones were met.

Figure 3 provides a description of typical blast results. As shown in Figures 5a and 5b, for representative gene TPT1, the number of ESTs at any given position on the gene is similar in both cancer and normal groups. Figure 5c shows that 489 out of 830 possible proportion tests meet the assigned criteria. Proportion tests that were statistically significant at the level of 5 % are shown in Figure 5d. An estimated error resulting from multiple testing, defined by the Location Based Estimator (C.Dalmasso, P.Broet (2005) Journal de la Société Française de Statistique, tome 146, n1-2, 2005) can be calculated. 26 proportion tests positives are due to base substitutions occurring in the normal tissue

(N>C) in figure 5d (n = 26; LBE = 33). This contrasts with the accumulation of statistically significant proportion tests occurring because of sequence variations in the cancer group (C>N), (n= 145; LBE = 15). A similar analysis was conducted for a second gene VIM (***Fig 5e-5h)***. Again the number of ESTs at any given gene position is super-imposable in cancer and normal groups. 752 out of 1847 proportion tests matched the criteria. Again, we observed a large difference in the number of positive proportion tests : n= 78 variations are due to the normal group (LBE = 50) and n=269 variations are due to the cancer group (LBE = 24). We repeated the same analysis on 17 genes that are abundantly present in the EST database. ***Fig 5i*** is a summary of the statistical results obtained. Out of 17 tested genes, 15 exhibit greater sequence variations in ESTs obtained from cancer tissues (***Fig 5j***). At this stage, statistical analysis covered only 9 to 91% of the positions of each gene (an average of 32%) due to constraints of the statistical test.

The conclusion of this first round of analysis is that ESTs of the same gene were different when compared according to the source of tissue (normal versus cancer) from which mRNA was extracted. Assuming that the rate of technical errors that yield the EST sequence variation was not different according to the normal or cancerous origins of tissue and considering the fact that 15 out of 17 tested genes showed sequence variations due to the source of tissue, we propose that these differences directly result from the status –normal versus cancerous– of the cells from which the ESTs were produced.

Similar experiments are performed to study insertions and deletions. Results are obtained after applying the F3 filter (see Example 4). A more stringent condition was used to avoid non biological events (sequencing errors, Megablast misalignment…): a gap or an insertion was only considered if no other modifications were present in a -10/+10 window. For each position, we compared the proportion of windows with deletion (or insertion) to other windows between normal and cancer groups using a two-sided proportion test as described above. ***Fig 5k*** is a summary of the statistical results obtained. Out of 17 tested genes, 14 (or 10 for insertion) exhibit greater sequence variations (deletions and insertions) in ESTs obtained from cancer tissues (***Fig 5l***).

Example 4: Verification of the variation excess in cancer set C>N versus normal set N>C

Libraries originating from cancer or from normal samples are processed essentially in the same manner. Therefore, random errors resulting from library construction or clone sequencing are expected to occur at the same rates in both sets and cannot account for the observed differences between the normal and cancer EST groups. Mathematical analysis is consistent with this interpretation (Brulliard M., et *al* PNAS **May 1, 2007** vol. 104 no. 18 **7522-7527** - see Supplementary material Fig.7).

We next searched to eliminate other sources of EST heterogeneity by sequentially applying filtering procedures based on the following rationale. Our initial requirements were that EST aligned to RefSeq with 100% identity on at least 16 consecutive bases, and with ≥ 90% identity on at least 50 bases. As shown in Figure 6b, this yields, when comparing cancer and normal sets for all 17 genes, 2281 and 725 statistically significant differences C>N and N>C respectively (column F1), and distinct from putative or biologically validated SNPs. The second filter (F2) required that each EST aligned to RefSeq continuously on more than 70% of its length. The third filter (F3) removed ESTs with sequence more closely related to paralogues and pseudogenes than to the *bona fide* RefSeq. The fourth filter (F4) deleted from analysis the first and last 50 bases of each EST alignment. We used this filter in order to remove mismatches at the 3' and 5' borders of EST that can be created by the MegaBLAST program to further optimize alignments and/or resulting from error accumulation at aligned EST extremities. The last filter (F5) normalized EST sequence lengths for each gene in order to remove any difference in length between normal and cancer sets. Indeed, we noticed after applying the first 4 filters, the average cancer EST length was greater than that of normal ESTs (640 ± 248 and 554 ± 229 bases, respectively). However, we also observed significantly greater cancer EST heterogeneity in 5 genes where the average EST length was not different between the normal and cancer sets (TPT1, VIM, HSPA8, LDHA, CALM2).

The number of statistically significant C>N tests remained greater than the number of N>C tests, but the ratio C/N decreased from 3.15 (F1) to 2.05 (F5) (***Fig 6c***).

To further ascertain that cancer EST heterogeneity was not due to accumulation of errors at the end of sequencing runs, we took into account only the information provided after the first 50 bases and no longer than 450 bases of any sequencing run. After this drastic filtering, statistically significant C>N tests were 455 (LBE = 92) and N>C tests were 292 (LBE = 119) for all 17 genes. It can therefore reasonably be assumed that sequence variations causing increased EST heterogeneity in the cancer set are a direct reflection of cancer cell mRNA heterogeneity (***Fig 6d***).

We next independently verified that a greater statistically significant EST heterogeneity persisted in the cancer set after removing ESTs produced from cultured cancer cells. After this filtering, statistically significant C>N tests were 1009 (LBE = 117) and N>C tests were 445 (LBE = 193) for all 17 genes (***Fig 6e***).

Example 5. Breaking the code of base substitution occurring because of transcription infidelity in cancer cells.

Our next goal was to determine whether the described phenomenon of base substitution due to transcription infidelity is a random phenomenon or follows specific rules. To achieve this, we focused on EST variations where C>N was statistically significant. To avoid bias that might be introduced by the filtering procedures, we used all available non-filtered data.The first indication that transcription infidelity is not a random process was that base substitutions rarely occurred in the 5' region of the tested genes. For all tested genes, we very rarely observed the phenomenon of base substitution occurring in the first 400-500 bases of the mature mRNA. The second observation indicating that transcription infidelity is not random stems from the observation that there is a difference in base composition on genomic DNA template observed when comparing sequences upstream and downstream of substitution events (heterogeneous, H, n = 2281) with those where no substitution event was detected (non heterogeneous, NH, n = 12273). The criteria for NH sites were cancer set variations lower than 0.5% and not statistically different from normal set variations (***Fig 7a***). In this analysis, we refer to the base undergoing substitution as b0, bases located on pmRNA 5' end are referred to as b-n, and bases located on 3' end as b+n. For the sake of clarity, we refer in this analysis to the base composition of pmRNA: this corresponds on the DNA to the non-

template strand (the strand not transcribed by RNAP). The data show first that not all 4 bases were equally susceptible to variation: b0 = A (33%) ≈ T (32%) >> C (21%) >> G (14%). Further, the compositions of the 4 bases upstream and 3 bases downstream of the site of event were statistically different (results of Chi-squared analysis) from those of the sites without significant EST variation (***Fig 7a***, the base composition is expressed in % and Gray shading show enriched bases; darker gray show paucity bases). Specifically, sites where variations occur were more frequently preceded and followed by A ≥ G > T ≈ C.

Thus, the occurrence of cancer EST heterogeneity is not random, but determined first by the nature of the base undergoing substitution and second by the nature of the bases that immediately precede and follow the event.

We next questioned whether the replacement base was selected randomly. It is clear from Figure 7b that it is not the case. Statistical significance of difference in proportions was calculated with the null hypothesis that replacement base is selected randomly (adjustment test of all three replacement bases to uniform distribution). A was preferentially replaced by C ($p = 2.8x10^{-125}$), T by G ($p = 5.7x10^{-29}$) and G by A ($p = 2.2x10^{-32}$). Substitution of C showed a more even distribution, with a slight paucity of T ($p = 0.007$).

We then sought for the causes underlying preferential base replacement. To achieve this, we distinguished two sets of informative and non-informative events.

Informative events were situations where the substituted base was different from either preceding base (b-1) or following base (b+1) (n = 1676) (***Fig 7c***). Non-informative events were situations not matching these criteria. When informative events were analysed, two cases were encountered: substituted base was replaced by b-1 or b+1 (79%) or by another base, different from b-1 and b+1 (21%). 1) In the first subset, replacement base was identical to b-1 (n = 799; ***Fig 7c*** panel B) or b+1 (n = 530; ***Fig 7c*** panel C). When replacement base was b-1, then b0 = A (36%) > C (30%) >> T (21%) >> G (13%) (***Fig 7d*** panel A). A was preferentially replaced by C (71% of the cases). When replacement base was b+1, then b0 = T (47%) >> A (21%) > C (19%) >> G (13%) (***Fig 7a*** panel B). T was preferentially replaced by G (71%). Interestingly, the statistical importance of the surrounding bases was also different (***Fig 7d*** panel A and

7d panel B). For b-1 substitutions, the pattern of relative influence of base composition was b0 > b-1 > b-2 > b+1 >b-3 > b+2. For b+1 substitutions, the relative influence of surrounding base followed a pattern of b0 > b+1 > b-1 > b+2 > b-3 > b-2. 2) In the second subset of informative events, the replacement base did not correspond to either b-1 or b+1 (n = 347; ***Fig 7c*** panel D). Affected bases were in the following order: A (47%) > T (29%) > C (14%) > G (10%). A was most commonly replaced by C (91% of the cases), T by C (50%) and A (42%), C by G (46%) and G by C (73%). Thus, when replacement base does not correspond to b-1 or b+1, the replacement base is not randomly selected but C is in large excess.

We next considered the set of non-informative events, i.e., situations where 1) b-1 = b+1 and where 2) b-1 = b0 = b+1 (***Fig 7c***). When b-1 and b+1 were identical but different from b0 (n = 339; ***Fig 7c*** panel E), substituted bases were in the following order: T (34.8%) > G (23.6%) > C (21.2%) > A (20.4%) and followed the same pattern of preference as in Figure 7b: T → G, G → A, A → C. Substitutions occurring on the central base of repeat of three identical bases (n = 266; ***Fig 7c*** panel F) were observed in the following order: A (46.2%) > T (36.9%) > G (10.5%) > C (6.4%). In this case, most common substitution events were: A → C and T → C and A. Rare GGG substitutions were most commonly replaced by GCG and CCC by CAC (***Fig 7c***).

Thus, when substitutions occur within three consecutive identical bases and when substitutions do not correspond to either b-1 or b+1, then C is the most common replacement base (***Fig 7c***). When the replacement base corresponds to b-1, most common substitution is A → C; when the replacement base corresponds to b+1, most common substitution is T → G.

It can therefore be concluded that neither the base undergoing substitution nor the replacement base are selected randomly. Both phenomena follow predictable patterns defined by the composition of the base undergoing substitution and that of bases located upstream and downstream of this event.

Specific algorithms can be defined to identify precisely the motif composition that determines the occurrence of base substitution in any given sequence.

We next separated the set C>N in two sets where N is stable (no deviation observed at the same position in the Normal set) or not. Deviation is considered only if it exceeds a certain threshold defined as the average percentage of deviation in the Normal set.

Figure 7e shows that the -4/+3 signature is carried by the C>N set where N deviates. Context length is less important in the other set. Figure 7f shows an accumulation in the 500-1000 interval that is more important in the C>N set where N deviates than in C>N set where N is stable. In the latter, it seems that earlier events (0-500 interval) appeared. Replacement bases for C>N sites were similar between the two sets (***Fig 7g***).

In case of deletion or insertion, results show that omitted bases (398 C>N cases) were in the following order: C (46.0%) > T (37.9%) >> A (9.3%) > G (6.8%) (***Fig 5m***), and that inserted bases (225 C>N cases) were in the following order: G (36.0%) > C (30.2%) > A (21.8%) > T (12.0%) (***Fig 5o***). We observed that deletion or insertion events often occur in conserved identical bases. A specific program designed to analyze such events shows effectively that 94.7% of deletions and 81% of insertions occur in stretches of variables length. Stretches can be bipolar and symmetrical (e.g. AAATAA) or not (e.g. CCCGG) or single nucleotides repeats (e.g. AAA) (see motifs on ***Fig 5m-5p***). In the deletion case, omitted bases are identical to the b-1 or b+1 base (stretch) in 84% of cases. Most motifs are doublet or triplet of C > T > A~G. Similar results are observed in the case of insertion. In fact, stretch motifs did not appear to be different between C>N and N>C. C>N or N>C specificity would come from other information around the stretches. Additional analyzes on more genes will confirm these observations.

Taking into account all 17 genes, we observed EST heterogeneity at the rate of 10 per 100 bases (***Fig 5j*** and ***Fig 8a – c***). This rate is in excess of any described rate of mutation occurring in genomic DNA. As a reference, one can estimate that single nucleotide polymorphisms occur randomly in the genome once every 300 bases. The rate of single nucleotide polymorphisms affecting transcribed DNA is much lower: it is estimated to occur once every 3000 bases. It is clear that DNA mutations occur more frequently in cancer. However, in depth efforts of breast and colon cancer DNA sequencing that included 14 out of 17 genes used in our study led to an estimated somatic mutation rate of 3.1 mutations per 10^6 bases (Sjoblom, T., et *al.* (2006) Science 314, 268-74). Therefore, transcription infidelity that is described in this invention occurs at a rate far greater than that of mutations affecting DNA. More importantly, most base substitutions on DNA have limited consequences at the protein level because less than 10% of genomic DNA is transcribed to mRNA. In contrast, base substitution due to transcription infidelity often has direct consequences on protein function. Indeed,

1179 out of 2281 substitutions described here (1548 CDS - 369 silent substitutions) led to base substitutions with immediate impact on protein AA primary sequence (***Fig 8d***). Most importantly, significant base substitutions affecting the stop codon were observed in 9 out of 17 tested genes. Before the concept of transcription infidelity, it had not, however, been proposed that human proteins would contain additional coding sequences encoded by RNA sequences considered thus far as "untranslated regions". We now show that base substitution occurring in natural stop codons because of transcription infidelity reveals novel coding regions that encode specific AA. This novel coding region is in phase with the native open reading frame. The natural stop codon is transformed into a coding codon. The next triplet of base is then read as an AA and the translation proceeds with a novel coding region until a new stop codon is reached. We have verified that this is indeed the case in 8 of 17 genes that demonstrated transcription infidelity. All 8 contain alternative stop codons in frame with the corresponding RefSeq (GAPDH does not contain an alternative STOP codon). This has immediate consequence because in each case, it created novel coding sequences of 14, 7, 13, 15, 13, 4, 9, 55 AA in TPT1, RPS6, RPL7A, VIM, RPS4X, FTH1, FTL and TPI1 respectively ***(Fig 10)***. The addition of these AA has the potential to create motifs that will be greatly enhanced in cancer; these motifs will or will not result in novel function of the proteins. Predicting this occurrence leads to possible development of useful tools that could be use in diagnostic, therapeutic or other goals. Predicting this occurrence leads also to possible development of specific antibodies that will recognize cancer specific sequences in the carboxy-terminal end of the protein. No analytical method is currently capable of direct protein sequencing at the carboxy-terminal end. It is, however, possible to cleave proteins enzymatically and sequence cleavage products from their NH_2 terminal end. It is also possible to analyze the AA content of peptides generated by proteolysis using mass spectrometry. We further show that these alternative stop codons are also affected by transcription infidelity (7/9 genes have the second alternative STOP codon affected). The same phenomenon described above can further expand the reading to a novel set of sequences. Annotation of all protein sequences using our method will reveal several unsuspected coding mRNA sequences that will be more or less effectively transcribed depending on codon usage as well as the ability of translation machinery to or not to correctly translate the base substitution. In fact, base substitutions can lead to mRNA structure changes, which can modify

ribosomal reading rate. Nevertheless, we assume that base substitutions don't involve RNA structure changes. On the basis of the occurrence of stop codon alterations, we estimate in affected genes that up to 4% of mRNA in cancer tissues contain these additional coding regions.

A specific program based on several filters can be used to annotate all protein sequences for the presence of a putative Post STOP Peptide (PSP). After retrieving nucleic sequence corresponding to the studied proteins, the program searches the presence or not of an in phase nucleic sequence after the canonical STOP, with another STOP in phase (the possibility to bypass one or more STOP in case of transcription infidelity affecting these alternative STOP codons can be take into account). A minimal length can be fixed (e.g. only sequences coding more than 12 amino acids, but eventually, the peptide length could be smaller) according to minimal criterion for a potential antigenic pattern prediction. Antigenic PSP are then stored (***Fig 18***). An additional step is used to validate the candidates with the annotation by training learning machine of the canonical STOP: indeed the probability that one or more bases of the STOP codon could be substituted by transcription infidelity can be determined. It is also possible to analyze the AA content of peptides generated by proteolysis using mass spectrometry.

We have also shown that besides creating conditions allowing translation of novel protein sequences, transcription infidelity can introduce premature stops in mRNA. Twenty-four novel stop codons occurring within the canonical open reading frame were identified within 13 out of 17 genes. This indicated that transcription infidelity could yield the production of shorter proteins that are lacking specific domains. These truncated proteins might result in an increase or loss of function. The 3-D structure of the protein is likely to be affected and would thus create novel entities that might be recognized by the immune system.

Finally, transcription infidelity in the 17 tested genes revealed that 50% of all identified C>N substitutions lead to AA changes. 17 % correspond to replacement of AA by another from the same family, and 33 % correspond to substitutions of AA from a different class of AA. Thus, transcription infidelity is capable of generating proteins with novel AA sequences with potentially modified functions or activity. Predicting

transcription infidelity using the rules described above will allow the rational prediction of changes in protein behavior and the outcome of proteomic experiments.

Example 6. Biological validations

Biological validations are performed at two levels: mRNA and proteins.

First, mRNA substitutions of the 17 genes of interest will be detected in human cancerous tissues. We used DHPLC (Denaturing High Performance Liquid Chromatography), which is a large scale chromatographic method to detect sequence mutations. The principle of the experiment is described in Figure 11a-c. First, we have developed a method to test the *Transgenomic DHPLC* threshold in order to estimate the percentage of mutated DNA in the sample which is sufficient to allow heteroduplex formation and detection. Indeed, 300bp PCR fragments with 1 and 3 bases substitutions were used. Different ratios of these fragments were prepared: 0%, 2.5%, 5%, 7.5%, 10%, 20% or 50%. The DHPLC allowed us to distinguish [normal] and [normal plus mutated] DNA as soon as the sample contained 2.5 to 5% mutated DNA (***Fig 11d***).

These results indicate that we are able to distinguish mRNA from normal and cancerous tissues for the genes of interest. mRNA extracted from normal and cancerous adjacent tissues (*Biochain Inc*) were used to test three genes: GAPDH, ENO1 and TMSB4X. As DHPLC works on DNA, mRNA samples are converted into cDNA using reverse transcriptase. We chose regions that exhibit far more significant substitutions in ESTs coming from cancer tissues than in ESTs coming from normal tissues. Primers used for amplification are shown in Figure 12.

In a first set of experiment, several cDNA from cancerous and adjacent normal tissue (liver, kidney, breast and colon) were amplified by PCR with said primers and injected on the DHPLC system. The temperature of the oven was selected with the Navigator software (Transgenomic). Profiles were obtained for the genes described above and a representative experiment is presented in figure 15 for the ENO1, the GAPDH and the TMSB4X genes. As shown in figure 15a to 15c, cancer profiles are clearly different from normal profiles for GAPDH and ENO1 genes. No differences were observed for the TMSB4X gene as we expected (far less transcription infidelity sites). The injections of the same PCR product and of 2 other PCR products were done in triplicate (figure

15b and 15c) and the profiles are very reproducible in the same experiment. However, the nature of the difference cannot be deduced from this experiment.

Consequently, the continuation of this biological validation based on mRNA is performed. In fact, sequencing of PCR products issued from cancerous tissues can allow to precisely detect most abundant mutations.

Classical Sanger sequencing method of reverse transcribed PCR amplified mRNA does not detect sequence variants occurring at rates lower than 15-30% at a specific position (***Fig 16***). Indeed, mutated base calling obtained by sequencing amplification products mixes from mutated and non mutated known sequences is obtained for 50-50% mixes and minor peaks are detected from 15%. Pyrosequencing and Emulsion PCR are more sensitive methods allowing the detection of cDNA heterogeneity, thereby analysis of RT-PCR products obtained from cancer and normal cells is possible (Thomas, R. K, et *al.* (2006) *Nat Med* **12,** 852-5).

A gene affected by transcription infidelity is cloned either entirely or a fragment comprising and not comprising its transcription infidelity site. The construct with and without canonical stop codon is ligated in frame with the gene encoding neomycin resistance. Cancer and normal cells are transfected with this chimeric construct first transiently then stably. We predict that transcription infidelity leads to change the canonical stop codon, thereby allowing translation of neomycin gene and creating neomycin resistant cell. We predict that the cancer cells that are neomycin resistant will grow more rapidly and that the shape of these cells will differ significantly from that of normal cells. Moreover, we predict that these cells are more invasive and can be compared to cells at a later stage of cancer progression. Therefore, this technique may be used to determine cancerous phase of different cells from an individual patient.

The final proof that neomycin resistance occurs as a result of transcription infidelity will be obtain by sequencing the inserted construct amplified from genomic DNA and showing that genomic information remains unchanged. We will also verify that mRNA from neomycin resistant cell contain a mutated stop codon due to transcription infidelity.

One technique that permits the detection of transcription infidelity sites is the construction of cDNA libraries that consists in cloning cDNA reverse-transcribed from RNA extracted from cancerous and normal patient tissues. Each cDNA amplified or not from a specific gene are cloned in separate plasmids that are transformed in *Escherichia coli*. Different *E. coli* colonies are picked and the plasmids are sequenced. The number of clones needed to be sequenced depends on the percentage of substitutions in the cloned cDNA fragment. A statistical analysis will give us the sequence variation in transcription infidelity sites. This technique can be ameliorated. Indeed, after reverse transcription and amplification with specific primers, cDNA can be cloned in a plasmid that carries a reporter gene e.g. *lacZ* gene. The cDNA and the reporter gene are cloned in phase. When a substitution appears in the STOP codon sequence of the mRNA, the reporter gene is expressed. When *E. coli* cells are transformed with this construction, colonies that carry the plasmid where the cDNA is mutated in the stop codon are selected with the expression of the reporter gene. After that, plasmid can be sequenced in order to verify if the substitution of the stop codon is present.

Another technique based on real-time PCR with specific primers is used to detect transcription infidelity sites. These primers are designed to match cDNA with sequence variation(s) based on statistical analysis. A second primer is designed to match the cDNA without the sequence variation (reference). The number of variation(s) in the primer sequence is very important and we have determined in a test experiment on known sequence variation that 2 mutations outside of the position that is studied are necessary to obtain a specific fluorescence signal. When the primer is complementary to the cDNA without a sequence substitution (reference), the fluorescence signal is detected at a specific number of amplification cycles. The same cDNA amplified with the primer that bears the sequence substitution leads to a fluorescence signal that appears with a greater number of amplification cycles. The difference between the numbers of amplification cycle with the 2 primers is a direct estimation of the presence of a cDNA with a sequence substitution. Moreover, we verified that the method is sensitive enough to detect 1% of mutated sequence in a mix of known sequences.

In a typical experiment, RNA extracted from cancerous and normal patient tissues are reverse transcribed and amplified with each of said specific primers. The difference between amplification cycle with both primers is then measured.

Finally, we can focus on novel proteins induced when the natural stop codon is affected. We showed that stop codon is significantly affected for 9 of our 17 genes of interest. That leads to distinct specific populations of proteins with a novel sequence in the carboxy-terminal end. We estimate that cancerous tissues for affected genes contain 4% proteins that are longer than normal ones.

In view of this hypothesis, we analysed 60 abundant plasma proteins (***Fig 13***) and searched for possible PSP. We found 22 genes for which a PSP was long enough and antigenic (***Fig 18***). We then searched for putative longer protein sequences induced (***Fig 18b-c***). We then searched for putative new peptides in plasma obtained from normal and cancerous individuals (***Fig 18***). We selected 3 out of these 22 interesting proteins: APOAI, APOAII and APOCII. Based on the above analysis, we expect to find longer proteins in cancerous-patient plasma (13, 16 and 17 AA longer). These transcription infidelity peptide sequences are predicted to be immunogenic, and antibodies directed against these novel sequences represent specific ligands to measure transcription infidelity (***Fig 14***). Since Kyte-Doolittle analysis indicates that these sequences are not hydrophobic, we expect these three novel proteins to be secreted into the circulation (see below).

Identification of post stop peptide (PSP) of ApoAII and ApoCII due to substitution in canonical stop codon

PSPs that result from base substitutions in the canonical stop codon can be identified and characterized in the following manner. Rabbit polyclonal antibodies are prepared that recognize an immunogenic portion(s) of the PSP in question. These anti-peptide antibodies can be checked by dot blot using the purified peptide to verify that they indeed recognize the PSP. Western blots can then be performed on plasma samples obtained from cancer patients using the antibodies directed against the PSP. Figure 17a (right panel) shows that the anti-PSP ApoAII antibody recognizes a band in Western blots performed on plasma samples obtained from prostate cancer patients,

that is not observed when using rabbit pre-immune serum as a negative control (Figure 17a, left panel). The PSP ApoAII band is also observed in cancer patients in the metastatic phase (Figure 17b, right panel). This band has a slightly greater molecular mass (11.4 kDa) as compared to that of the native monomer form of ApoAII (Figure 17a, middle panel, 17b, left panel). This molecular mass corresponds to that predicted based on the additional peptide sequence. Two-dimensional gels can also be performed in order to further characterize this band.

Affinity chromatography experiments can be carried out to isolate the PSP form of ApoAII using the anti-PSP antibody (***Fig 17e***). The anti-PSP antibody is immobilized on matrix beads and the following column is incubated in presence of plasma or delipidated HDL then sequentially washed to remove aspecifically bound proteins and finally eluted with detergent or chaotropic reagents. The eluted fraction is analysed by Western blotting using both the anti-PSP and the commercial anti-ApoAII antibodies. Two bands at 9 kDa and 11.4 kDa are recognized by the commercial anti-apoAII antibody whereas only the 11.4 kDa band is recognized by the anti-PSP antibody. Therefore, the 9 kDa band corresponds to the native ApoAII form and the 11.4 kDa one corresponds to the PSP form of ApoAII. The presence of native ApoAII in the eluted fraction suggests that the PSP ApoAII protein can form dimers with the native ApoAII protein under native non-reducing conditions. Indeed, ApoAII exists normally in plasma as a dimer of 2 monomers linked by a single disulfide bridge.

ApoAII is located primarily on the HDL fraction in plasma, which can be isolated by sequential ultracentrifugation. Western blotting shows that the PSP ApoAII is not detected in the d<1.07 g/ml fraction containing VLDL and LDL, but rather in the d>1.07 g/ml fraction containing HDL and plasma proteins (***Fig 17c***). After further ultracentrifugation steps to purify and wash HDL (d1.07-1.21 g/ml), Western blotting reveals that the PSP ApoAII remains associated with the HDL fraction rather than the d>1.21 g/ml fraction, also referred to as lipoprotein deficient serum (LPDS). A corresponding amount of plasma is shown on the Western blot to demonstrate that this is not due to a dilution of the d>1.21g/ml fraction during the purification steps. The separation of lipoproteins from plasma by gel filtration using Superose 6B (Amersham,

GE Healthcare) also demonstrates that the PSP ApoAII elutes in a similar manner to that of ApoAII associated with HDL.

PSP ApoAII association with HDL allowed the purification of the PSP form of ApoAII in a manner similar to that of ApoAII. HDL is first delipidated to remove all lipids. The resulting lipid-free protein pellet is then resuspended in a 10 mM Tris buffer containing urea or guanidine in presence or absence of reducing agent, and applied to a gel filtration column (example, Superdex 200, Amersham-GE Healthcare). Western blotting (Figure 17d) shows that the PSP form of ApoAII is still present in the delipidated HDL. Further purification was achieved by preparative electrophoresis (example: DE52). The PSP form of ApoAII was tracked by Western blotting. The purified PSP form of ApoAII was then cleaved enzymatically (trypsin) and the resulting peptides are analyzed on MS-MS for full AA sequencing. Results show that canonical STOP is replaced preferentially by Arginine with either (U to C) or (U to A) substitution followed by Serine, Valine, Glutamic acid, Threonine, Isoleucine, Valine, Phenylalanine, Glutamic acid, Proline, Glutamic acid, Leucine, Alanine, Serine, Arginine. This is the exact sequence of amino acid predicted to occur following bypass of ApoAII canonical STOP. Thus canonical UGA STOP of ApoAII is substituted to AGA or CGA leading to Arginine. A yet unexplored hypothesis is that UGA is converted to GGA leading to Glycine. But detection of this variant by mass spectrometry is currently beyond technological limits.

Another example is illustrated by the PSP of ApoCII. Figure 17f shows a similar experiment to that of Figure 17a, with the exception that Westerns were performed with commercially available anti-ApoCII or anti PSP ApoCII antibodies. A procedure similar to that for the PSP of ApoAII can be followed for ApoCII. However, ApoCII is less abundant in the plasma. A much larger quantity of plasma that contains the PSP ApoCII is therefore required in order to obtain sufficient amounts for analysis on MS-MS.

Conclusions

We describe here a novel mechanism leading to base substitutions occurring mostly in the 3' end of coding sequence and untranslated region of mRNA. These base substitutions lead to changes in protein AA sequences due to changes of AA identity, introduction of premature stop codons as well as the modification of naturally-occurring stop codons resulting in the introduction of novel coding regions. This phenomenon of transcription infidelity could also affect the ncRNA world and disturb the regulation mediated by these RNA. It occurs in most genes at a rate that exceeds any described phenomenon leading to DNA mutation. Transcription infidelity is greatly increased in cancer cells. Transcription infidelity provides a novel paradigm to understand cancer pathology, disease severity and disease progression. It has important implications not only for the design of novel transcriptomic and proteomic experiments, but also for the development of specific diagnostics and therapeutics.

CLAIMS

1. A method of detecting the presence or stage of a disease in a subject, the method comprising assessing in vitro or ex vivo the presence or rate of transcription infidelity in a sample from said subject, said presence or rate being an indication of the presence or stage of a disease in said subject.

2. A method of screening, identifying or optimising a drug, the method comprising a step of assessing whether said drug alters the rate of transcription infidelity of a gene.

3. The use of a compound that alters the rate of transcription infidelity of a mammalian gene for the manufacture of a pharmaceutical composition for use in a method of treatment of a human or animal, particularly for treating a proliferative cell disorder.

3. A method of modifying a mammalian gene to make it resistant or sensitive to transcription infidelity for use in a method of treatment of a human or animal, particularly for treating a proliferative cell disorder.

4. A method of assessing the efficacy of a drug or candidate drug, the method comprising a step of assessing whether said drug alters the rate of transcription infidelity of a mammalian gene, such an alteration being an indication of drug efficacy.

5. A method of identifying or producing biomarkers, the method comprising identifying, in a sample from a subject, the presence of transcription infidelity site(s) in a target protein, RNA or gene and, optionally, determining the sequence of said transcription infidelity site(s).

6. A method of identifying or producing ligands specific for a trait or pathological condition, the method comprising identifying, in a sample from a subject having said trait or pathological condition, the presence of at least one transcription infidelity site in one or several target proteins or genes, optionally determining the sequence of said at least one transcription infidelity site, and producing a ligand that specifically binds to said at least one transcription infidelity site.

7. The method of any one of claims 1 to 6, wherein transcription infidelity is assessed using at least one ligand specific for a transcription infidelity site of a protein or nucleic acid.

8. The method of any one of claims 1 to 7, wherein the disease or pathological condition is selected from cell proliferative disorders, such as cancers, immune diseases, inflammatory diseases or aging.

9. A synthetic peptide comprising a domain of a TI-protein, particularly of a mammalian protein, more preferably of a human protein.

10. The peptide of claim 9, wherein the transcription infidelity site comprises at least one modification encoded by a nucleic acid alteration obeying to the following rules:

. Substitution of A occurs preferentially when A is preceded or followed by C,
. Substitution of T occurs preferentially when T is preceded or followed by G,
. Substitution of C occurs preferentially when C is preceded by G or A, and
. Substitution of G occurs preferentially when G is preceded or followed by A.

11. The peptide of any one of claims 9 to 10, wherein said peptide comprises the sequence of a C-terminal fragment of said protein comprising at least one amino acid variation caused by transcription infidelity, and/or comprises less than 100, 80, 75, 70, 65, 60, 50, 45, 40, 35, 30, 25 or even 20 amino acids.

12. The peptide of any one of claims 9 to 11, wherein the protein is a cell surface protein (e.g., a receptor) or a secreted protein (e.g., a plasma protein).

13. A peptide comprising a sequence selected from SEQ ID NOs: 1 to 32.

14. The use of a peptide of any one of claims 9 to 13 as an immunogen.

15. A vaccine composition comprising a peptide of any one of claims 9 to 13 and optionally a suitable carrier, excipient and/or adjuvant.

16. A synthetic nucleic acid molecule comprising the sequence of a transcription infidelity site of an RNA molecule.

17. A method of producing an antibody, the method comprising immunizing a non-human mammal with a peptide of any one of claims 9 to 13 and recovering antibodies that bind said peptide, or corresponding antibody-producing cells.

18. A method of producing an antibody, the method comprising (i) identifying a domain of a TI-protein, (ii) producing an antibody that specifically binds said domain and (iii) optionally, producing derivatives of the antibody which have substantially the same antigen specificity.

19. The method of claim 18, wherein identifying a transcription infidelity site comprises identifying at least one modification encoded by a nucleic acid alteration obeying to the following rules:

. Substitution of A occurs preferentially when A is preceded or followed by C,

. Substitution of T occurs preferentially when T is preceded or followed by G,

. Substitution of C occurs preferentially when C is preceded or followed by G or A, and

. Substitution of G occurs preferentially when G is preceded or followed by A.

20. A ligand that specifically binds a transcription infidelity site of a protein or RNA.

21. The ligand of claim 20, wherein said ligand is an antibody (or any fragment or derivative thereof), a nucleic acid primer or a nucleic acid probe.

22. An antibody that specifically binds a domain of a TI-protein, or a derivative of such an antibody having substantially the same antigen specificity.

23. An antibody or derivative thereof, according to claim 22, wherein said antibody is conjugated to a molecule, such as a drug, a label, a toxic molecule or a radioisotope.

24. The use of a ligand of claim 20 or 21 for detecting or quantifying in vitro transcription infidelity of a gene.

25. The use of a conjugated antibody or derivative thereof of claim 23, as a medicament or diagnostic reagent.

26. The use of a nucleic acid comprising a transcription infidelity site of a gene for the manufacture of a pharmaceutical composition for use in a method of treatment of a human or animal.

27. A device or product comprising, immobilized on a support, a ligand of claim 20 or 21.

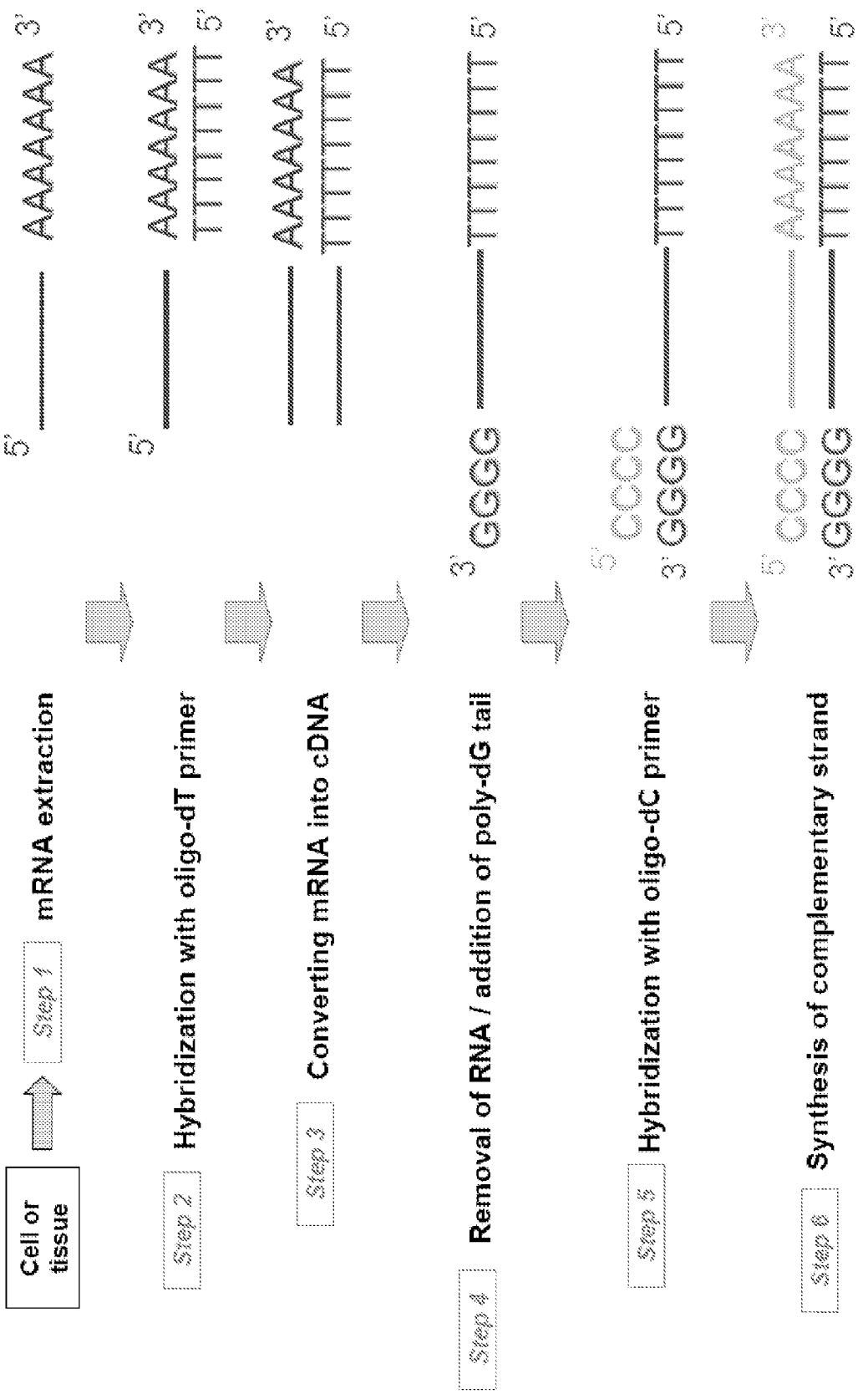

Cell or tissue
Step 1
mRNA extraction
Step 2
Hybridization with oligo-dT primer
Step 3
Converting mRNA into cDNA
Step 4
Removal of RNA / addition of poly-dG tail
Step 5
Hybridization with oligo-dC primer
Step 6
Synthesis of complementary strand
5' AAAAAAA 3'
5' AAAAAAA 3'
TTTTTTT 5'
3' AAAAAAA 5'
TTTTTTT 5'
3' GGGGG TTTTTTT 5'
5' CCCC
3' GGGGG TTTTTTT 5'
5' CCCC AAAAAAA 3'
3' GGGGG TTTTTTT 5'

Figure 1a

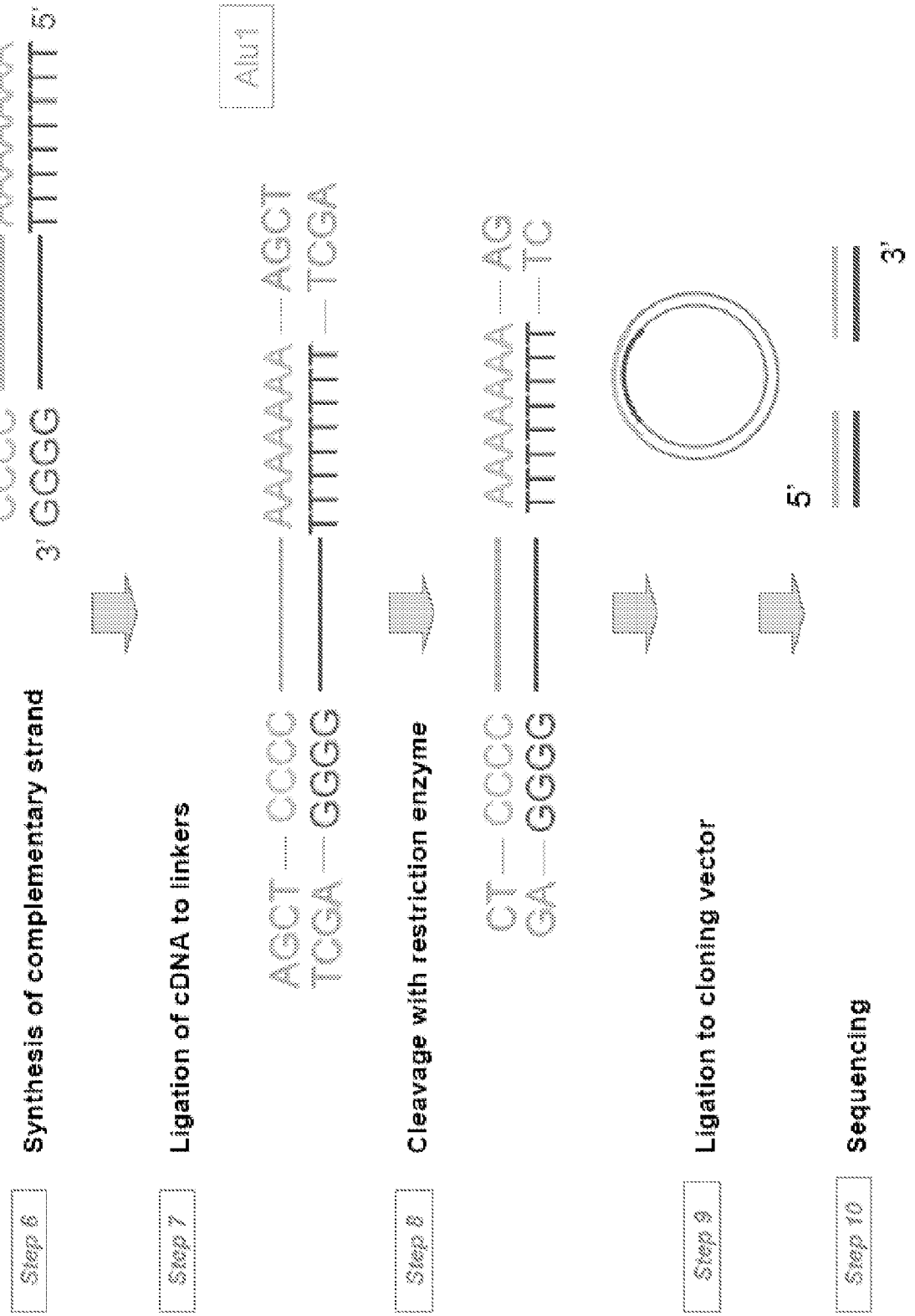

Step 6
Synthesis of complementary strand
5' CCCC ——— AAAAAA 3'
3' GGGG ——— TTTTTTTT 5'
Step 7
Ligation of cDNA to linkers
Alu1
AGCT — CCCC ——— AAAAAA — AGCT
TCGA — GGGG ——— TTTTTTTT — TCGA
Step 8
Cleavage with restriction enzyme
CT — CCCC ——— AAAAAA — AG
GA — GGGG ——— TTTTTTTT — TC
Step 9
Ligation to cloning vector
Step 10
Sequencing
5'
3'

Figure 1b

>gi|16507965|ref|NM_001428.2| Homo sapiens enolase 1, (alpha) (ENO1), mRNA

TAGCTAGGCAGGAAGTCGGCGCGGGCGGCGCGGACAGTATCTGTGGGTACCCGGAGCACGGAGATCTCGCCGGCTTTACGTTCACCTCGG
TGTCTGCAGCACCCTCCGCTTCCTCTCCTAGGCGACGAGACCCAGTGGCTAGAAGTTCACCATGTCTATTCTCAAGATCCATGCCAGGGA
GATCTTTGACTCTCGCGGGAATCCCACTGTTGAGGTTGATCTCTTCACCTCAAAAGGTCTCTTCAGAGCTGCTGTGCCCAGTGGTGCTTC
AACTGGTATCTATGAGGCCCTAGAGCTCCGGGACAATGATAAGACTCGCTATATGGGGAAGGGTGTCTCAAAGGCTGTTGAGCACATCAA
TAAAACTATTGCGCCTGCCCTGGTTAGCAAGAAACTGAACGTCACAGAACAAGAGAAGATTGACAAACTGATGATCGAGATGGATGGAAC
AGAAAATAAATCTAAGTTTGGTGCGAACGCCATTCTGGGGGTGTCCCTTGCCGTCTGCAAAGCTGGTGCCGTTGAGAAGGGGGTCCCCCT
GTACCGCCACATCGCTGACTTGGCTGGCAACTCTGAAGTCATCCTGCCAGTCCCGGCGTTCAATGTCATCAATGGCGGTTCTCATGCTGG
CAACAAGCTGGCCATGCAGGAGTTCATGATCCTCCCAGTCGGTGCAGCAAACTTCAGGGAAGCCATGCGCATTGGAGCAGAGGTTTACCA
CAACCTGAAGAATGTCATCAAGGAGAAATATGGGAAAGATGCCACCAATGTGGGGGATGAAGGCGGGTTTGCTCCCAACATCCTGGAGAA
TAAAGAAGGCCTGGAGCTGCTGAAGACTGCTATTGGGAAAGCTGGCTACACTGATAAGGTGGTCATCGGCATGGACGTAGCGGCCTCCGA
GTTCTTCAGGTCTGGGAAGTATGACCTGGACTTCAAGTCTCCCGATGACCCCAGCAGGTACATCTCGCCTGACCAGCTGGCTGACCTGTA
CAAGTCCTTCATCAAGGACTACCCAGTGGTGTCTATCGAAGATCCCTTTGACCAGGATGACTGGGGAGCTTGGCAGAAGTTCACAGCCAG
TGCAGGAATCCAGGTAGTGGGGGATGATCTCACAGTGACCAACCCAAAGAGGATCGCCAAGGCCGTGAACGAGAAGTCCTGCAACTGCCT
CCTGCTCAAAGTCAACCAGATTGGCTCCGTGACCGAGTCTCTTCAGGCGTGCAAGCTGGCCCAGGCCAATGGTTGGGGCGTCATGGTGTC
TCATCGTTCGGGGGAGACTGAAGATACCTTCATCGCTGACCTGGTTGTGGGGCTGTGCACTGGGCAGATCAAGACTGGTGCCCCTTGCCG
ATCTGAGCGCTTGGCCAAGTACAACCAGCTCCTCAGAATTGAAGAGGAGCTGGGCAGCAAGGCTAAGTTTGCCGGCAGGAACTTCAGAAA
CCCCTTGGCCAAGTAAGCTGTGGGCAGGCAAGCCCTTCGGTCACCTGTTGGCTACACAGACCCCTCCCCTCGTGTCAGCTCAGGCAGCTC
GAGGCCCCCGACCAACACTTGCAGGGGTCCCTGCTAGTTAGCGCCCCACCGCCGTGGAGTTCGTACCGCTTCCTTAGAACTTCTACAGAA
GCCAAGCTCCCTGGAGCCCTGTTGGCAGCTCTAGCTTTGCAGTCGTGTAATTGGCCCAAGTCATTGTTTTTCTCGCCTCACTTTCCACCA
AGTGTCTAGAGTCATGTGAGCCTCGTGTCATCTCCGGGGTGGCCACAGGCTAGATCCCCGGTGGTTTTGTGCTCAAAATAAAAAGCCTCA
GTGACCCATGAG

Figure 2a

>gi|56682958|ref|NM_002032.2| Homo sapiens ferritin, heavy polypeptide 1 (FTH1), mRNA

ATAAGAGACCACAAGCGACCCGCAGGGCCAGACGTTCTTCGCCGAGAGTCGTCGGGGTTTCCTGCTTCAACAGTGCTTGGACGGAAC
CCGGCGCTCGTTCCCCACCCCGGCCGGCCGCCCATAGCCAGCCCTCCGTCACCTCTTCACCGCACCCTCGGACTGCCCCAAGGCCCC
CGCCGCCGCTCCAGCGCCGCGCAGCCACCGCCGCCGCCGCCGCCTCTCCTTAGTCGCCGCCATGACGACCGCGTCCACCTCGCAGGT
GCGCCAGAACTACCACCAGGACTCAGAGGCCGCCATCAACCGCCAGATCAACCTGGAGCTCTACGCCTCCTACGTTTACCTGTCCAT
GTCTTACTACTTTGACCGCGATGATGTGGCTTTGAAGAACTTTGCCAAATACTTTCTTCACCAATCTCATGAGGAGAGGGAACATGC
TGAGAAACTGATGAAGCTGCAGAACCAACGAGGTGGCCGAATCTTCCTTCAGGATATCAAGAAACCAGACTGTGATGACTGGGAGAG
CGGGCTGAATGCAATGGAGTGTGCATTACATTTGGAAAAAAATGTGAATCAGTCACTACTGGAACTGCACAAACTGGCCACTGACAA
AAATGACCCCCATTTGTGTGACTTCATTGAGACACATTACCTGAATGAGCAGGTGAAAGCCATCAAAGAATTGGGTGACCACGTGAC
CAACTTGCGCAAGATGGGAGCGCCCGAATCTGGCTTGGCGGAATATCTCTTTGACAAGCACACCCTGGGAGACAGTGATAATGAAAG
CTAAGCCTCGGGCTAATTTCCCCATAGCCGTGGGGTGACTTCCCTGGTCACCAAGGCAGTGCATGCATGTTGGGGTTTCCTTTACCT
TTTCTATAAGTTGTACCAAAACATCCACTTAAGTTCTTTGATTTGTACCATTCCTTCAAATAAAGAAATTTGGTACCCAGGTGTTGT
CTTTGAGGTCTTGGGATGAATCAGAAATCTATCCAGGCTATCTTCCAGATTCCTTAAGTGCCGTTGTTCAGTTCTAATCACACTAAT
CAAAAAGAAACGAGTATTTGTATTTATTAAACTCATTAGTTTGGGCAGTATACTAAGGTGTGGCTGTCTTGGATTCAGATAGAACTA
AGGGTTCCCGACTCTGAATCCAGAGTCTGAGTTAAATGTTTCCAATGGTTCAGTCTAGCTTTCACAGTTTTTATGAATAAAAGGCAT
TAAAGGCTGA

Figure 2b

>gi|56682960|ref|NM_000146.3| Homo sapiens ferritin, light polypeptide (FTL), mRNA

GCAGTTCGGCGGTCCGCGGTCTGTTCTTGCTTCAACAGTGTTTGGACGGAACAGATCCGGGGACTCTCTTCCAGCCTCCGACCGC
CCTCCGATTCCTCTCCTCCGCTTGCAACGTCCGGGACCATCTTCTTCGGCCATCTCCTGCTTCTGGGACCTGCCAGCACCGTTTTTGTGGT
TAGCTCCTTCTTGCCAACCAACCATGAGCTCCCAGATTCGTCAGAATTATTCCACCGACGTGGAGGCAGCCCGTCAACAGCCTGGTCAA
TTTGTACCTGCAGGCCTCCTACACCTACTCTCTGGGCTTCTATTTCGACCGCGATGATGTGGCTCTGGAAGGCGTGAGCCACTTC
TTCCGCGAATTGGCCGAGGAGAAGGCTACGAGGGCTACGAGAAGATGCAAAACCAGCGTGGCGGCCGCGCTCTTCTTCCAGG
ACATCAAGAAGCCAGCTGAAGATGAGTGGGGTAAAACCCCAGACGCCATGAAAGCTGCCATGGCCTTGGAGAAAAAGCTGAACCAGGCC
CTTTTGGATCTTCATGCCTTGGCTTCTGGCACGGACCCCATCTCTGTGACTTCCTGGAGACTCACTTCCTAGATGAGGAAGTG
AAGCTTATCAAGAAGATGGGTGACCACCTGACCAACCTCCACAGGCTGGGTGGCCCGGAGGCTGGGCTGGGCGAGTATCTCTTCGAAA
GCCTCACTCTCAAGCACGACTAAGAGCCTTCTGAGCCCAGCGACTTCTTGCAAAGTAATAGGGCTTCTGCCTAAGCC
TCTCCCTCCAGCCAATAGGCAGCTTTCTTAACTATCCTAACAAGCCTTGGACCAAATGGAAATAAAGCTTTTTGATGC

Figure 2c

>gi|83641890|ref|NM_002046.3| Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), mRNA

AAATTGAGCCCGCAGCCTCCCGCTTCGCTCTCTGCTCCTCCTGTTCGACAGTCAGCCGCATCTTCTTTTGCGTCGCCAGCCGAGCCACA
TCGCTCAGACACCATGGGGAAGGTGAAGGTCGGAGTCAACGGATTTGGTCGTATTGGGCGCCTGGTCACCAGGGCTGCTTTTAACTCTG
GTAAAGTGGATATTGTTGCCATCAATGACCCCTTCATTGACCTCAACTACATGGTTTACATGTTCCAATATGATTCCACCCATGGCAAA
TTCCATGGCACCGTCAAGGCTGAGAACGGGAAGCTTGTCATCAATGGAAATCCCATCACCATCTTCCAGGAGCGAGATCCCTCCAAAAT
CAAGTGGGGCGATGCTGGCGCTGAGTACGTCGTGGAGTCCACTGGCGTCTTCACCACCATGGAGAAGGCTGGGGCTCATTTGCAGGGGG
GAGCCAAAAGGGTCATCATCTCTGCCCCCTCTGCTGATGCCCCCATGTTCGTCATGGGTGTGAACCATGAGAAGTATGACAACAGCCTC
AAGATCATCAGCAATGCCTCCTGCACCACCAACTGCTTAGCCCCAAGGTCATCCATGACAACTTTGGTATCGTGGAAGGACT
CATGACCACAGTCCATGCCATCACTGCCACCCAGAAGACTGTGGATGGCCCCTCCGGGGAAACTGTGGCGTGATGGCCGCGGGGCTCTCC
AGAACATCATCCCTGCCTCTACTGGCGCTGCCAAGGCTGTGGGCAAGGTCATCCCTGAGCTGAACGGGAAGCTCACTGGCATGGCCTTC
CGTGTCCCCACTGCCAACGTGTCAGTGGTGGACCTGACCTGCCGTCTAGAAAAACCTGCCAAATATGATGACATCAAGAAGGTGGTGAA
GCAGGCGTCGGAGGGCCCCTCAAGGGCATCCTGGGCTACACTGAGCACCAGGTGGTCTCCTCTGACTTCAACAGCGACACCCACTCCT
CCACCTTTGACGCTGGGGCTGGCATTGCCCTCAACGACCACTTTGTCAAGCTCATTTCCTGGTATGACAACGAATTTGGCTACAGCAAC
AGGGTGGTGGACCTCATGGCCCACATGGCCTCCAAGGAGTAAGACCCCTGGACCACCAGCCCCAGCAAGAGCACAAGAGGAAGAGAGAG
ACCCTCACTGCTGGGGAGTCCCTGCCACACTCAGTCCCCCACCACACTGAATCTCCCCTCCTCACAGTTGCCATGTAGACCCCTTGAAG
AGGGGAGGGCCTAGGGAGCCGCACCCTTGTCATGTACCATCAATAAAGTACCCTGTGCTCAACC

Figure 2d

>gi|24234684|ref|NM_006597.3| Homo sapiens heat shock 70kDa protein 8 (HSPA8), transcript variant 1, mRNA

CTCATTGAACTCGCCTGCAGCTCTTGGTTTTTTGTGGTTCCTCGTTATTGGAGCCAGGCCTACACGCCAGCATGTCCAAGGGAC
CTGCAGTTGGTATTGATCTTGGCACCACCTACTCTTGTGGGTGTTTTCCAGCACGGAAAAGTCGAGATAATTGCCAATGATCAGGGAAA
CCGAACCACTCCAAGCTATGTCGCCTTTACGGACACTGGTTGATCGGTGCGCAAAGAATCAAGTTGCAATGAACCCCACCAAC
ACAGTTTTTGATGCCAAACGTCTGATTGGACAGATTTGATGATGCTGTTGTCCAGTCTGATATGAAACATTGGCCCTTTATGGTGGTGA
ATGATGCTGGCAGGCCCAAGGTCCAAGTAGAATACAAGGGAGACCAAAAGCTTCTATCCAGAGGAGGTGTCTTCTATGGTCTGACAAA
GATGAAGGAAATTGCAGAAGCCTACCTTGGGAAGACTGTTACCAATGCTGTGGTCACAGTGCCAGCTTACTTTAATGACTCTCAGCGTCAG
GCTACCAAAGATGCTGGAACTATTGCTGGTCTCAATGTACTTAGAATTATAATGAGCCAACTGCTGCTGCTACGGCTTAGACA
AAAAGGTTGGAGCAGAAGAAACGTGCTCATCTTTGACCTGGGAGGTGGCACTTTTGATGTCAGTCTCACTATTGAGGATGGAATCTT
TGAGGTCAAGTCTACAGCTGGAGACACCCACTTGGGTGGAGAAGATTTTGACAACCGAATGGTCAACCATTTTATTGCTGAGTTTAAGCGC
AAGCATAAGAAGGACATCAGTGAGAACAAGAGAGCTGTAAGACGCCTCCGTACTGCTTGTGAACGTGCTAACCCTCTTCTTCCAGCA
CCCAGGCCAGTATTGAGATCGATTCTCTCTATGAAGGAATCGACTTCTATACCTCCATTACCCGTGCCGGATTTGAAGAACTGAATGCTGA
CCTGTTCCGTGGCACCCTGGACCCAGTAGAGGAAGCCCTTCGAGATGCCAAACCTAGACAAGTCACAGATTCATGATATTGTCCTGGTTGGT
GGTTCTACTCCGTATCCCCAAGATTCAGAAGCTTCTCCAAGACTTCTTCAATGGAAAAGAACTGAATAAGAGCATCAACCCTGATGAAGCTG
TTGCTTATGGTGCAGCTGTCCAGGCCATCTTGTCTGGAGACAAGTCTGAGAATGTTCAAGATTTGCTGCTCTTGGATGTCACTCCTCT
TTCCCTTGGTATTGAAACTGCTGGTGGAGTCATGACTGTCCTCATCAAGCGTAATACCACCATTCCTACCAAGCAGACAGACTTCACT
ACCTATTCTGACAACCAGCCTGGTGTGCTTATTCAGGTTTATGAGGGCGAGCGTGCCATGACAAAGGATAACAACCTGCTTGGCAAGTTTG
AACTCACAGGCATACCTCCCTGCACCCCGAGGTGTTCCTCAGATTGAAGTCACTTTTGACATTGATGCCAATGGTATACTCAATGTCTCTGC
TGTGGACAAGAGTACGGGAAAAGAGAACAAGATTACTATCACTAATGACAAGGGCCGTTTGAGCAAGGAAGACATTGAACGTATGGTCCAG
GAAGCTGAGAAGTACAAAGCTGAAGATGAGAAGCAGAGGGACAAGGTGTCATCCAAGAATTCACTTGAGTCCTATGCCTTCAACATGAAAG
CAACTGTTGAAGATGAGAAACTTCAAGGCAAGATTAACGATGAGGACAAACAGAAGATTCTGGACAAGTGTAATGAAATTATCAACTGGCT
TGATAAGAATCAGACTGCTGAGAAGGAAGAATTTGAACATCAACAGAAAGAGCTGGAGAAAGTTTGCAACCCATCATCACAGCAAGCTGTAC
CAGAGTGCAGGAGGCATGCCAGGAGGAATGCCTGAGGGATTTCCTGGTGAGCCTCCTCCTCTGGTGGTGCTTCCTCAGGGCCACCA
TTGAAGAGGTTGATTAAGCCAACCAAGTGTAGATGTGCATTGTTCCACACATTTAAAACATTTGAAGGACCTAAATTCGTAGCAAATTCT
GTGGCAGTTTTAAAAAGTTAAGCTGCTATAGTAAGTTACTGGGCATTCTCAATACTTGAATATGGAACATATGCACAGGGGAAGGAAATAA
CATTGCACTTTATAAACACTGTATTGTAAGTGGAAAATGCAATGTCTTAAATAAAACTATTTAAAATTGGCACCAT

Figure 2e

>gi|18390348|ref|NM_000972.2| Homo sapiens ribosomal protein L7a (RPL7A), mRNA

TTTCCTTTCTCTCCTCCGCCGCCCAGATGCCGAAGGAAGAGAAGGCCAAGGAAGAAGAGGTGGCTCCGGCCCCAGCTGTCGTGAA
GAAGCAGGAGGCTAAGAAAGTGGTGAATCCCCTGTTTGAGAAAAGGCCTAAGAATTTTGGCATTGGACAGGACATCCAGCCCAAAGAG
ACCTCACCCGCTTTGTGAAATGGCCCCGCTATATCAGGTTCAGCGGCAGAGAGCCATCCTCTATAAGCGGCTGAAAGTGCCTCCTGCG
ATTAACCAGTTGACCCAGGGCCCTGGACCGCCAAACAGCTACTCAGCTGCTTAAGCTGGCCCACAAGTACAGAGCCAGAGACAAAGGCAAGA
GAAGAAGCAGAGACTGTTGGCCCGAGAAGAAGGCTGCTGGCAAAGGGGACGTCCCCAACGAAGAGACCACCTGTCCTTCGAGCCAG
GAGTTAACACCGTCACCACCTTGGTGGAGAACAAGAAAGCTCAGCTGGTGGTGATTGCACACGACGTGGATCCCATCGAGCTGGTTGTC
TTCTTGCCTGCCCTGTGTCGTAAAATGGGGGTCCCTTACTGCATTATCAAGGGAAGGCAAGACTGGGACGTCTAGTCCACAGGAAGAC
CTGCCACCACTGTCGCCTTCACACAGGTGAACTCGGAAGACAAAGGCGCTTTGGCTAAGCTGGTGGAAGCTATCAGGACCAATTACAATG
ACAGATACGATGAGATCCGCCGTCACTGGGGTGGCAATGTCCTGGGTCCTAAGTCTGTGGCCTCGTATCGCCAAGCTCGAAAGGCAAAG
GCTAAAGAACTTGCCACTAAACTGGGTTAAATGTACACTGTTGAGTTTTCTGTACATAAAAATAATTGAAATAATACAAATTTTGCCTTC

Figure 2f

>gi|39812410|ref|NM_001007.3| Homo sapiens ribosomal protein S4, X-linked (RPS4X), mRNA

GGGCGGAGCAGCTGAAAAATCCGGCGCGCAGTCTCCAGCCCCAATTTCTACGCGCACCCGGAAGACGGAGGTCCTCTTTCCTTGCGT
AACGCAGCCATGGCTCGTGGTCCCAAGAAGCATCTGAAGCGGGTGGCAGCTCCAAAGCATTGGATGCCTGGATAAATTGACCGGTGTG
TTTGCTCCTCGTCCATCCACCGGTCCCCACAAGTTGAGAGAGTGTCTCCCCCTCATCATTTTCCTGAGGAACAGACTTAAGTATGCC
CTGACAGGAGATGAAGTAAAGAAGATTTTGCATGCAGCGGTTCATTAAAATCGATGGCAAGGTCCGAACTGATATAACCTACCCTGCT
GGATTCATGGATGTCATCAGCATTGACAAGACGGGAGAGAAATTTCGTCTGATCTATGACACCAAGGGTCGCTTTGCTGTACATCGT
ATTACACCTGAGGAGGGCCAAGTACAAGTTGTGCAAAGTGAGAAAGATCTTTGTGGGCACAAAAGGAATCCCTCATCTGGTGACTCAT
GATGCCGGCACCATCCGCTACCCCGATCCCCTCATCAAGGTGAATGATACCATTCAGATTGATTTGGAGACTGGCAAGATTACTGAT
TTCATCAAGTTCGACACTGGTAACCTGTGTATGGTGACTGGAGGTGCTAACCTAGGAAGAATTGGTGTGATCACCAACAGAGAGAGG
CACCCTGGATCTTTTGACGTGGTTCACGTGAAAGATGCCAACAGCTTTGCCACTCGACTTTCCAACATTTTTGTTATTGGC
AAGGGCAACAAACCATGGATTTCTCTTCCCCGAGGAAAGGGTATCCGCCTCACCATTGCTGAAGAGACAAAAGACTGGCGGC
AAACAGCAGCAGTGGGTGAAATGGGTCCCTGGGTGACATGTCAGATCTTTGTACGTAATTAAAAATATTGTGGCAGGATTAATAGC

Figure 2g

>gi|17158043|ref|NM_001010.2| Homo sapiens ribosomal protein S6 (RPS6), mRNA

CCTCTTTTCCGTGGCGCCTCGGAGGCGTTCAGCTGCTTCAAGATGAAGCTGAACATCTCCTTCCCAGCCACCGGCTGCCAGAAACTCA
TTGAAGTGGACGATGAACGCAAACTTCGTACTTTCTATGAGAAGCGTATGGCCACAGAAGTTGCTGCTGACGGCTCTGGGTGAAGAATG
GAAGGGTTATGTGGTCCGAATCAGTGGTGGGAACGACAAGAACAAGGTTTCCCCATGAAGCAGGGTGTCTTGACCCATGGCCGTGTCCGC
CTGCTACTGAGTAAGGGGCATTCCTGTTACAGACCAAGGAGACTGGAGAAAGAAGAGAAGAAATCAGTTCGTGGTTGCATTGTGGATG
CAAATCTGAGCGTTCTCAACTTGGTTATTGTAAAAAAGCGAGAAGGCATATTCCTGGACTGACTGATACTACAGTGCCTCGCGCGCCT
GGGCCCAAAAGAGCTAGCAGAATCCGCAAACTTTCAATCTCTCTAAAGAAGATGATCCGCCAGTATGTTGTAAGAAAGCCCTTA
AATAAAGAAGGTAAGAAACCTAGGACCCAAAGCACCCAAGATTCAGCGTCCTTGTTACTCCACGTGTCCTGCAGCACAAACGGCGGCGTA
TTGCTCTGAAGAAGCAGCGTACCAAGAAAAATAAAGAAGGCTGCAGAATATGCTAAACTTTTTGGCCAAGAGAATGAAGGAGGCTAA
GGAGAAGCGCCAGGAACAAATTGCGAAGAGACGCAGACTTTCCTCTCTCGAGCTTCTACTTCTAAGTCTGAATCCAGTCAGAAATAA
GATTTTTTGAGTAACAAATAAATAAGATCAGACTCTG

Figure 2h

>gi|34328943|ref|NM_021109.2| Homo sapiens thymosin, beta 4, X-linked (TMSB4X), mRNA

ACAACTCCGGTGGTGGCCACTGCGCAGACCAGACTTCGGTACTCGTGCGCTCGCTTTCCTCCGCAACCATGTCTGACAAACCC
GATATGGCTGAGATCGAGAAATTCGATAAGTCGAAACTGAAGAAGACAGAGACGCAAGAGAAAAATCCACTGCCTTCCAAAGAAACGATTGA
ACAGGAGAAGCAAGCAGGCGAATCGTAATGAGGCGTGCGCCGGCCAATATGCACTGTACATTCCACAAGCATTGCCTTCTTATTTTACTTCTT
TTAGCTGTTTTAACTTTGTAAGATGCAAAGAGGTTGGATCAAGTTTAAATGACTGTGCTGCCCCTTTCACATCAAAGAACTACTGACAACGAA
GGCCGCGCCTGCCTTTCCCATCTGTCTATCTGGCAGGAAGGAAAGAACTTGCATGTTGGTGAAGGAAGGAAGTGGGTGGAAGA
AGTGGGGTGGGACGACAGTGAAATCTAGAGTAAAAACCAAGCTGGCCCCAAGGGTGTCCTGCAGGCTGTAATGCAGTTTAATCAGAGTGCCATTT
TTTTTTTTGTTCAAATGATTTTAATTATTGGAATGCACAATTTTTTTAATATGCAAATAAAAAAGTTTAAAAAACTT

Figure 2i

>gi|4507668|ref|NM_003295.1| Homo sapiens tumor protein, translationally-controlled 1 (TPT1), mRNA

CCCCCCGAGCGCCGCTCCGGCTGCACCGCGCTCGTTCAGGCTCGTGCTAAGCTAGCGCCGTCGTCGTCCTTCAGT
CGCCATCATGATTATCTACCGGGACCTCATCAGCCACGATGAGATGTTCTCCGACATCTACAAGATCCGCGAGATCGCCGGCGACGGGTT
GTGCCTGGAGGTGGAGGGAAGATGGTCAGTAGGACAGAAGGTAACATTGATGACTCGCTCATTGGTGGAAATGCCTCCGCCTGAAGG
CCCCGAGGCCGAAGCCTACCGAAAGCCACAGTAATCACTGGTGTCGATATTGTCATGAACCATCACCTGCAGCAAACAAGTTTCACAAA
AGAAGCCTACAAGAAGTACATCAAAGAATTACATGAAATCAATCAAAAGGGAAACTTGAAGAACAGAGACCAGAAAGAGTAAAAACCTTT
TATGACAGGGGCCTGCAGAACAAATCAAGCACATCCTTGCTAATTTCAAAAACTACCAGTTCTTTATTGGTGAAAACATGAATCCAGA
TGGCATGGTTGCCTCTATTGGACTACCGTGAGGATGGTGTGACCCCATATATGATTTTCTTTAAGGATGGTTTAGAAATGGAAAAATG
TTAACAAATGTGGCAATTATTTTGGATCTATCACCTGTCATAACTGGCTTCTGCTTGTCATCCACACAACACCAGGACTTAAGA
CAAATGGGACTGATGTCATCTTGAGCTCTTCATTTATTTTGACTGTGATTTATTTGGAGTGGAGGCATTGTTTTTAAGAAAAACATG
TCATGTAGGTTGTCTAAAAATAAAATGCATTTTAAACTCATTTGAGAG

Figure 2j

>gi|62414288|ref|NM_003380.2| Homo sapiens vimentin (VIM), mRNA

GTCCCCGCGCCAGAGACGCAGCCGCGCTCCCACCACCCACACCCACCGCGCCCTCGTTCGCCTCTTCTCCGGGAGCCAGTCCGCGCC
ACCGCCGCCGCCCAGGCCATCGCCACCCTCCGCAGCCATGTCCACCAGGTCCGTGTCCTCGTCCTCCTACCGCAGGATGTTCGGCGG
CCCGGGCACCGCGAGCCGGCCGAGCTCCAGCCGGAGCTACGTGACTACGTCCACCCGCACCTACAGCCTGGGCAGCGCGCTGCGCCC
CAGCACCAGCCGCAGCCTCTACGCCTCGTCCCCGGGCGGCGTGTATGCCACGCGCTCCTCTGCCGTGCGCCTGCGGAGCAGCGTGCC
CGGGGTGCGGCTCCTGCAGGACTCGGTGGACTTCTCGCTGGCCGACGCCATCAACACCGAGTTCAAGAACACCCGCACCAACGAGAA
GGTGGAGCTGCAGGAGCTGAATGACCGCTTCGCCAACTACATCGACAAGGTGCGCTTCCTGGAGCAGCAGAATAAGATCCTGCTGGC
CGAGCTCGAGCAGCTCAAGGGCCAAGGCAAGTCGCGCCTGGGGGACCTCTACGAGGAGGAGATGCGGGAGCTGCGCCGGCAGGTGGA
CCAGCTAACCAACGACAAAGCCCGCGTCGAGGTGGAGCGCGACAACCTGGCCGAGGACATCATGCGCCTCCGGGAGAAATTGCAGGA
GGAGATGCTTCAGAGAGAGGAAGCCGAAAACACCCTGCAATCTTTCAGACAGGATGTTGACAATGCGTCTCTGGCACGTCTTGACCT
TGAACGCAAAGTGGAATCTTTGCAAGAAGAGATTGCCTTTTTGAAGAAACTCCACGAAGAGGAAATCCAGGAGCTGCAGGCTCAGAT
TCACGAACACCATGTCCAAATCCATGTGCATCTTTCCAAGCCTCACCTCACCGCTCCCCTCCCTCACCTACCTCACCAATATCAAAC
TGTGGCTGCCAAGAACCTGCAGGAGGCAGAAGAATGGTACAAATCCAAGTTTGCTGACCTCTCTGAGGCTGCCAACCGGAACAATGA
CGCCCTGCGCCAGGCAAAGCAGGAGTCCACTGAGTACCGGAGACAGGTGCAGTCCCTCACCTGTGAAGTGGATGCCCTTAAAGGAAC
CAATGAGTCCCTGGAACGCCAGATGCGTGAAATGGAAGAGAACTTTGCCGTTGAAGCTGCTAACTACCAAGACACTATTGGCCGCCT
GCAGGATGAGATTCAGAATATGAAGGAGGAAATGGCTCGTCACCTTCGTGAATACCAAGACCTGCTCAATGTTAAGATGGCCCTTGA
CATTGAGATTGCCACCTACAGGAAGCTGCTGGAAGGCGAGGAGAGCAGGATTTCTCTGCCTCTTCCAAACTTTTCCTCCCTGAACCT
GAGGGAAACTAATCTGGATTCACTCCCTCTGGTTGATACCCACTCAAAAAGGACACTTCTGATTAAGACGGTTGAAACTAGAGATGG
ACAGGTTATCAACGAAACTTCTCAGCATCACGATGACCTTGAATAAAAATTGCACACACTCAGTGCAGCAATATATTACCAGCAAGA
ATAAAAAAAGAAATCCATATCTTAAAGAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGATAGATTTGGAATAGGA
ATAAGCTCTAGTTCTTAACAACCGACACTCCTACAAGATTTAGAAAAAAGTTTACAACATAATCTAGTTTACAGAAAAATCTTGTGC
TAGAATACTTTTTAAAAGGTATTTGAATACCATTAAAACTGCTTTTTTTTTTCCAGCAAGTATCCAACCAACTTGGTTCTGCTTCA
ATAAATCTTTGGAAAAACTC

Figure 2k

>gi|8392890|ref|NM_000477.3| Homo sapiens albumin (ALB), mRNA

AGCTTTTCTCTTCTGTCAACCCCACACGCCTTTGGCACAATGAAGTGGGTAACCTTTATTTCCCTTCTTTTTCTCTTTAGCTCGGCTTATT
CCAGGGGTGTGTTTCGTCGAGATGCACACAAGAGTGAGGTTGCTCATCGGTTTAAAGATTTGGGAGAAGAAAATTTCAAAGCCTTGGTGTT
GATTGCCTTTGCTCAGTATCTTCAGCAGTGTCCATTTGAAGATCATGTAAAATTAGTGAATGAAGTAACTGAATTTGCAAAAACATGTGTT
GCTGATGAGTCAGCTGAAAATTGTGACAAATCACTTCATACCCTTTTTGGAGACAAATTATGCACAGTTGCAACTCTTCGTGAAACCTATG
GTGAAATGGCTGACTGCTGTGCAAAACAAGAACCTGAGAGAAATGAATGCTTCTTGCAACACAAAGATGACAACCCAAACCTCCCCCGATT
GGTGAGACCAGAGGTTGATGTGATGTGCACTGCTTTTCATGACAATGAAGAGACATTTTTGAAAAAATACTTATATGAAATTGCCAGAAGA
CATCCTTACTTTTATGCCCCGGAACTCCTTTTCTTTGCTAAAAGGTATAAAGCTGCTTTTACAGAATGTTGCCAAGCTGCTGATAAAGCTG
CCTGCCTGTTGCCAAAGCTCGATGAACTTCGGGATGAAGGGAAGGCTTCGTCTGCCAAACAGAGACTCAAGTGTGCCAGTCTCCAAAAATT
TGGAGAAAGAGCTTTCAAAGCATGGGCAGTAGCTCGCCTGAGCCAGAGATTTCCCAAAGCTGAGTTTGCAGAAGTTTCCAAGTTAGTGACA
GATCTTACCAAAGTCCACACGGAATGCTGCCATGGAGATCTGCTTGAATGTGCTGATGACAGGGCGGACCTTGCCAAGTATATCTGTGAAA
ATCAAGATTCGATCTCCAGTAAACTGAAGGAATGCTGTGAAAAACCTCTGTTGGAAAAATCCCACTGCATTGCCGAAGTGGAAAATGATGA
GATGCCTGCTGACTTGCCTTCATTAGCTGCTGATTTTGTTGAAAGTAAGGATGTTTGCAAAAACTATGCTGAGGCAAAGGATGTCTTCCTG
GGCATGTTTTTGTATGAATATGCAAGAAGGCATCCTGATTACTCTGTCGTGCTGCTGCTGAGACTTGCCAAGACATATGAAACCACTCTAG
AGAAGTGCTGTGCCGCTGCAGATCCTCATGAATGCTATGCCAAAGTGTTCGATGAATTTAAACCTCTTGTGGAAGAGCCTCAGAATTTAAT
CAAACAAAATTGTGAGCTTTTTGAGCAGCTTGGAGAGTACAAATTCCAGAATGCGCTATTAGTTCGTTACACCAAGAAAGTACCCCAAGTG
TCAACTCCAACTCTTGTAGAGGTCTCAAGAAACCTAGGAAAAGTGGGCAGCAAATGTTGTAAACATCCTGAAGCAAAAAGAATGCCCTGTG
CAGAAGACTATCTATCCGTGGTCCTGAACCAGTTATGTGTGTTGCATGAGAAAACGCCAGTAAGTGACAGAGTCACCAAATGCTGCACAGA
ATCCTTGGTGAACAGGCGACCATGCTTTTCAGCTCTGGAAGTCGATGAAACATACGTTCCCAAAGAGTTTAATGCTGAAACATTCACCTTC
CATGCAGATATATGCACACTTTCTGAGAAGGAGAGACAAATCAAGAAACAAACTGCACTTGTTGAGCTCGTGAAACACAAGCCCAAGGCAA
CAAAAGAGCAACTGAAAGCTGTTATGGATGATTTCGCAGCTTTTGTAGAGAAGTGCTGCAAGGCTGACGATAAGGAGACCTGCTTTGCCGA
GGAGGGTAAAAAACTTGTTGCTGCAAGTCAAGCTGCCTTAGGCTTATAACATCTACATTTAAAAGCATCTCAGCCTACCATGAGAATAAGA
GAAAGAAAATGAAGATCAAAAGCTTATTCATCTGTTTTTCTTTTTCGTTGGTGTAAAGCCAACACCCTGTCTAAAAAACATAAATTTCTTTA
ATCATTTTGCCTCTTTTCTCTGTGCTTCAATTAATAAAAAATGGAAAGAATCTAATAGAGTGGTACAGCACTGTTATTTTTCAAAGATGTG
TTGCTATCCTGAAAATTCTGTAGGTTCTGTGGAAGTTCCAGTGTTCTCTCTTATTCCACTTCGGTAGAGGATTTCTAGTTTCTGTGGGCTA
ATTAAATAAATCACTAATACTCTTCTAAGTT

Figure 2l

>gi|34577108|ref|NM_000034.2| Homo sapiens aldolase A, fructose-bisphosphate (ALDOA), transcript variant 1, mRNA

CCTAGCTTGGCGGAATCCGTGAATTGCCCGGCGGAGGGTGCAGCTCCCGGACTGACTTGGCCTCTGGCCTTCCCCATGGAGGCCTCCT
CTAGCCCGTGGAATCCAACCCCGGCTCCTGTCAGCAGCCCTCCCCTGCTCCCCATCCCTGCCATCGTCTTCATGCCTGTGGGCATCTA
TTTGTTGCTGCTGGGTCTAGTCCTGCTGCTGGCCCAGGGCTGCTGGCGCGACGGTAGCTCCCCCTGCAGGAAGCAA
GGTTCCTCCGGGCCCCCAGACTGCTGCTGGACCTGTGCAGAAGCCTGCAACTTTCCTCTGCCTAGCCGGGCCCACTTCCTGGATGCTTGCT
GCCCCCAGCCCACCAGAGCTGACTGGGCACCTCGCTGCCCCGCTGCTGCCCACTCTGCGACTGTGCCTGTACGGTGCCAGCTCCCCGACTG
CCAGAGCCCTCAACTGTCTCTTGCCTTCGAGATCAAGCTCCGATGAGGACCCAGGGCCCCTGCCCTCTGGGGAGCGGCCAGCCCCCAGGGCCCA
TGTGCCCCTCCCTCCCTGAAGAGCCTTTCCCCACCGCCACTGGAACCACACCCAGGCCTGGGAACTGGCAGC
GCAGGGCCTGGCTCCCTGCAGGGCAGGACTCTTGGCCGGCTGGAGGGCCAGCTCCTCTGGAGGGCCAGAAAAGAGAGGGCTAGTGCTCGGGG
CAGGTGCCCTGGCTTCCCCTTCCCCTCCACACGATCAACGATTCTATTTGAAGTTGGGCAGGGGGGTGGCGCTGCTCACCACACACAAGTGTT
ATAGGAGGAGTCTGGCCCTTGAGTACGGTACGGCAGGGGTGCCTCAACCACACTCCCGTCCACGGACTCTCCGTTATTTTAGGAGGTCCCT
GGCCAAAGATTTATTTCTTTGACAACCAAGGGCCTCCGTCTGGATTTCCAAGGAAGAATTTCCTCTGAAGCACCGGAACTTGCTACTACC
AGCACCATGCCCTACCAATATCCAGCACTGACCCCGGAGCAGAAGAAGGAGCTGTCTGACATCGCTCACGGCATCGTGGCACCTGGCAAGG
GCATCCTGGCTGCAGATGAGTCCACTGGGAGCATTGCCCAAGCGGCTGCAGTCCATTGGCACCGAGAACACCGGAGGAGAACCGGGCGCTTCTA
CCGCCAGCCTGCTGCTGACAGCTGACGACCGCGTGAACCCCTGCATTTGGGGTGTCATCCTCTTCCATGAGACACTCTACCAGAAGGCGGAT
GATGGGCGTCCCCTTCCCCAAGTTATCAAATCCAAGGGCGGTGTTGGGCATCAAGGTAGACAAGGGCGTGGTCCCCCTGGCAGGGACAA
ATGGCGAGACTACCACCCAAGGGTTGGATGGGCTGTCTGAGCGCTGTGCCCAGTACAAGAAGGACGGAGCTGACTTCGCCAAGTGGCGTTG
TGTGCTGAAGATTGGGGAACACACCCCCTCAGCCCTCGCCATCATGGAAAATGCCAATGTTCTGGCCGTTATGCCAGTATCTGCCAGCAG
AATGGCATTGTGCCCATCGTGGAGCCTGAGATCCTCCTCGATGGGGACCATGACTTGAAGCGCTGGCCAGTATGTGACCGACAAGGTGCCTGG
CTGCTGTCTACAAGGCTCTGAGTGACCACCATCTACCTGGAAGGCACCTTGCTGAAGCCCAACATGGTCACCCCAGGCCATGCTTGCAC
TCAGAAGTTTTCTCATGAGGAGATTGCCATGGCGACCGTCACACAGCGCTGCGCCCCCGCTGTCACTGGGATCACCTTCCTG
TCTGGAGGCCAGAGTGAGGAGGAGGCGTCCATCAACCTCAATGCCATTAACAAGTGCCCCCTGCTGAAGCCCTGGGCCCTGACCTTCTCCT
ACGGCCGAGCCCTGCAGGCCTCTGCCTGAAGGCCTGGGGCGGGAAGAAGGAGACCTGAAGGCTGCGCAGGAGGAGTATGTCAAGCGAGC
CCTGGCCAACAGCCTTGCCTGTCAAGGAAAGTACACTCCGAGCGGTCAGGCTGGGGCTGCTGCCAGCGAGTCCCTCTTGGTCTCTAACCAC
GCCTATTAAGCGGAGGTGTCCCAAGCTGCCCCAACACTCCCAGGCCCTCCCCACTCTTGAAGAGGAGCCGCCCTCCTCGGGGCT
CCAGGCTGGCTTGCCCGCGCTCTTTCTTCCCTCCGTGACAGTGGTGTCTGTGAATGCTAAGTCCATCACCCTTTCCGGCACA
CTGCCAAATAAACAGCTATTTAAGGGGG

Figure 2m

>gi|50345983|ref|NM_001001937.1| Homo sapiens ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle (ATP5A1), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA

TCTGGCATTGCAAGCCTCGCTTCGTTGCCACTTCCCAGCTCTTCCCGCCTTCCGCGGTATAATCAACACTACGAGAGATAGAGCCGCCTAG
AACCAGTCCGGAGGCTGCGGCTGCAGAAGTACCGCCTGCGGAGTAACTGCAAAGATGCTGTCCGTGCGGCGTTGCTGCGGCCGTGGTCCGCG
CCCTTCCTCGGCGGGCCGGACTGGTCTCCAGAAATGCTTTGGGTTCATCTTTCATTGCTGCAAGGAACTTCCATGCCTCTAACACTCATCT
TCAAAAGACTGGGACTGCTGAGATGTCCTCTATTCTTGAAGAGCGTATTCTTGGAGCTGATACCTCTGTTGATCTTGAAGAAACTGGGCGT
GTCTTAAGTATTGGTGATGGTATTGCCCGCGTACATGGGCTGAGGAATGTTCAAGCAGAAGAAATGGTAGAGTTTTCTTCAGGCTTAAAGG
GTATGTCCTTGAACTTGGAACCTGACAATGTTGGTGTTGTCGTGTTTGGAAATGATAAACTAATTAAGGAAGGAGATATAGTGAAGAGGAC
AGGAGCCATTGTGGACGTTCCAGTTGGTGAGGAGCTGTTGGGTCGTGTAGTTGATGCCCTTGGTAATGCTATTGATGGAAAGGGTCCAATT
GGTTCCAAGACGCGTAGGCGAGTTGGTCTGAAAGCCCCCGGTATCATTCCTCGAATTTCAGTGCGGGAACCAATGCAGACTGGCATTAAGG
CTGTGGATAGCTTGGTGCCAATTGGTCGTGGTCAGCGTGAACTGATTATTGGTGACCGACAGACTGGGAAAACCTCAATTGCTATTGACAC
AATCATTAACCAGAAACGTTTCAATGATGGATCTGATGAAAAGAAGAAGCTGTACTGTATTTATGTTGCTATTGGTCAAAAGAGATCCACT
GTTGCCCAGTTGGTGAAGACACTTACAGATGCAGATGCCATGAAGTACACCATTGTGGTGTCGGCTACGGCCTCGGATGCTGCCCCACTTC
AGTACCTGGCTCCTTACTCTGGCTGTTCCATGGGAGAGTATTTTAGAGACAATGGCAAACATGCTTTGATCATCTATGACGACTTATCCAA
ACAGGCTGTTGCTTACCGTCAGATGTCTCTGTTGCTCCGCCGACCCCCTGGTCGTGAGGCCTATCCTGGTGATGTGTTCTACCTACACTCC
CGGTTGCTGGAGAGAGCAGCCAAAATGAACGATGCTTTTGGTGGTGGCTCCTTGACTGCTTTGCCAGTCATAGAAACACAGGCTGGTGATG
TGTCTGCTTACATTCCAACAAATGTCATTTCCATCACTGACGGACAGATCTTCTTGGAAACAGAATTGTTCTACAAAGGTATCCGCCCTGC
AATTAACGTTGGTCTGTCTGTATCTCGTGTCGGATCCGCTGCCCAAACCAGGGCTATGAAGCAGGTAGCAGGTACCATGAAGCTGGAATTG
GCTCAGTATCGTGAGGTTGCTGCTTTTGCCCAGTTCGGTTCTGACCTCGATGCTGCCACTCAACAACTTTTGAGTCGTGGCGTGCGTCTAA
CTGAGTTGCTGAAGCAAGGACAGTATTCTCCCATGGCTATTGAAGAACAAGTGGCTGTTATCTATGCGGGTGTAAGGGGATATCTTGATAA
ACTGGAGCCCAGCAAGATTACAAAGTTTGAGAATGCTTTCTTGTCTCATGTCGTCAGCCAGCACCAAGCCTTGTTGGGCACTATCAGGGCT
GATGGAAAGATCTCAGAACAATCAGATGCAAAGCTGAAAGAGATTGTAACAAATTTCTTGGCTGGATTTGAAGCTTAAACTCCTGTGGATT
CACATCAAATACCAGTTCAGTTTTGTCATTGTTCTAGTAAATTAGTTCCATTTGTAAAAGGGTTACTCTCATACTCCTTATGTACAGAAAT
CACATGAAAAATAAAGGTTCCATAATGCATAGTT

Figure 2n

>gi|20428653|ref|NM_001743.3| Homo sapiens calmodulin 2 (phosphorylase kinase, delta) (CALM2), mRNA

AGTCCGAGTGGAGAGAGAGCGAGCTGAGTGGTTGTTGTCGCGTCTCGGAAACCGGTAGCGCTTGGCAGCATGGCTGACCAACTGACTGAAGA
GCAGATTGCAGAATTCAAAGAAGCTTTTTCACTATTTGACAAAGATGGTGATGGAACTATAACAACAAAGAATTGGGAACTGTAATGAGA
TCTCTTGGGCAGAATCCCACAGAAGCAGAGTTACAGGACATGATTAATGAAGTAGATGCTGATGGTAATGGCACAATTGACTTCCCTGAAT
TTCTGACAATGATGGCAAGAAAAATGAAAGACACAGACAGTGAAGAAGAAATTAGAGAAGCATTCCGTGTTTTGATAAGGATGGCAATGG
CTATATTAGTGCTGCAGAACTTCGCCATGTGATGACAAACCTTGGAGAGAAGTTAAGAAGAAGTTGATGAAATGATCAGGGAAGCA
GATATTGATGGTGATGGTCAAGTAAACTATGAAGAGTTTGTACAAATGATGAAAGACCTTGTACAGAATGTGTTAAATTTC
TTGTACAAAAATTGTTTATTTGCCTTTTCTTTGTTTTGTTTTCTCCCTACTGTCAAAAAAATATGCATGTATAGTA
ATTAGGACTTCATTCCTCCATGTTTTCTTTCCCTTATCTTACTTGTCATTGTCCTAAAACCTTATTTTTAGAAAATTGATCAAGTAACATGTTG
CATGTGGCTTACTCTGGATATATCTAAGCCCTTCTGCACATCTAAACTTAGATGGAGTTGGTCAAATGAGGGAACATCTGGGTTATGCCTT
TTTTAAGGTAGTTTTCTTTTAGGAACTGTTGCAGCATGTTGTTGTTGAAGTGTGGAGTTGTAACTCTGCGTGGACTATGGACAGTCAACAATAT
GTACTTAAAAAGTTGCACTATTTGCAAAACGGGTGTATTATCCAGGTACTCGTACACTATTTTTTTGTACTGCTGGTCCTGTACCAGAAACAT
TTTCTTTTTATTGTTACTTGCTTTTTAAACTTTGTTTAGCCACTTTAAAATCTGCTTATGGCCTCAAAAATCCATTCCAAGTTGT
ATATTTGTTTTCCAATAAAAAAATTACAATTTACCC

Figure 2o

>gi|50318561ref|NM_005566.1| Homo sapiens lactate dehydrogenase A (LDHA), mRNA

TGCTGCAGCCGCTGCCGCCGGATTCCGGATCTCATTGCCACGCCGCCCGACGACCGCCGACGTGCATTCCCTTTGGTTCCAAG
TCCAATATGGCAACTCTAAAGGATCAGCTGATTTATAATCTTCTAAAGGAAGAACAGACCCCCCAGAATAAGATTACAGTTGTTGGGTTG
GTGCTGTTGGCATGGCCTGTGCCATCAGTATCTTAATGAAGGACTTGCTCTTGTTGATGTCATCGAAGACAAATTGAA
GGGAGAGATGGATCTCCAACATGGCAGCCTTTTCCTTAGAACACCAAAGATTGTCTCTGGCAAAGACTATAAATGTAACTGCAAACTCC
AAGCTGGTCATTATCACGGCTGGGGCACGTCAGCAAGAGGGAAAGCCGTCTTAATTTGGTCCAGCGTAACGTGACATATTTAAATTCA
TCATTCCTAATGTTGTAAAATACAGCCCGAACTGCAAGTTGCTTATTGTTTCAAATCCAGTGGATATCTTGACCTACGTGGCTTGGAAGAT
AAGTGGTTTTTCCCAAAAACCGTGTTATTGGAAGTGGTTGCAATCTGGATTCCAGCCCGATTCCGTTACCTGATGGGGGAAAGGCTGGGAGTT
CACCCATTAAGCTGTCATGGGTGGGTCCTTGGGGAACATGGAGATTCCAGTGTGCCTGTATGGAGTGGAATGAATGTTGCTGGTGTCTCTC
TGAAGACTCTGCACCCAGATTTAGGGACTGATAAAGATAAGGAAACAGTGGAAAGAGGTTCACAAGCAGGTGGTTGAGAGTGCTTATGAGGT
GATCAAACTCAAAGGCTACACATCCTGGCTATTGGACTCTCTGTAGCAGATTTGGCAGAGATATAATGAAGAATCTTAGGCGGGTGCAC
CCAGTTTCCACCATGATTAAGGGTCTTTACGGAATAAAGGATGATGTCTTCCTTAGTGTTCCATTTTGGACAGAATGGAATCTCAG
ACCTTGTGAAGGTGACTCTGACTTCTGAGGAAGAGGCCCGTTTGAAGATACACTTTGGGGGATCCAAAAGGAGCTGCAATT
TTAAAGTCTTCTGATGTCATATCATTTCACTGTCTAGGCTACAACAGGATTCTAGGTGGAGGTTGTGCATGTTGTCCTTTTTATCTGATCT
GTGATTAAAGCAGTAATATTTTAAGATGGACCTGGGAAAAACATCAACTCCTGAAGTTAGAAAATAAGAATGGTTTGTAAAAATCCACAGCTAT
ATCCTGATGCTGGATGGTATTAATCTTGTGTAGTCTTCAACTGGTTAGTGTGAAATAGTTCTGCCACCTCTGACGCACCACTGCCAATGCT
GTACGTACTGCATTTGCCCCTTGAGCCAGGTGGATGTTTACCGTGTGTTATATAACTTCCTGGCTCCTTCACTGAACATGCCTAGTCCAAC
ATTTTTTCCCAGTGAGTCACATCCTGGGATCCAGTGTATAAATCCAATATCATGTCTTGTGCATAATTCTTCCAAAGGATCTTATTTTGTG
AACTATATCAGTAGTGTACATTACCATATAATGTAAAAAGATCTACATACAAACAATGCAACCAACTATCCAAGTGTTATACCAACTAAAA
GCCCCAAATAAACCTTGAACAGTG

Figure 2p

>gi|52851446|ref|NM_000365.4| Homo sapiens triosephosphate isomerase 1 (TPI1), mRNA

CCTTCAGCGCCTCGCTCCAGCGCCATGGCGCCCTCCAGGAAGTTCTTCGTTGGGGGAAACTGGAAGATGAACGGGCGGGAAGCAGAGTCTG
GGGGAGCTCATCGGCACTCTGAAGGCGGCCAAGGTGCCGGCCGACACCGAGGTGGTTTGTGCTCCCCCTACTGCCTATATCGACTTCGCCC
GGCAGAAGCTAGATCCCAAGATTGCTGTGGCTGCGCAGAACTGCTACAAAGTGACTAATGGGGCTTTTACTGGGGAGATCAGCCCTGGCAT
GATCAAAGACTGCGGAGCCACGTGGGTGCCTGGGGCACTCAGAGAGAAGGCATGTCTTTGGGGAGTCAGATGAGCTGATTGGGCAGAAA
GTGGCCCATGCTCTGGCAGAGGGACTCGGAGTAATCGCCTGCATTGGGGAGAAGCTAGATGAAAGGGCTGGCATCACTGAGAAGGTTG
TTTCGAGCAGACAAAGGTCATCGCAGATAACGGTGAAGGACTGGAGCAAGGTCGTCCTGGCCTATGAGCCTGTGTGGGCCATTGGTACTGG
CAAGACTGCAACACCCCAACAGGCCCAGGAAGTACACGAGAAGCTGCGAGGATGGCCTGAAGGTCCAACGGTCTCTGATGCGGTGGCTCAGAGC
ACCCGTATCATTTATGGAGGCTCTGTGACTGGGGCAACCTGCAAGGAGCTGGCCAGGCCAGCCTGATGTGGATGGCTTCCTTGTGGGTGGTG
CTTCCCTCAAGCCCGAATTCGTGGACATCATCAATGGCCAAAACAATGAGCCCCATCCATCTTCCCTACCCTTCCTGCCAAGCCAGGGACTAA
GCAGCCCAGAAGCCCAGTAACTGCCCTTTCCCTGCATATGCTTCTTGGTGTCATCTGCTCCTTCCTGTGGCCTCATCCAAACTGTATCT
TCCTTTACTGTTTATATCTTCACCCTGTAATGGTTGGACCAGGCCAATCCCTTCTCCACTTACTATAATGGTTGGAACTAAACGTCACCA
AGGTGGCTTCTCCTTGGCTGAGAGATGAAGGCGTGGTGGGATTTGCTCCTGGGTTCCCTAGTGAGGGCAGAAGAGAAACCATC
CTCTCCCTTCTTACACCGTGAGGCCAAGATCCCCTCAGAAGGCAGGAGTGCTGCCCTCTCCCATGGTGCCCGTGCCTCTGTGCTGTATG
TGAACCACCCATGTGAGGGAATAAACCTGGCACTAGG

Figure 2q

TPT1 = Tumor protein, translationnally controlled, 1, 830bp, chr 13 q12-q14.

VIM = Vimentin,1847bp, chr 10 p13.

FTH1 = Ferritin, heavy polypeptide 1, 1228bp, chr 11 q13.

RPS4X = Ribosomal proteinS4, X-linked, 955bp, chr X q13.1.

RPL7A = Ribosomal protein L7a, 890bp, chr 9 q34.

ENO1 = Enolase 1, 1812bp, chr 1 p36.3-p36.2.

HSPA8 = Heat shock 70kDa protein 8, 2260bp, chr 11 q24.1.

GAPDH = Glyceraldehyde-3-phosphate dehydrogenase, 1310bp, chr 12p13.

TMSB4X = Thymosin, beta 4, X linked, 627bp, chr X q21.3-q22.

RPS6 = Ribosomal protein S6, 829bp, chr 9 p21.

FTL = Ferritin, light polypeptide, 870bp, chr 19 q 13.3-q13.4.

ALB = Albumin, 2215bp, chr 4 q11-q13.

ALDOA = Aldolase A, fructose-bisphosphate, 2303bp, chr 16q22-q24.

ATP5A1 = ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle, 1945bp, chr 18q12-q21.

CALM2 = Calmodulin 2 (phosphorylase kinase, delta), 1128bp, chr 2p21.

LDHA = Lactate dehydrogenase A, 1661bp, chr 11p15.4.

TPI1 = Triosephosphate isomerase 1, 1220bp, chr 12p13.

Figure 2r

```
>gi|47009767|emb|BX437546.2| BX437546 Homo sapiens THYMUS Homo sapiens cDNA clone CS0CAP007YK06
          5-PRIME, mRNA sequence.
          Length = 868

 Score = 1531 bits (829), Expect = 0.0
 Identities = 829/829 (100%)
 Strand = Plus / Plus


Query: 2   cccccgagcgccgctccggctgcaccgcgctcgctccgagtttcaggctcgtgctaagc 61
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 12  cccccgagcgccgctccggctgcaccgcgctcgctccgagtttcaggctcgtgctaagc 71


Query: 62  tagcgccgtcgtcgtctccttcagtcgccatcatgattatctaccgggacctcatcagc 121
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 72  tagcgccgtcgtcgtctccttcagtcgccatcatgattatctaccgggacctcatcagc 131


Query: 122 cacgatgagatgttctccgacatctacaagatccgggagatcgcggacgggttgtgcctg 181
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 132 cacgatgagatgttctccgacatctacaagatccgggagatcgcggacgggttgtgcctg 191


Query: 182 gaggtggaggggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 241
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 192 gaggtggaggggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 251


Query: 242 ggaaatgcctccgctgaaggccccgagggcgaaggtaccgaaagcacagtaatcactggt 301
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 252 ggaaatgcctccgctgaaggccccgagggcgaaggtaccgaaagcacagtaatcactggt 311


Query: 302 gtcgatattgtcatgaaccatcacctgcaggaaacaagtttcacaaaagaagcctacaag 361
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 312 gtcgatattgtcatgaaccatcacctgcaggaaacaagtttcacaaaagaagcctacaag 371
```

Figure 3a

```
Query: 362 aagtacatcaaagattacatgaaatcaaatcaaagggaaaacttgaagaacagaggaccagaa 421
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 372 aagtacatcaaagattacatgaaatcaaatcaaagggaaaacttgaagaacagaggaccagaa 431

Query: 422 agaagtaaaaaccttttatgacagggcctgcagaacaaatcaagcacatcctttgctaaattc 481
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 432 agaagtaaaaaccttttatgacagggcctgcagaacaaatcaagcacatcctttgctaaattc 491

Query: 482 aaaactaccagtttcttattggtgaaaacatgaatccagatggcatggtttgctctattg 541
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 492 aaaactaccagtttcttattggtgaaaacatgaatccagatggcatggtttgctctattg 551

Query: 542 gactaccgtgaggatggtgtgacccatatatgatttctttaaggatggtttagaaatg 601
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 552 gactaccgtgagggatggtgtgacccatatatgatttctttaagggatggtttagaaatg 611

Query: 602 gaaaaatgttaacaaatggtggcaatttatttggatctatcacctgtcatcataactggct 661
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 612 gaaaaatgttaacaaatggtggcaatttatttggatctatcacctgtcatcataactggct 671

Query: 662 tctgcttgtcatcacacaacaccaggacttaagacaaaatgggactgatgtcatcttgag 721
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 672 tctgccttgtcatcacacaacaccaggacttaagacaaaatgggactgatgtcatcttgag 731

Query: 722 ctcttcatttatttgactgtgatttatttggagtggaggcattgtttttaagaaaaaca 781
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 732 ctcttcatttatttgactgtgatttatttggagtggaggcattgtttttaagaaaaaca 791

Query: 782 tgtcatgtaggtgtcttaaaaatataaaatgcatttaaactcatttgagag 830
           ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 792 tgtcatgtaggtgtcttaaaaatataaaatgcatttaaactcatttgagag 840
```

Figure 3b

```
>gi|45716222|emb|AL540618.3| AL540618 Homo sapiens PLACENTA Homo sapiens cDNA clone CS0DE002YF2 5
          5-PRIME, mRNA sequence.
         Length = 952

 Score = 1531 bits (829), Expect = 0.0
 Identities = 829/829 (100%)
 Strand = Plus / Plus

Query: 2   cccccgagcgccgctccggctgcaccgcgctcgctccgagtttcaggctcgtgctaagc 61
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 18  cccccgagcgccgctccggctgcaccgcgctcgctccgagtttcaggctcgtgctaagc 77

Query: 62  tagcgccgtcgtcgtctcccttcagtcgccatcatgattatctaccgggacctcatcagc 121
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 78  tagcgccgtcgtcgtctcccttcagtcgccatcatgattatctaccgggacctcatcagc 137

Query: 122 cacgatgagatgttctccgacatctacaagatccgggagatcgcggacgggttgtgcctg 181
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 138 cacgatgagatgttctccgacatctacaagatccgggagatcgcggacgggttgtgcctg 197

Query: 182 gaggtggaggggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 241
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 198 gaggtggaggggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 257

Query: 242 ggaaatgcctccgctgaaggccccgagggcgaaggtaccgaaagcacagtaatcactggt 301
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 258 ggaaatgcctccgctgaaggccccgagggcgaaggtaccgaaagcacagtaatcactggt 317

Query: 302 gtcgatattgtcatgaaccatcacctgcaggaaacaagtttcacaaaagaagcctacaag 361
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 318 gtcgatattgtcatgaaccatcacctgcaggaaacaagtttcacaaaagaagcctacaag 377

Query: 362 aagtacatcaaagattacatgaaatcaatcaaagggaaacttgaagaacagagaccagaa 421
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 378 aagtacatcaaagattacatgaaatcaatcaaagggaaacttgaagaacagagaccagaa 437
```

Figure 3c

```
Query:  482  aaaaactaccagttctcttattggtgaaaacatgaatccagatggcatggtgctctattg  541
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  498  aaaaactaccagttctcttattggtgaaaacatgaatccagatggcatggtgctctattg  557

Query:  542  gactaccgtgaggatggtgtgaccccatatatgatttctcttaaggatggtttagaaatg  601
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  558  gactaccgtgaggatggtgtgaccccatatatgatttctcttaaggatggtttagaaatg  617

Query:  602  gaaaaatgtttaacaaatggccaatttatttggaatctatcacctgtcatcataactggct  661
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  618  gaaaaatgtttaacaaatggccaatttatttggaatctatcacctgtcatcataactggct  677

Query:  662  tctggcttgtccatccacacaaccaaccaggacttaagacaaaatgggactgatggtccatcttcgag  721
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  678  tctggcttgtccatccacacaaccaaccaggacttaagacaaaatgggactgatggtccatcttcgag  737

Query:  722  cttttcatttatttgactgtgatttatttggagtgaaggcattgttttaaggaaaaaaca  781
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  738  cttttcatttatttgactgtgatttatttggagtgaaggcattgttttaaggaaaaaaca  797

Query:  782  tgtcatgtaggttgtccaaaaatcaaaaatgcatttaaactcatttgaggag  830
             ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  798  tgtcatgtaggttgtccaaaaatcaaaaatgcatttaaactcatttgaggag  846
```

Figure 3d

```
>gi|45696213|emb|AL520698.3| AL520698 Homo sapiens NEUROBLASTOMA COT 10-NORMALIZED Homo sapiens
          cDNA clone CS0DB002YF24 5-PRIME, mRNA sequence.
         Length = 909

 Score = 1530 bits (828), Expect = 0.0
 Identities = 828/829 (99%)
 Strand = Plus / Plus

Query: 2   cccccccgagcgccgctccggctgcaccgcgctcgctccgagtttcaggctcgtgctaagc 61
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 18  cccccccgagcgccgctccggctgcaccgcgctcgctccgagtttcaggctcgtgctaagc 77

Query: 62  tagcgccgtcgtcgtctcccttcagtcgccatcatgattatctaccgggacctcatcagc 121
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 78  tagcgccgtcgtcgtctcccttcagtcgccatcatgattatctaccgggacctcatcagc 137

Query: 122 cacgatgagatgttctccgacatctacaagatccgggagatcgcggacgggttgtgcctg 181
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 138 cacgatgagatgttctccgacatctacaagatccgggagatcgcggacgggttgtgcctg 197

Query: 182 gaggtggaggggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 241
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 198 gaggtggaggggaagatggtcagtaggacagaaggtaacattgatgactcgctcattggt 257

Query: 242 ggaaatgcctccgctgaaggccccgagggcgaaggtaccgaaagcacagtaatcactggt 301
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 258 ggaaatgcctccgctgaaggccccgagggcgaaggtaccgaaagcacagtaatcactggt 317

Query: 302 gtcgatattgtcatgaaccatcacctgcaggaaacaagtttcacaaaagaagcctacaag 361
           |||||||||||||||||||||||||||||||||||||||||||  ||||||||||||
Sbjct: 318 gtcgatattgtcatgaaccatcacctgcaggaaacaagtttcacaaaaraagcctacaag 377

Query: 362 aagtacatcaaagattacatgaaatcaatcaaagggaaacttgaagaacagagaccagaa 421
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 378 aagtacatcaaagattacatgaaatcaatcaaagggaaacttgaagaacagagaccagaa 437
```

Figure 3e

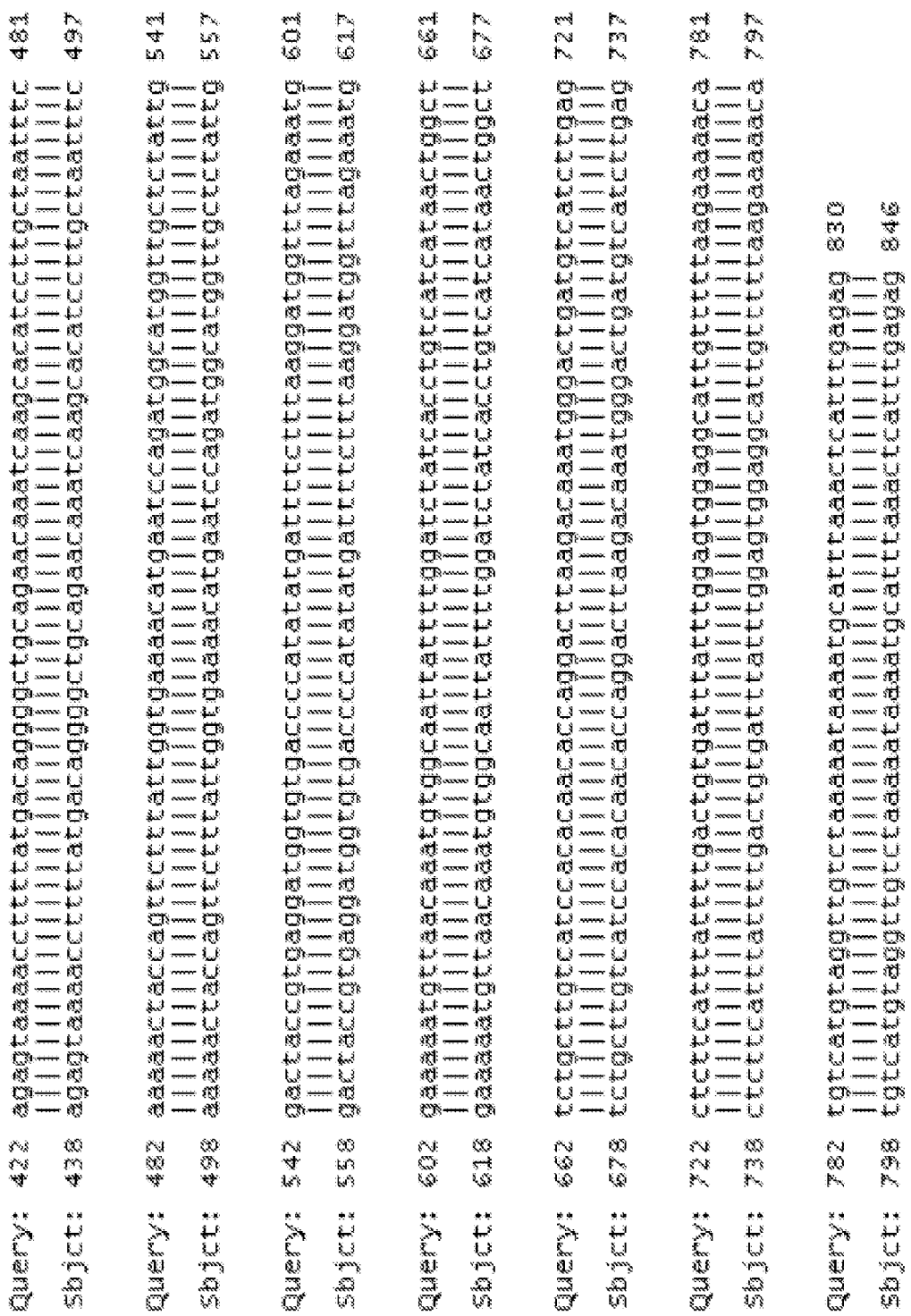


Figure 3f

```
>gi|10105790|gb|BE717525.1| RC4-HT0778-190700-014-e04 HT0778 Homo sapiens cDNA, mRNA sequence
          Length = 341

 Score =  564 bits (305), Expect = e-158
 Identities = 314/318 (98%), Gaps = 1/318 (0%)
 Strand = Plus / Plus


Query: 347 aaagaagcctacaagaagtacatcaaagattacatgaaatcaatcaaagggaaacttgaa 406
           |||||||| |||||||||||||||||| |||||||||||||||||||||||||||||||
Sbjct: 12  aaagaagcatacaagaagtacatcaaaga-tacatgaaatcaatcaaagggaaacttgaa 70


Query: 407 gaacagagaccagaaagagtaaaaccttttatgacaggggctgcagaacaaatcaagcac 466
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 71  gaacagagaccagaaagagtaaaaccttttatgacaggggctgcagaacaaatcaagcac 130


Query: 467 atccttgctaatttcaaaaactaccagttctttattggtgaaaacatgaatccagatggc 526
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 131 atccttgctaatttcaaaaactaccagttctttattggtgaaaacatgaatccagatggc 190


Query: 527 atggttgctctattggactaccgtgaggatggtgtgacccatatatgattttctttaag 586
           |||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 191 atggttgctctattgggctaccgtgaggatggtgtgacccatatatgattttctttaag 250


Query: 587 gatggtttagaaatggaaaaatgttaacaaatgtggcaattattttggatctatcacctg 646
           |||||||||||||||||||||||||||||||||||||||||| ||||||||||||||
Sbjct: 251 gatggtttagaaatggaaaaatgttaacaaatgtggcaattatttgggatctatcacctg 310


Query: 647 tcatcataactggcttct 664
           ||||||||||||||||||
Sbjct: 311 tcatcataactggcttct 328
```

Figure 3g

```
>gi|10224966|gb|BE771308.1| QV1-FT0082-210600-249-A12 FT0082 Homo sapiens cDNA, mRNA sequence
          Length = 338

 Score =  562 bits (304), Expect = e-157
 Identities = 316/321 (98%), Gaps = 3/321 (0%)
 Strand = Plus / Plus

Query: 346 aaaagaagcctacaagaa-gtacatca-aagattacatgaaaatcaatcaaagggaaactt 403
           ||||||||||||||||| |||||||| |||||||||||||||||||||||||||||||||
Sbjct: 6   aaaagaagcctacaagaacgtacatcagctgattacatg-aatcaatcaaagggaaactt 64

Query: 404 gaagaacagaggaccagaaaagtaaaaacctcctatgacagggggctgcagaaacaaatcaag 463
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 65  gaagaacagaggaccagaaaagtaaaaacctcctttatgacagggggctgcagaaacaaatcaag 124

Query: 464 cacatccttgctaatttcaaaaaactaccagttcttattgtgaaaacatgaatccagat 523
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 125 cacatccttgctaatttcaaaaaactaccagttcttactggtgaaaacatgaatccagat 184

Query: 524 ggcatggctgcctattggactaccgtgaggatggtgtgaccccatatatgatttcttt 583
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 185 ggcatggctgcctattggactaccgtgaggatggtgtgaccccatatatgatttcttt 244

Query: 584 aaggatggtttagaaatggaaaaatgttaacaaatggcaatttatttggatctatcac 643
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 245 aaggatggtttagaaatggaaaaatgttaacaaatggcaatttatttggatctatcac 304

Query: 644 ctgtcatcataactggcttct 664
           ||||||||||||||||||||
Sbjct: 305 ctgtcatcataactggcttct 325
```

Figure 3h

```
>gi|10246833|gb|BE814599.1| MRO-BN0070-070800-033-b04_1 BN0070 Homo sapiens cDNA, mRNA
          sequence.
         Length = 140

 Score =  259 bits (140), Expect = 2e-066
 Identities = 140/140 (100%)
 Strand = Plus / Plus


Query: 451  agaacaaatcaagcacatccttgctaatttcaaaaactaccagttctttattggtgaaaa 510
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1    agaacaaatcaagcacatccttgctaatttcaaaaactaccagttctttattggtgaaaa 60


Query: 511  catgaatccagatggcatggttgctctattggactaccgtgaggatggtgtgaccccata 570
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 61   catgaatccagatggcatggttgctctattggactaccgtgaggatggtgtgaccccata 120


Query: 571  tatgattttctttaaggatg 590
             ||||||||||||||||||||
Sbjct: 121  tatgattttctttaaggatg 140
```

Figure 3i

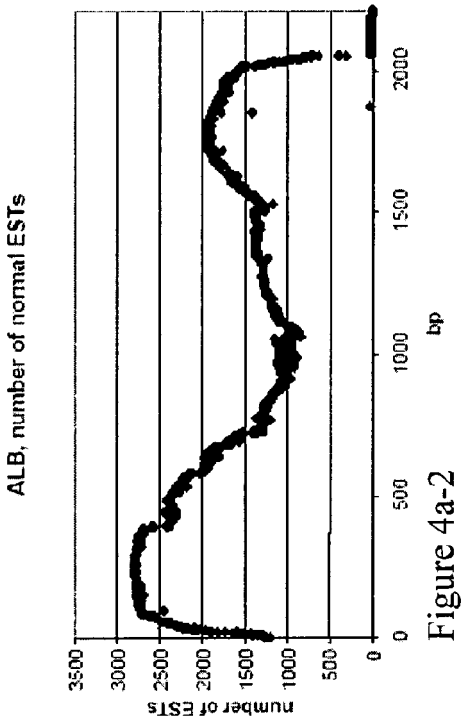


Figure 4a-2

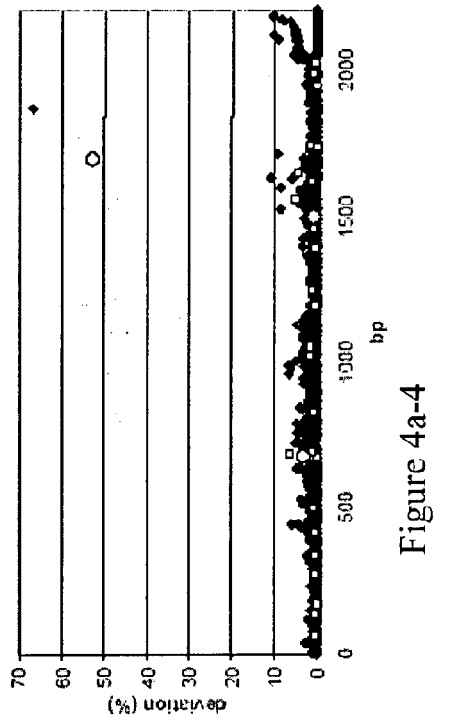


Figure 4a-4

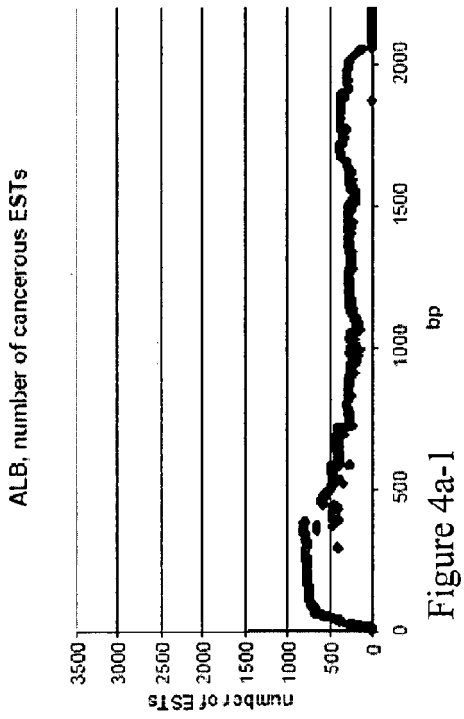


Figure 4a-1

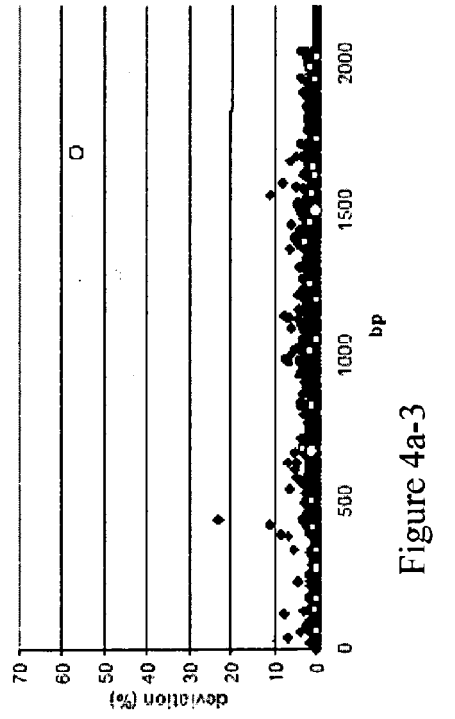


Figure 4a-3

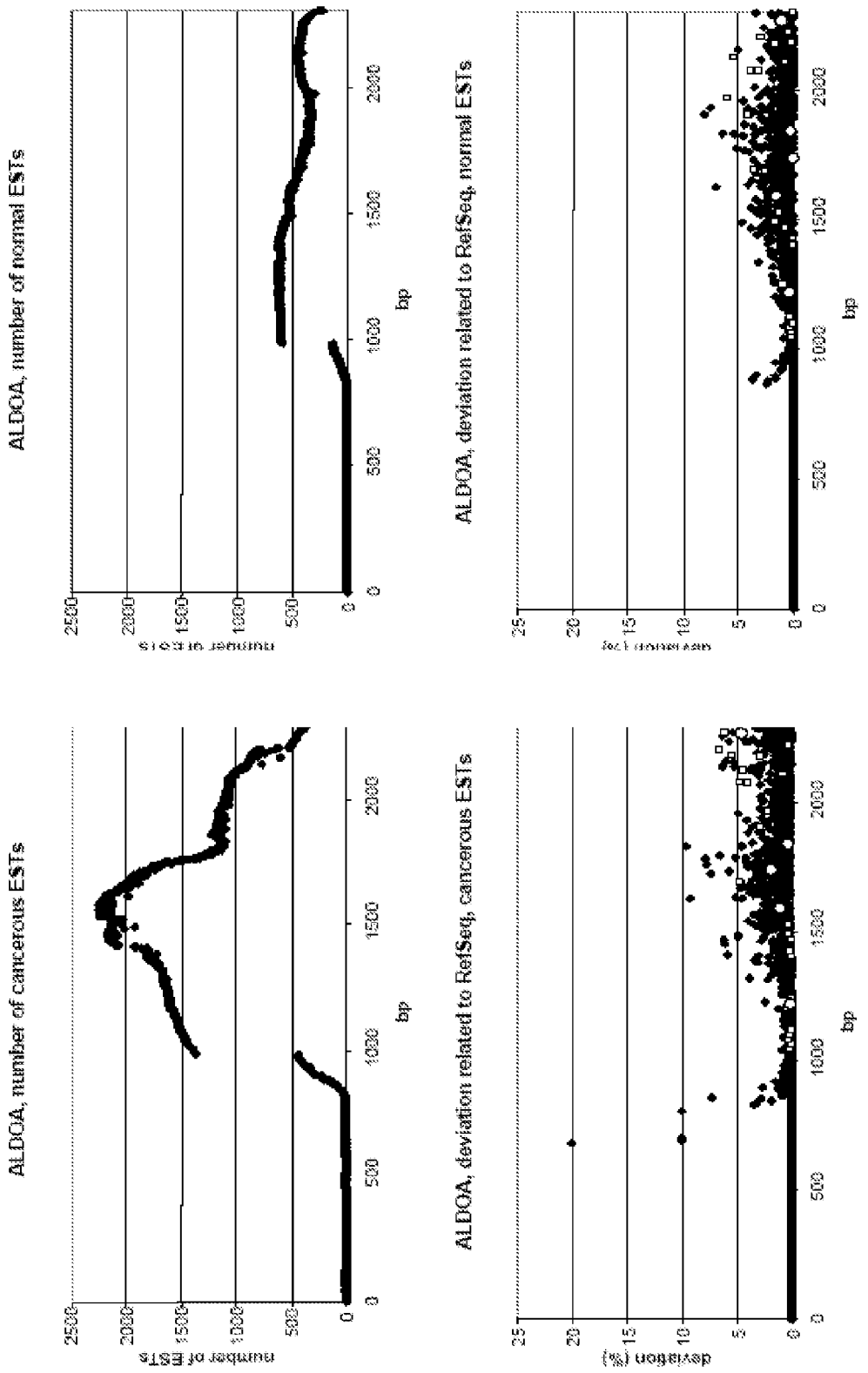


Figure 4a (following)

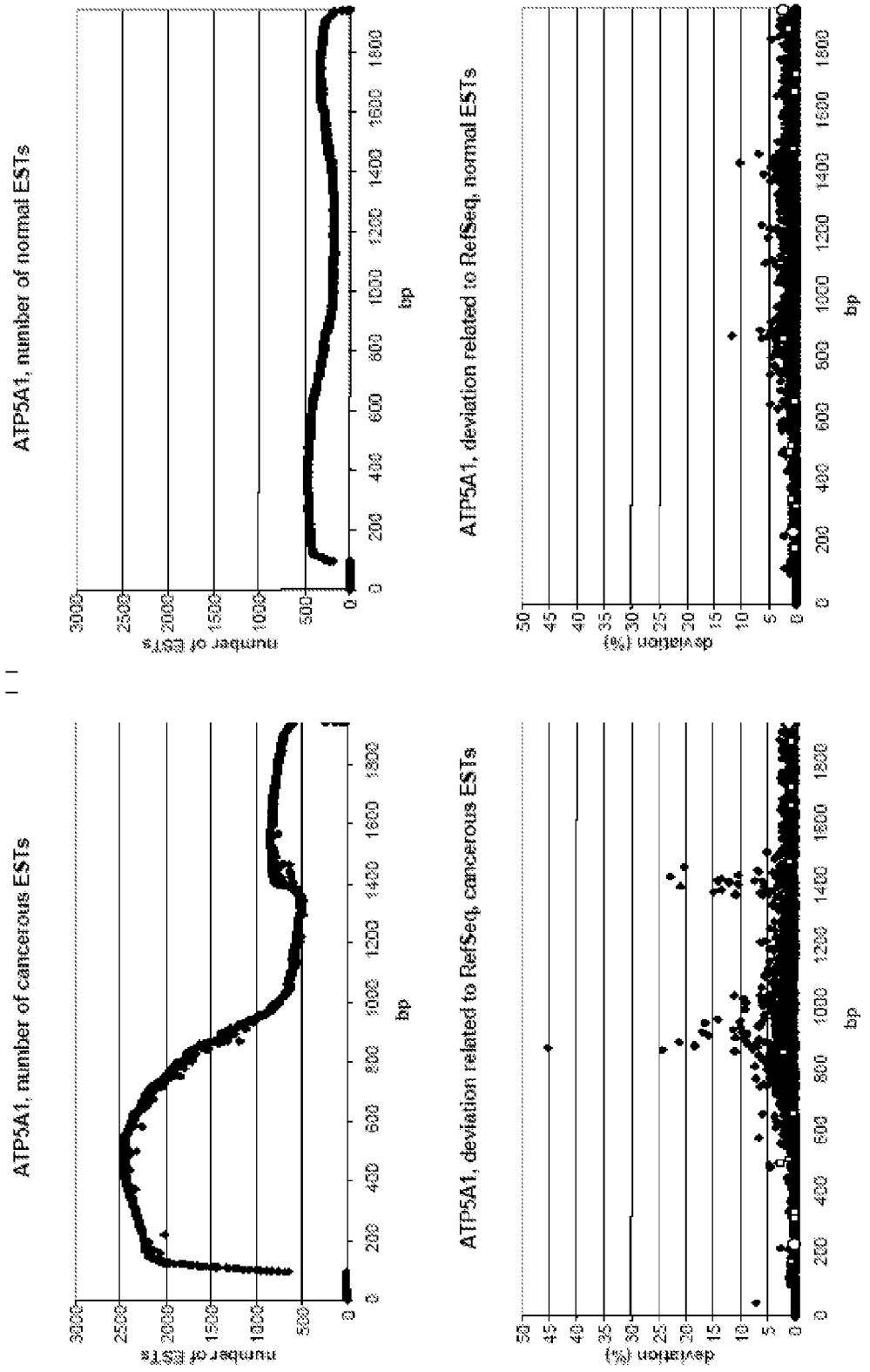


Figure 4a (following)

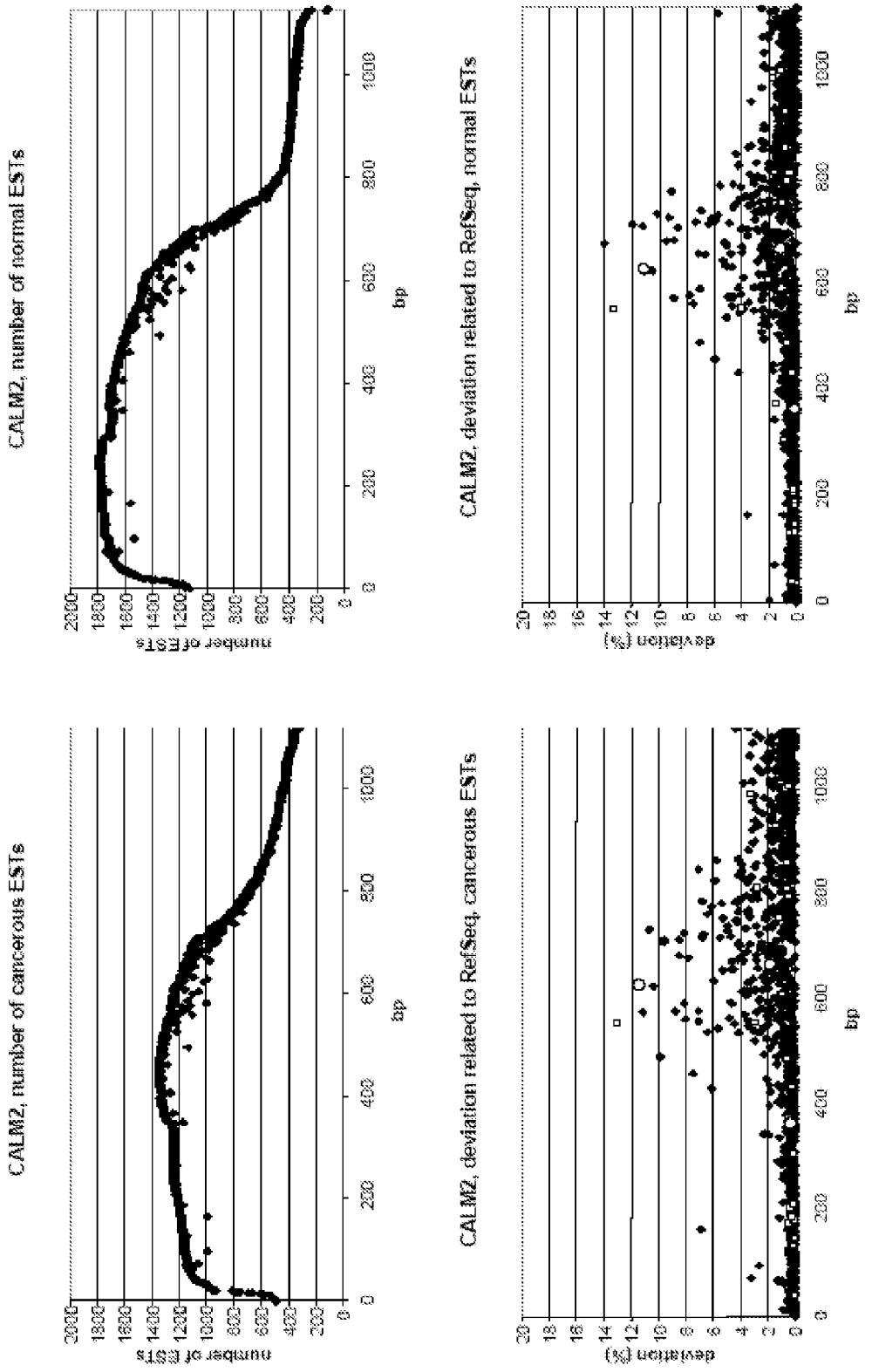


Figure 4a (following)

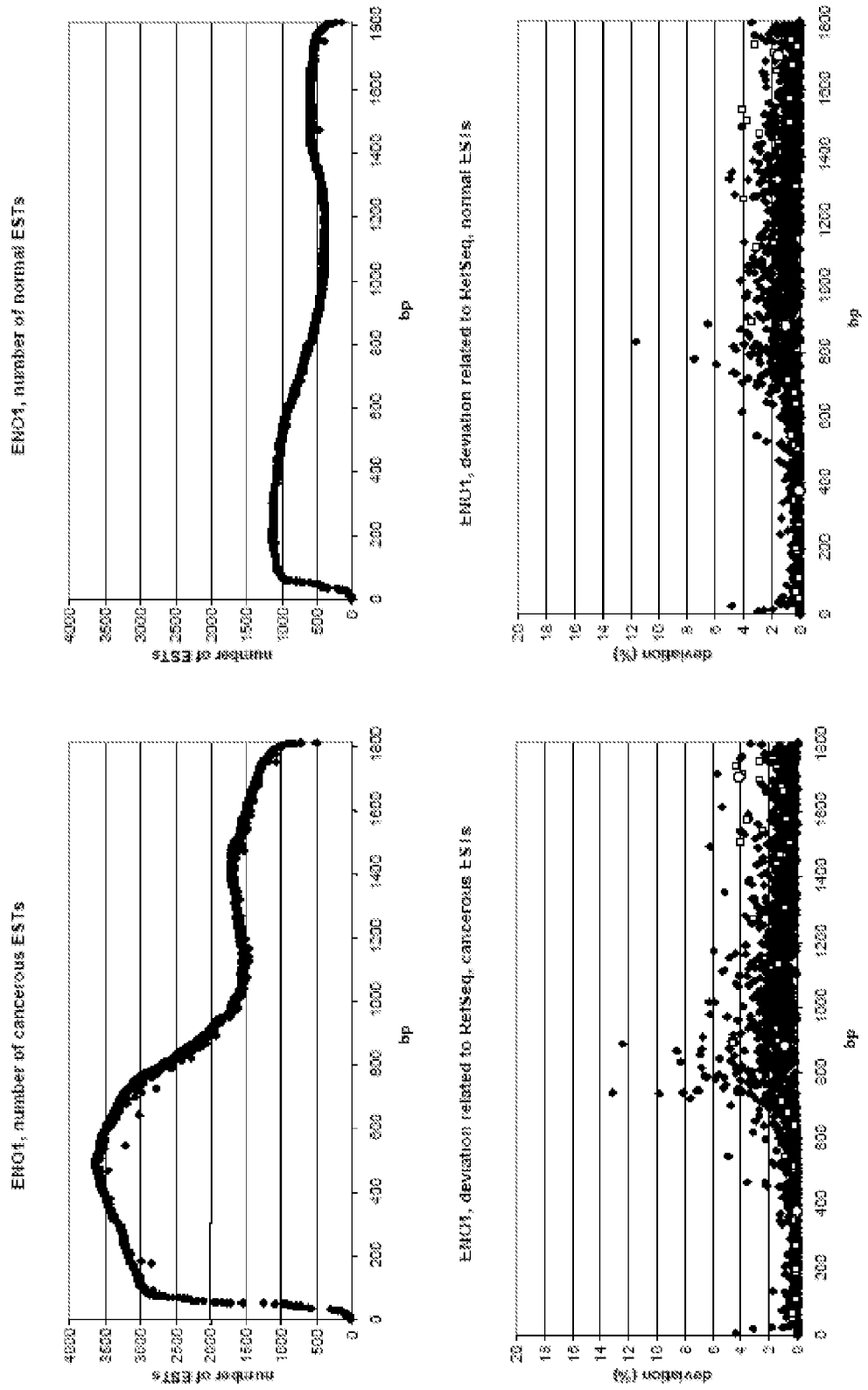

EHO1, number of normal ESTs
number of ESTs
bp
EHO1, deviation related to KetSeq, normal ESTs
deviation (%)
bp
EHO1, number of cancerous ESTs
number of ESTs
bp
EHO1, deviation related to KetSeq, cancerous ESTs
deviation (%)
bp

Figure 4b

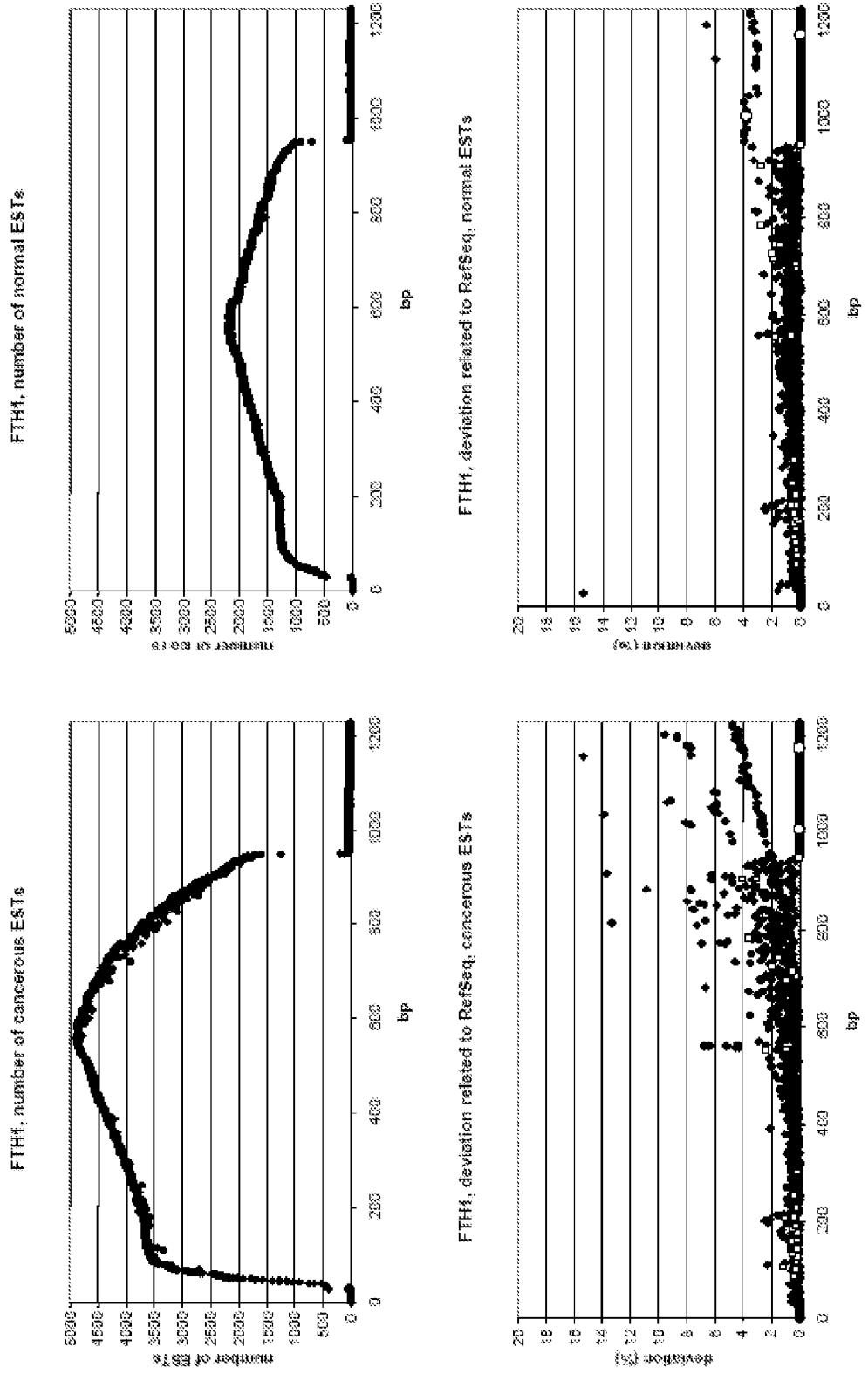


Figure 4b (following)

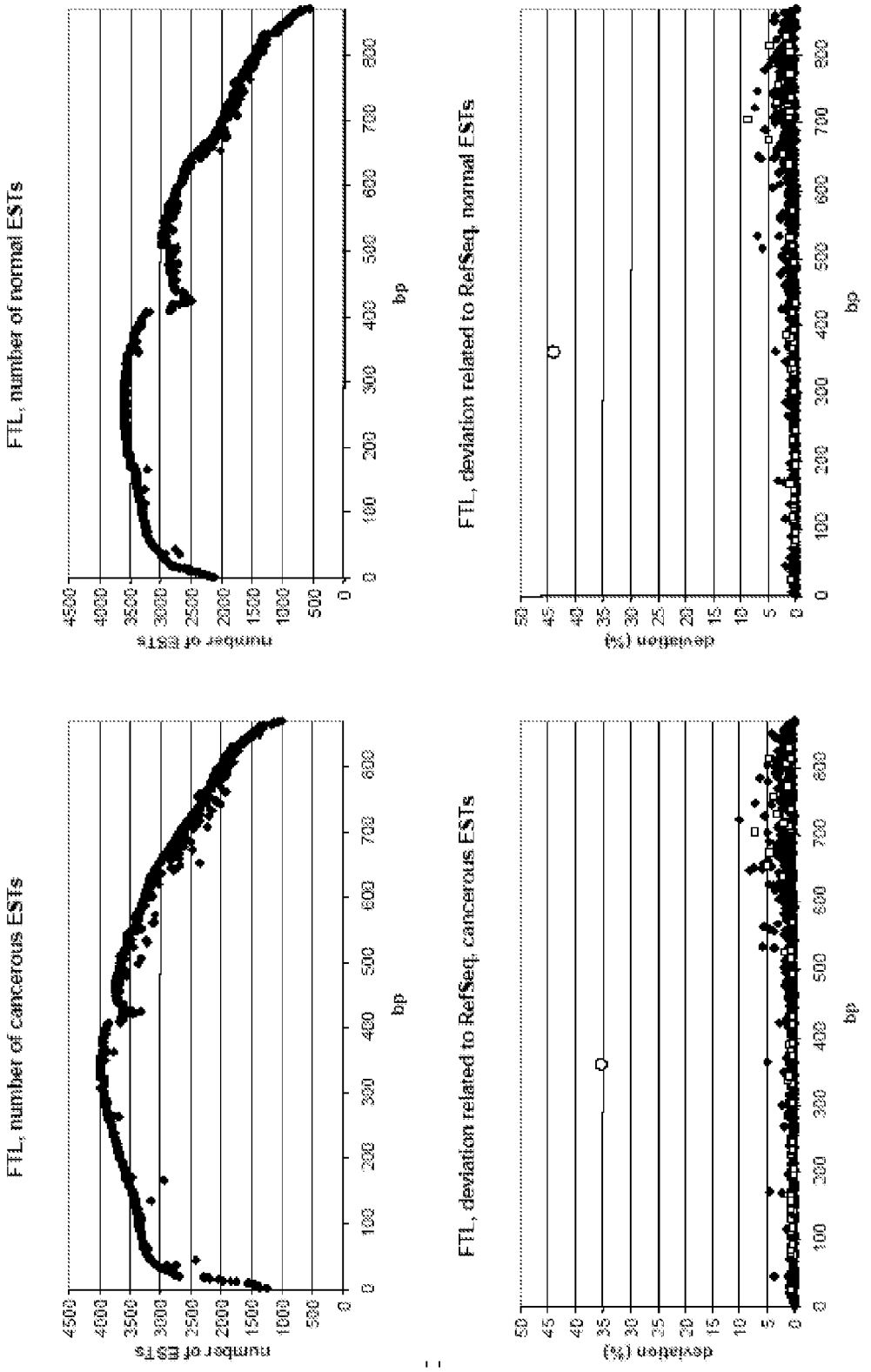


Figure 4b (following)

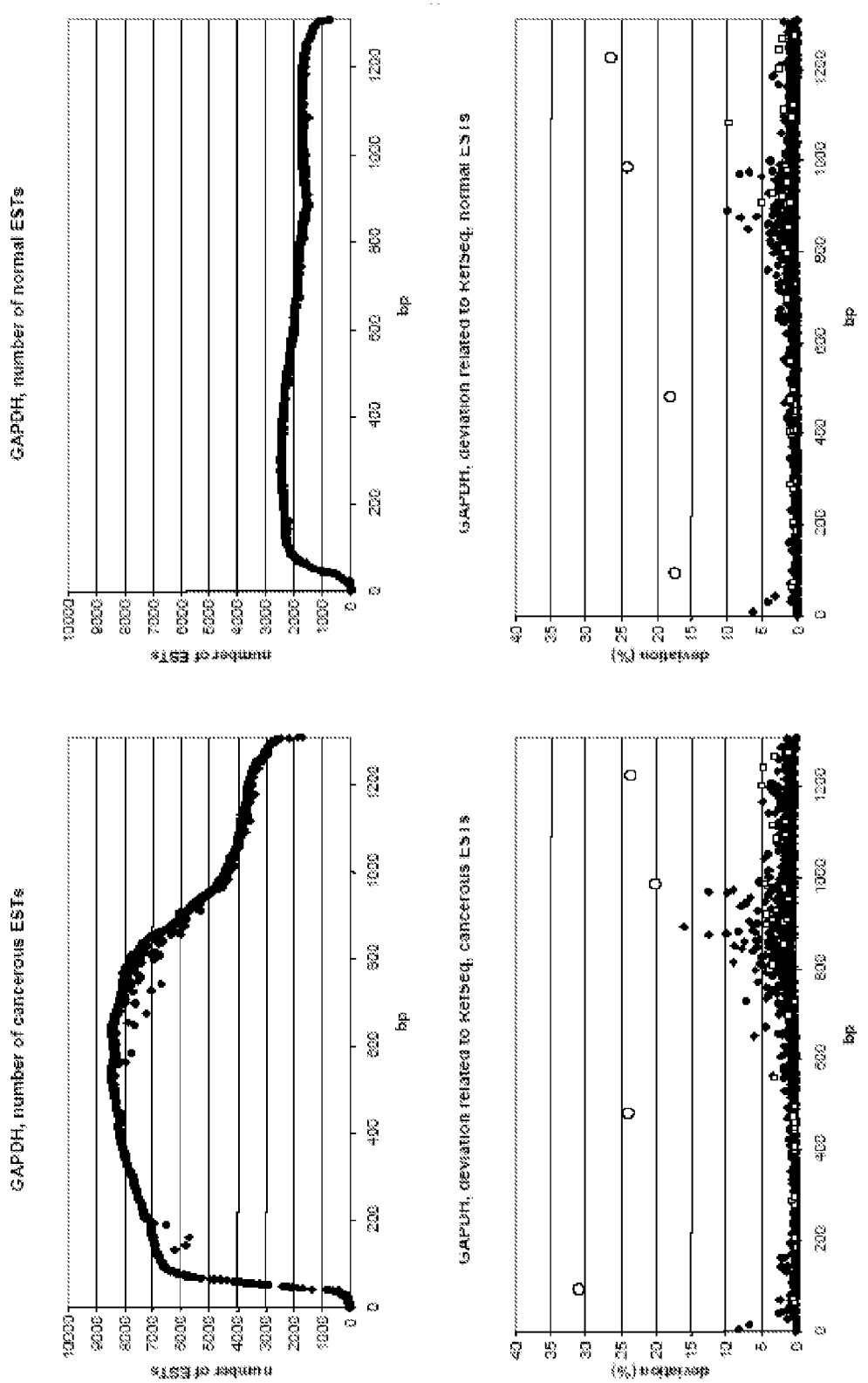


Figure 4b (following)

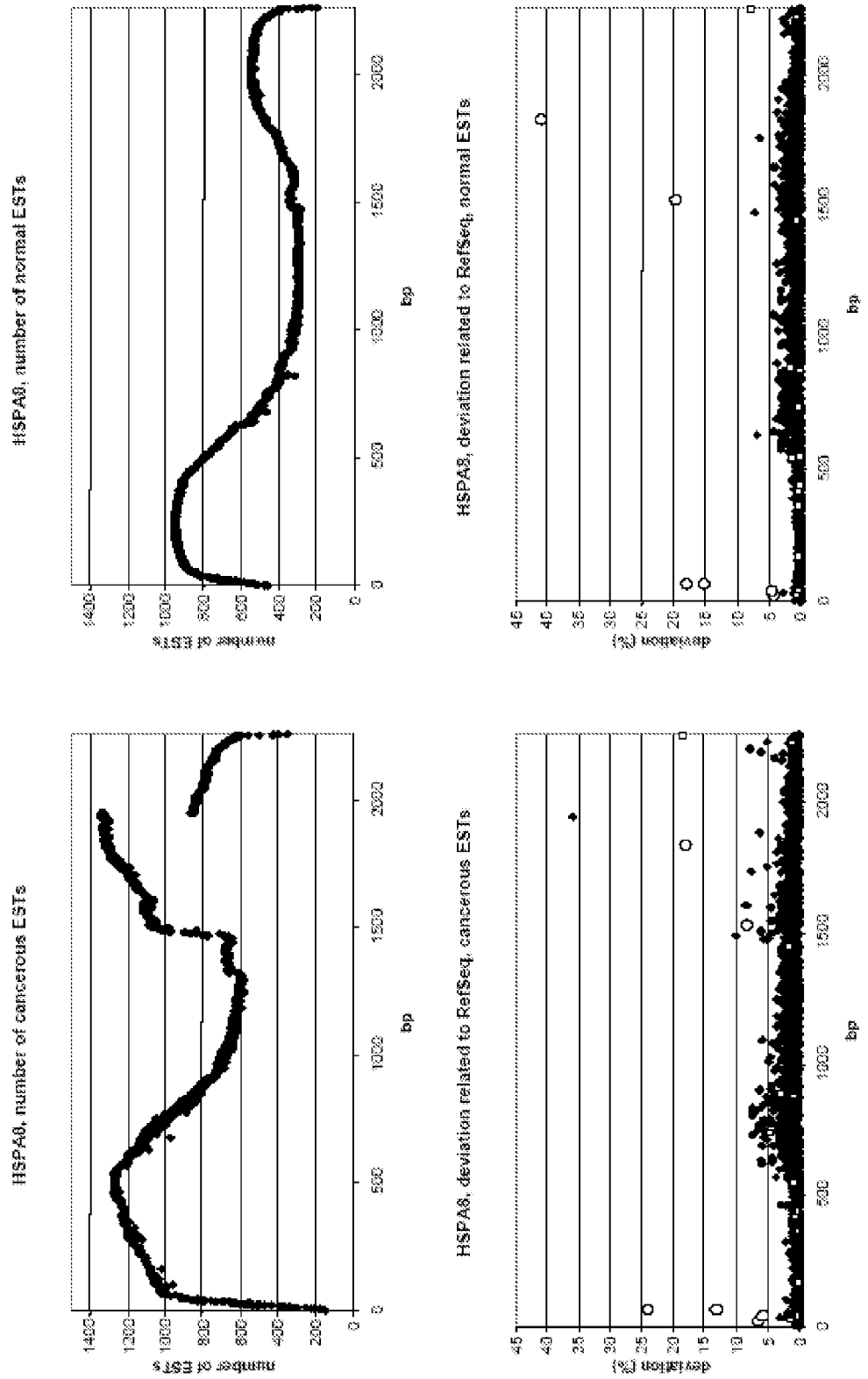

HSPA8, number of normal ESTs
number of ESTs
bp
HSPA8, deviation related to RefSeq, normal ESTs
deviation (%)
bp
HSPA8, number of cancerous ESTs
number of ESTs
bp
HSPA8, deviation related to RefSeq, cancerous ESTs
deviation (%)
bp

Figure 4c

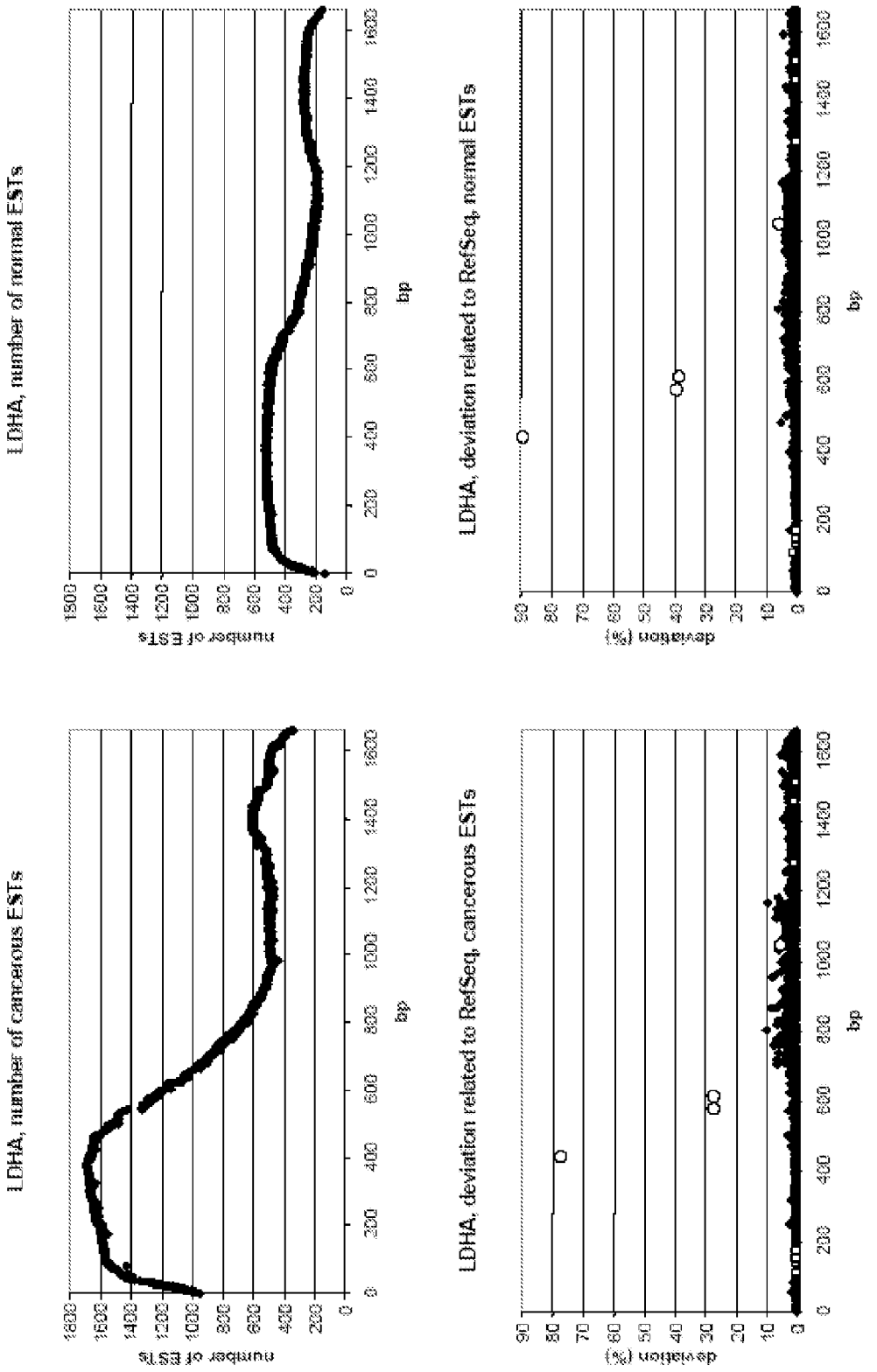


Figure 4c (following)

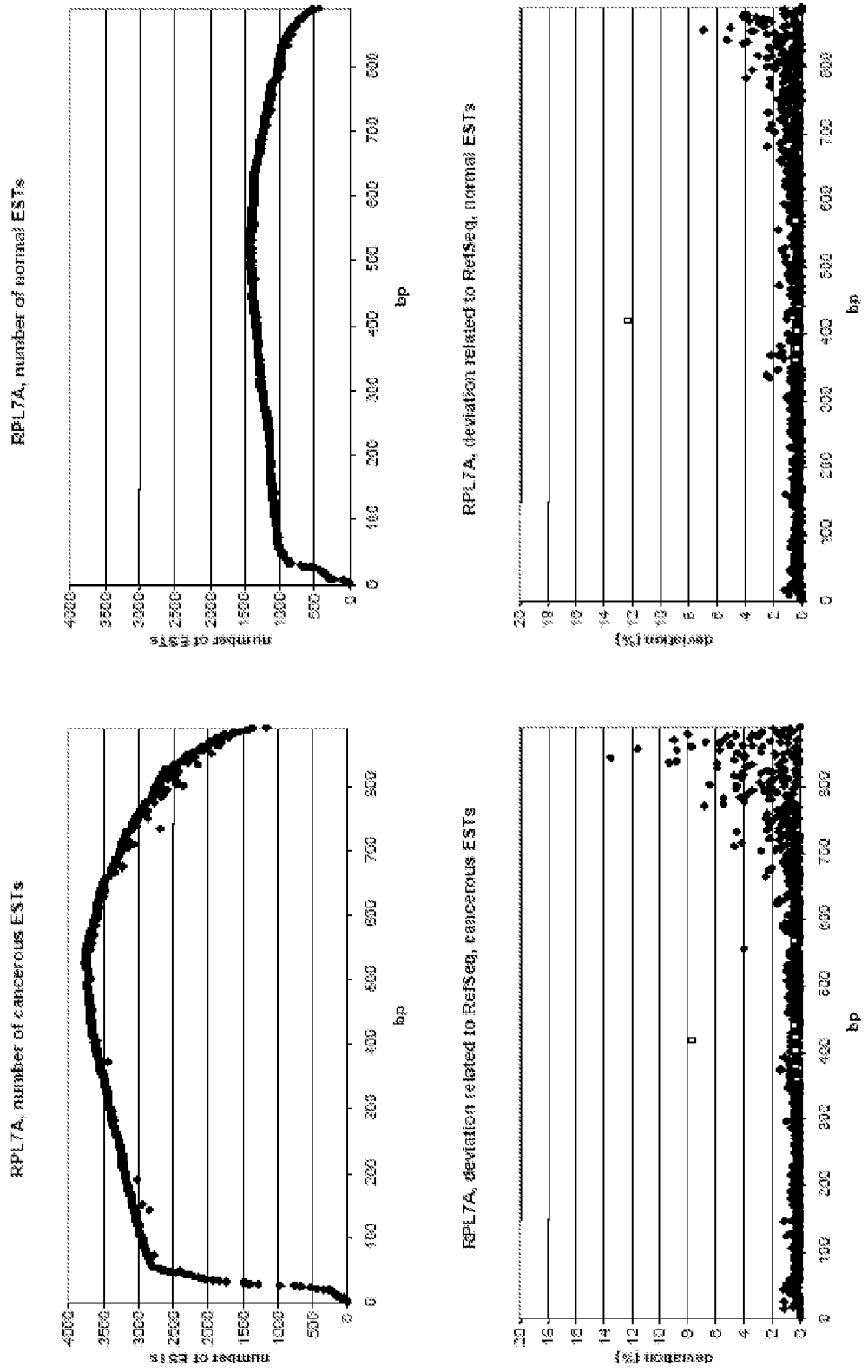


Figure 4c (following)

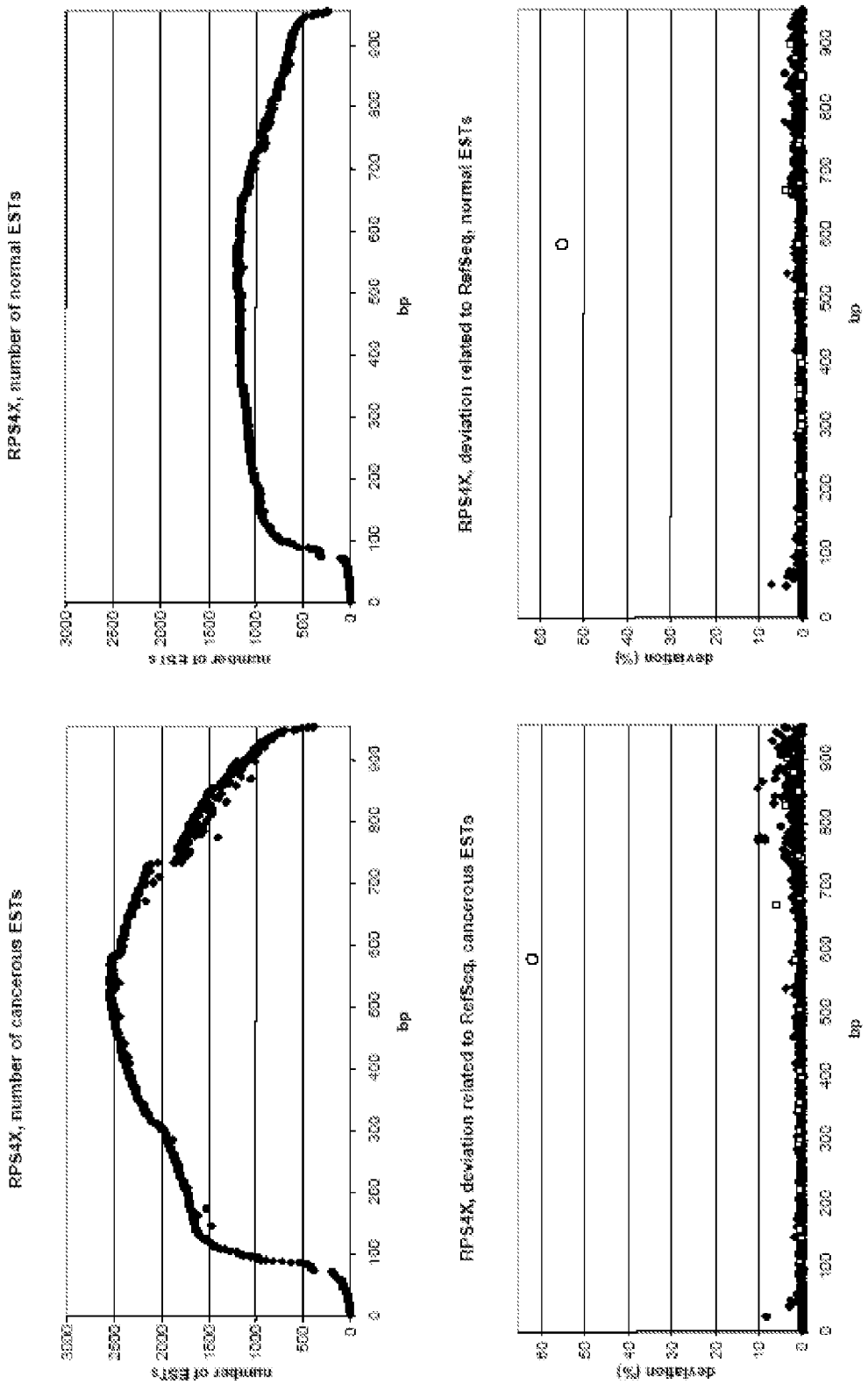


Figure 4c (following)

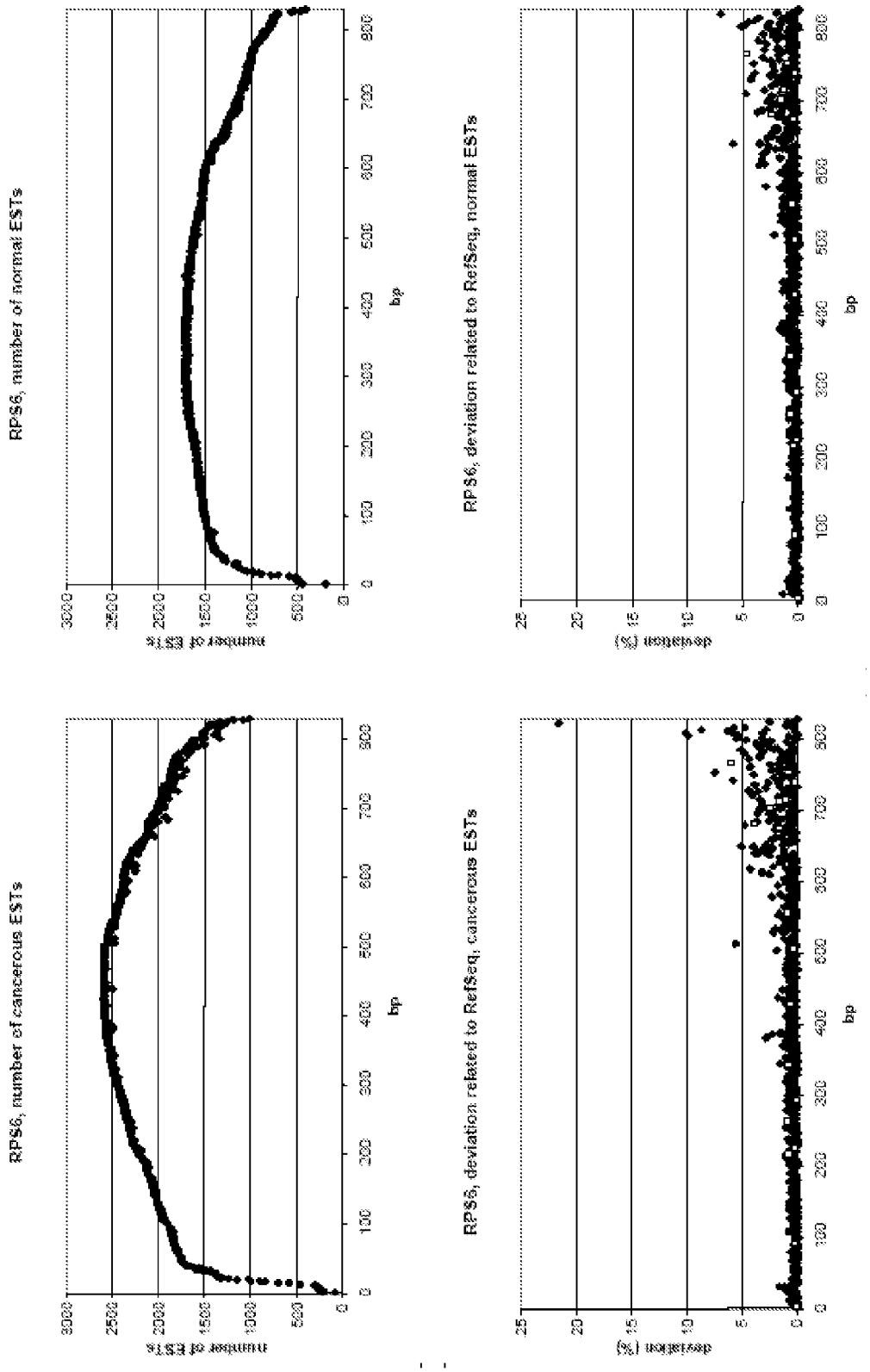

RPS6, number of normal ESTs
number of ESTs
bp
RPS6, deviation related to RefSeq, normal ESTs
deviation (%)
bp
RPS6, number of cancerous ESTs
number of ESTs
bp
RPS6, deviation related to RefSeq, cancerous ESTs
deviation (%)
bp

Figure 4d

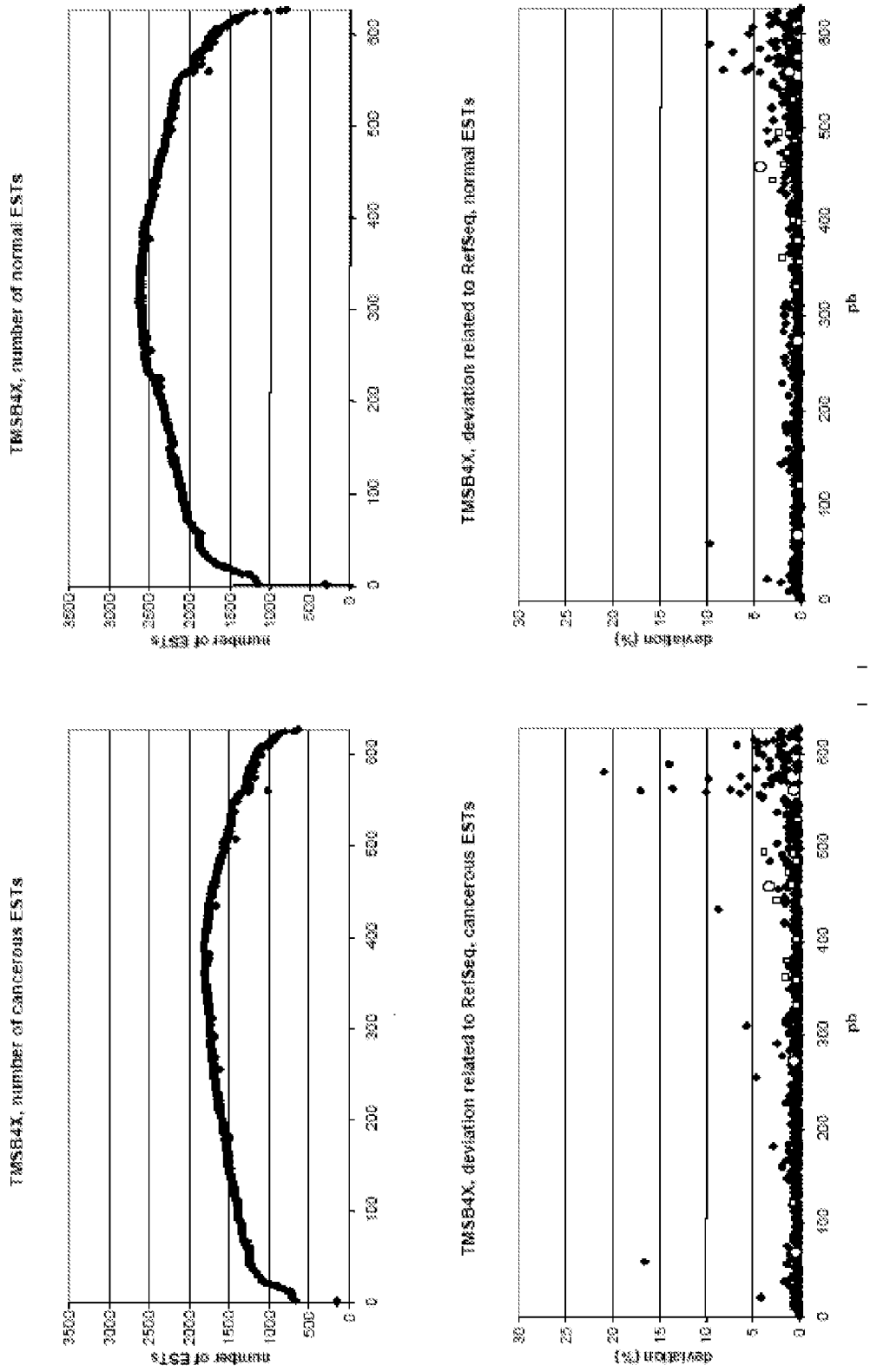


Figure 4d (following)

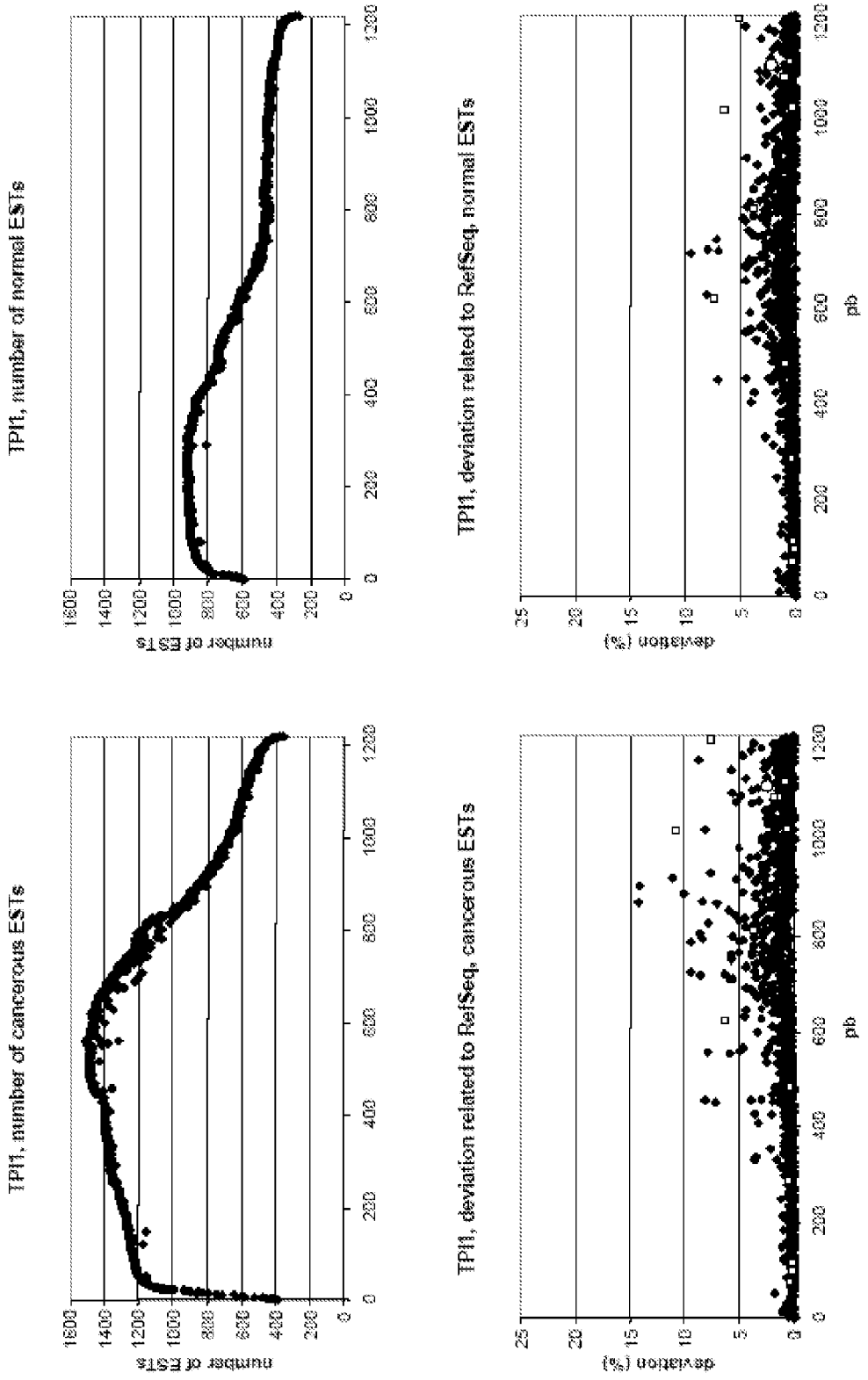


Figure 4d (following)

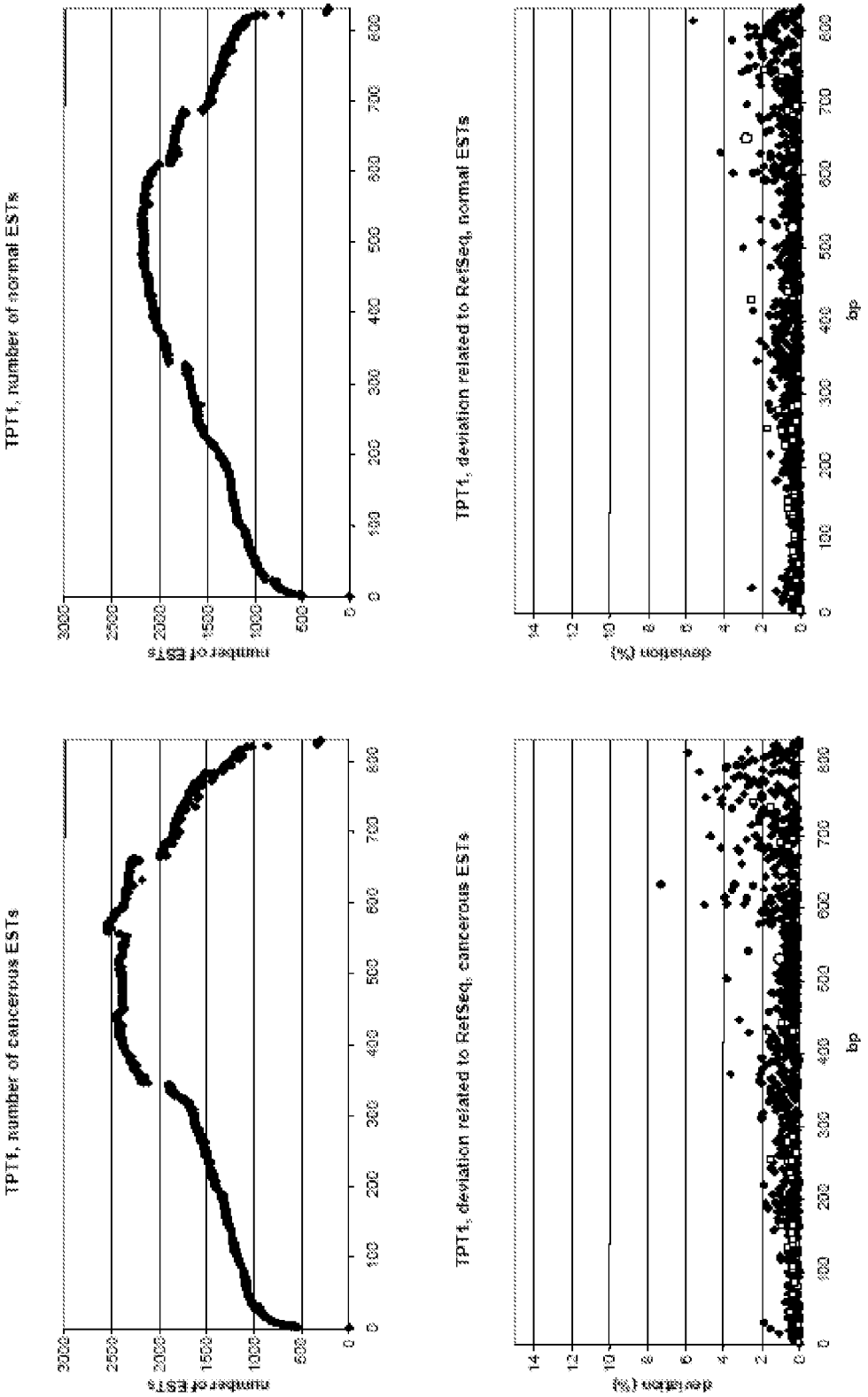


Figure 4d (following)

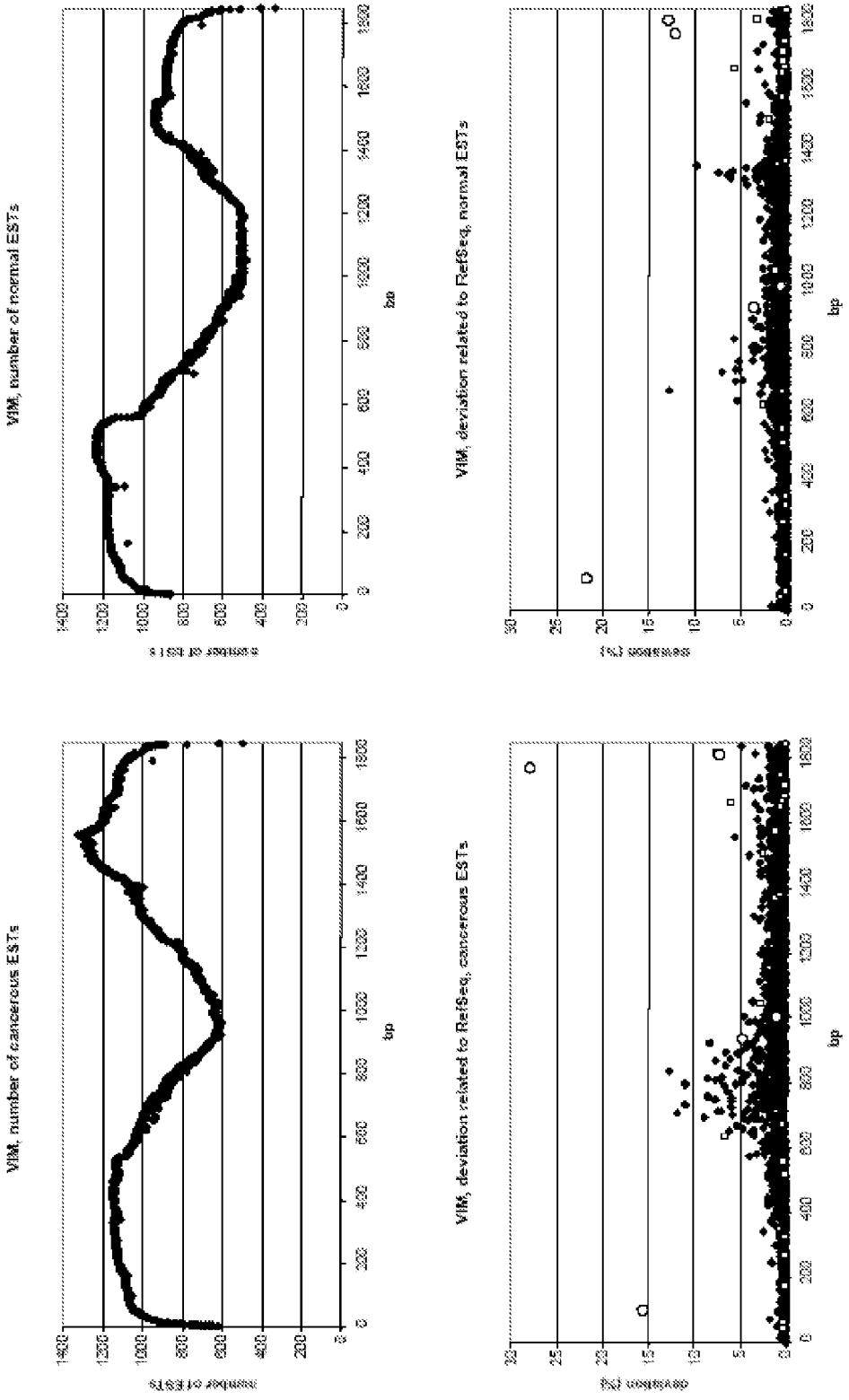


Figure 4d (following)

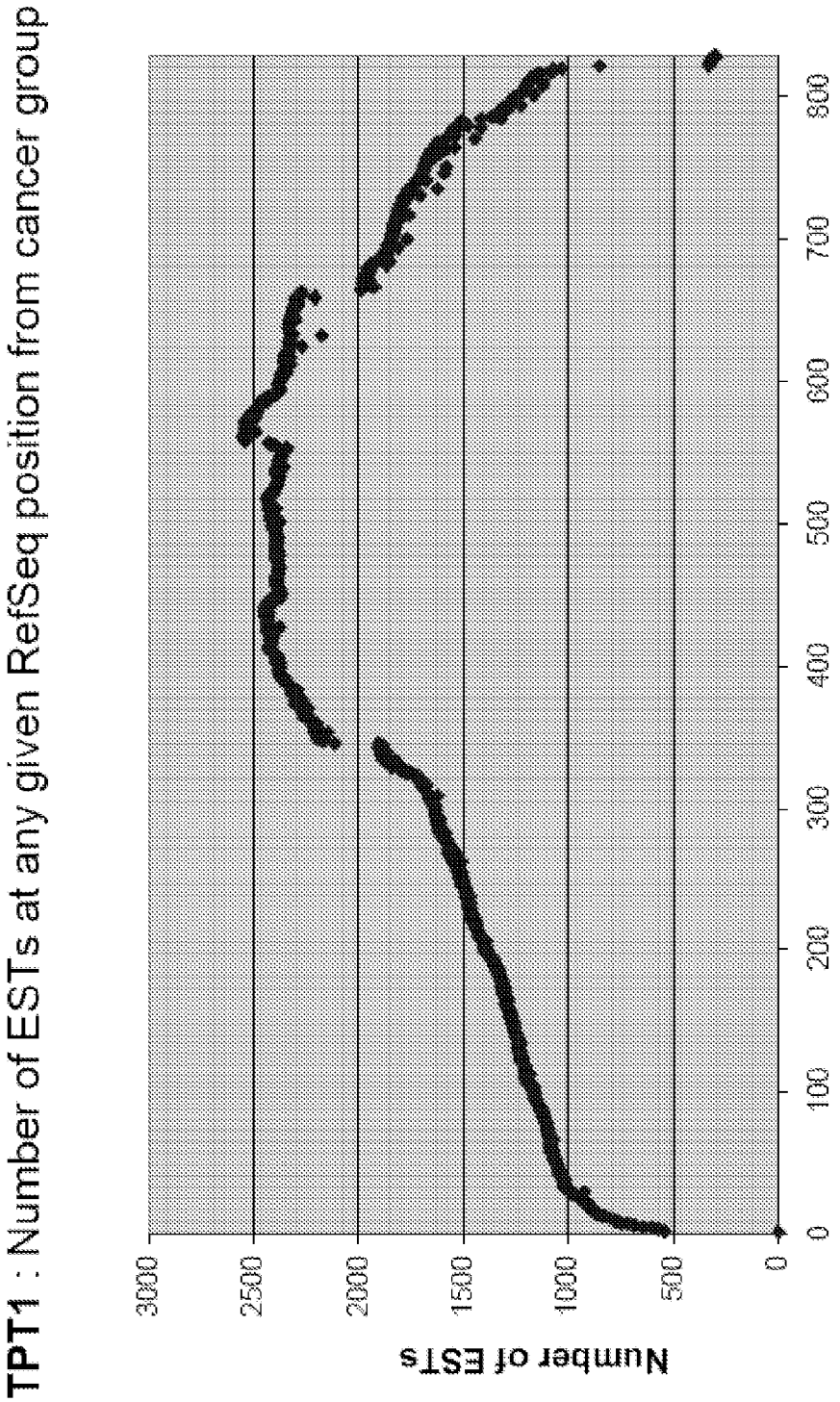

TPT1 : Number of ESTs at any given RefSeq position from cancer group
Number of ESTs
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
bp

Figure 5a

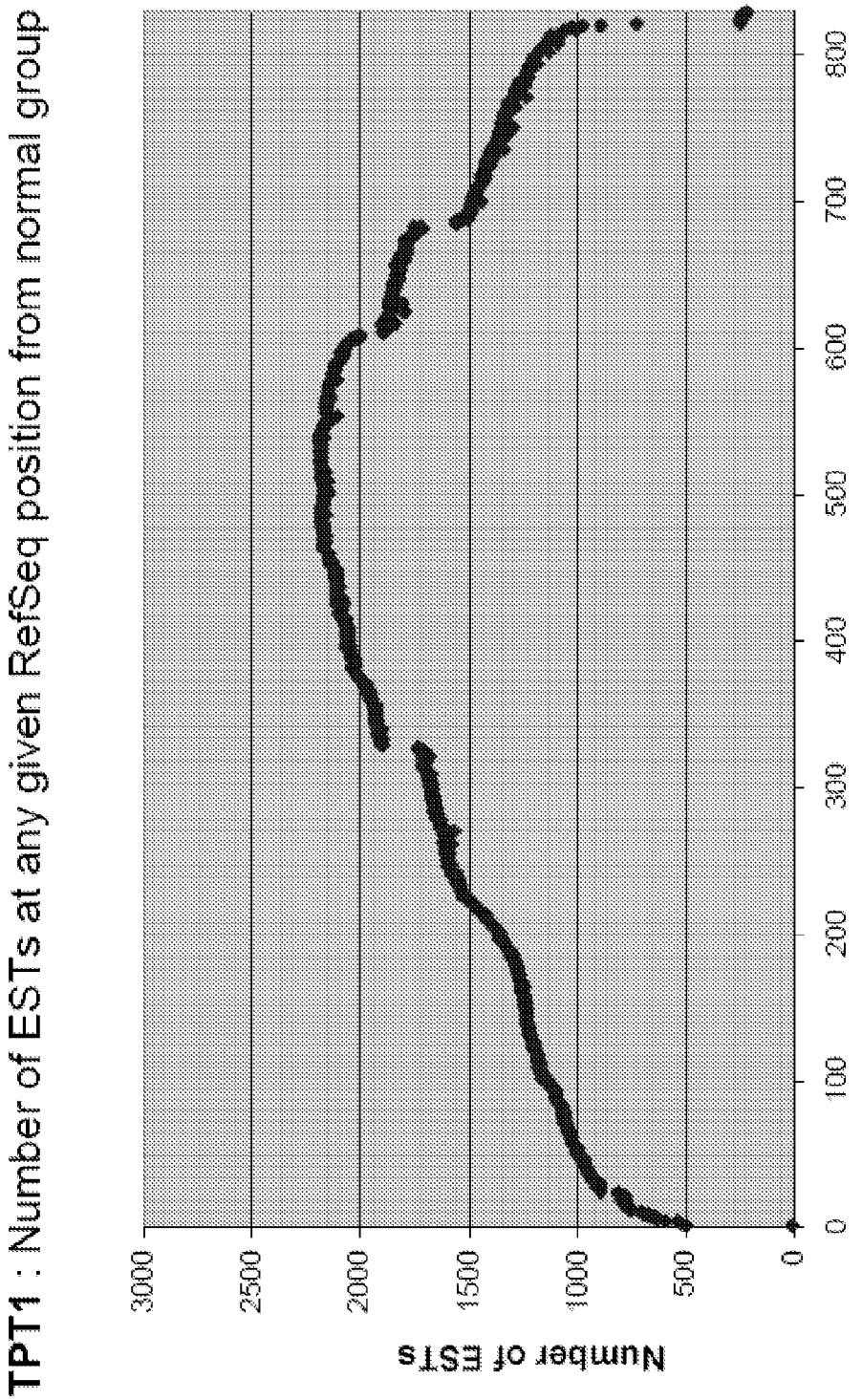

TPT1 : Number of ESTs at any given RefSeq position from normal group
Number of ESTs
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
bp

Figure 5b

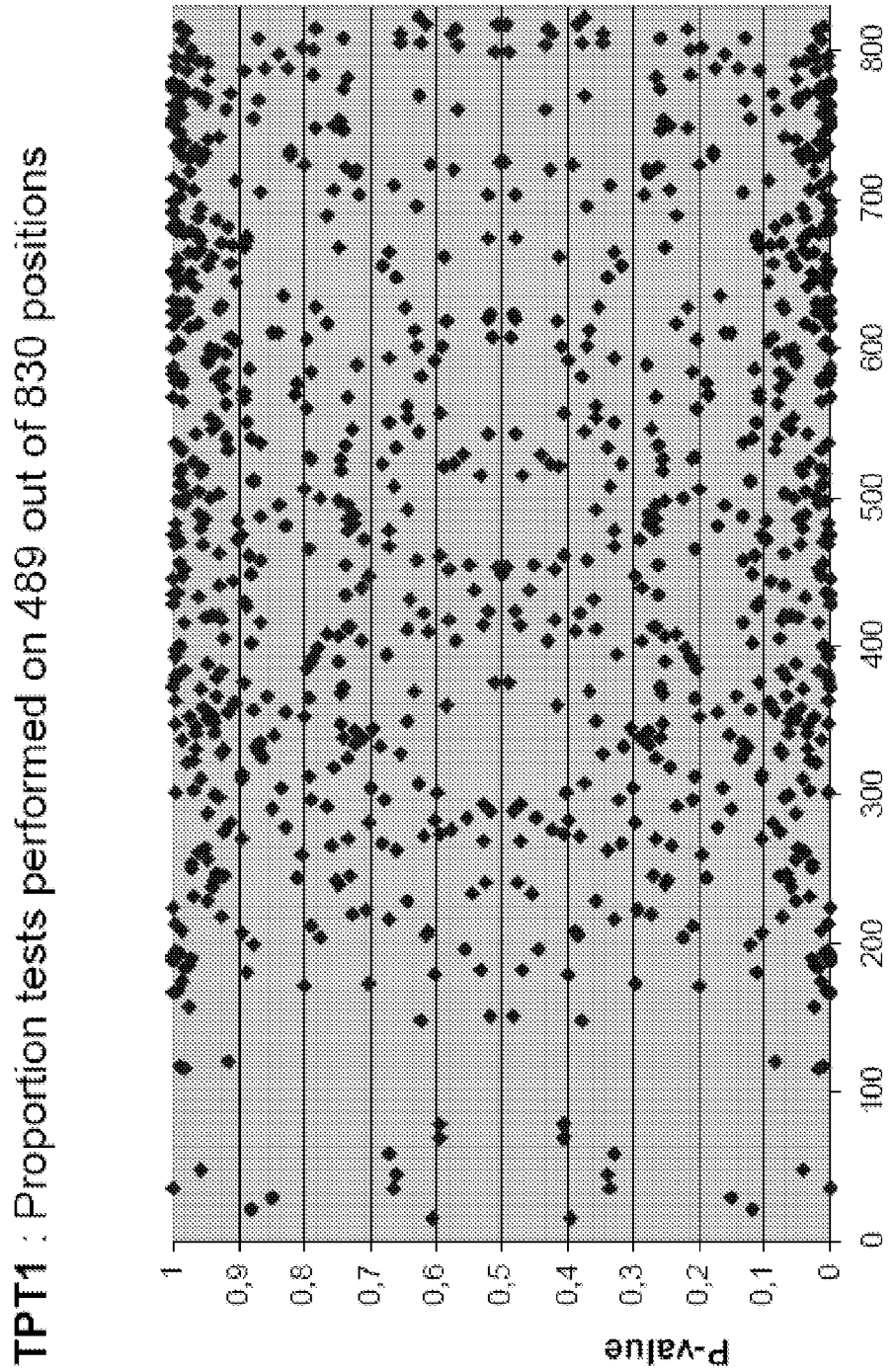

TPT1 : Proportion tests performed on 489 out of 830 positions
P-value
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
100
200
300
400
500
600
700
800
bp

Figure 5c

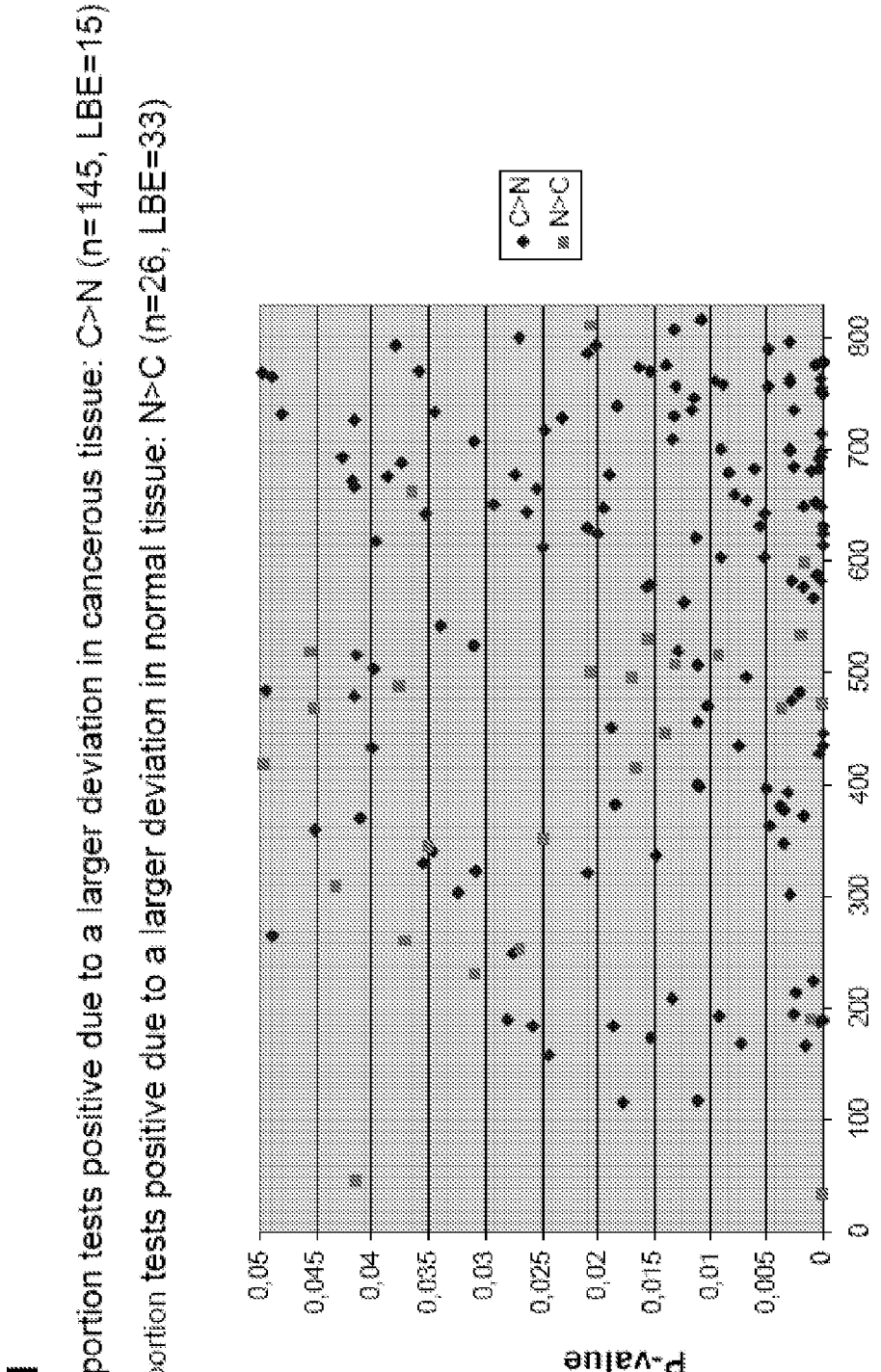

TPT1
✓ Proportion tests positive due to a larger deviation in cancerous tissue: C>N (n=145, LBE=15)
✓ Proportion tests positive due to a larger deviation in normal tissue: N>C (n=26, LBE=33)
C>N
N>C
P-value
0,05
0,045
0,04
0,035
0,03
0,025
0,02
0,015
0,01
0,005
0
0
100
200
300
400
500
600
700
800
bp

Figure 5d

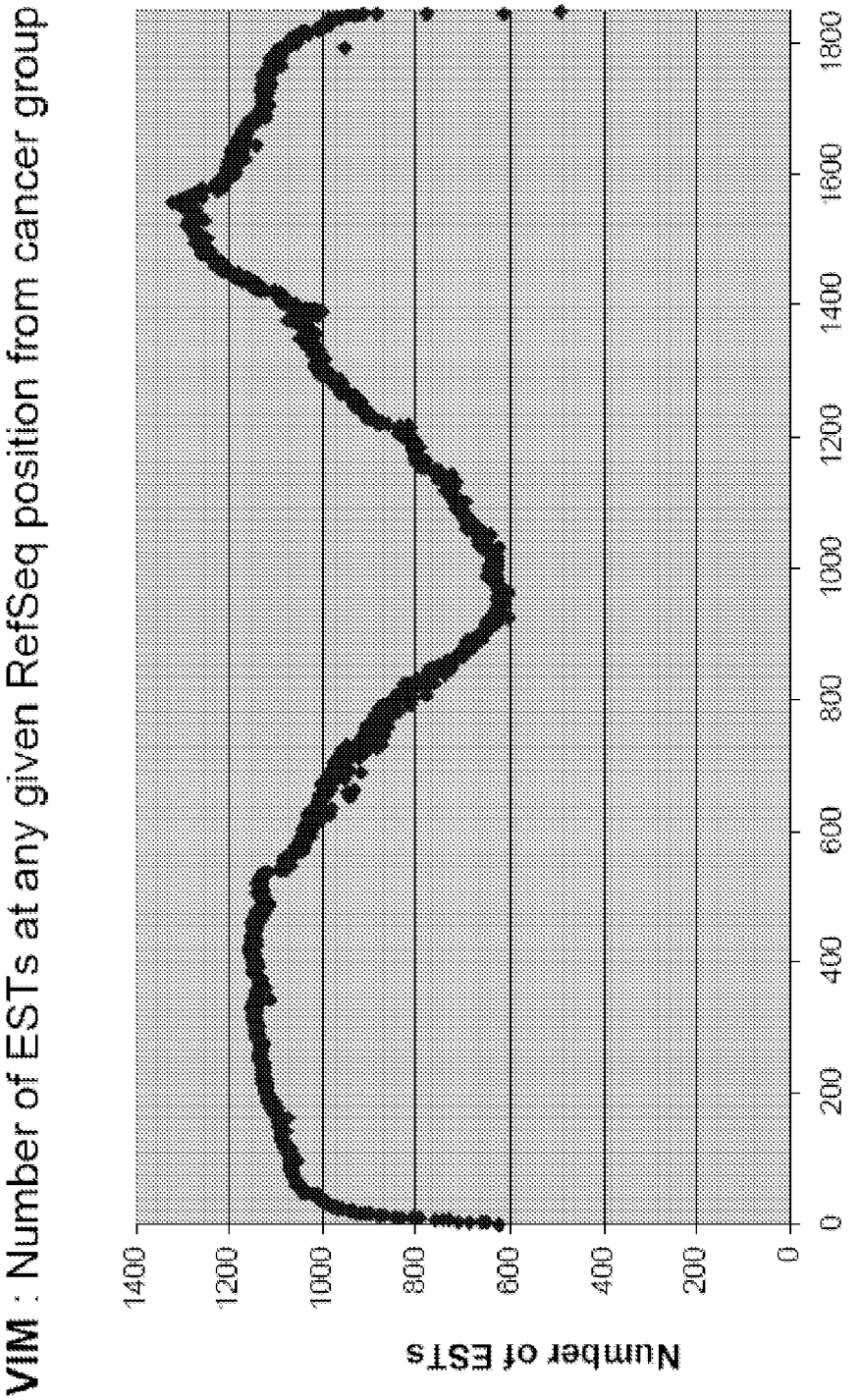

VIM : Number of ESTs at any given RefSeq position from cancer group
Number of ESTs
0
200
400
600
800
1000
1200
1400
bp
0
200
400
600
800
1000
1200
1400
1600
1800

Figure 5e

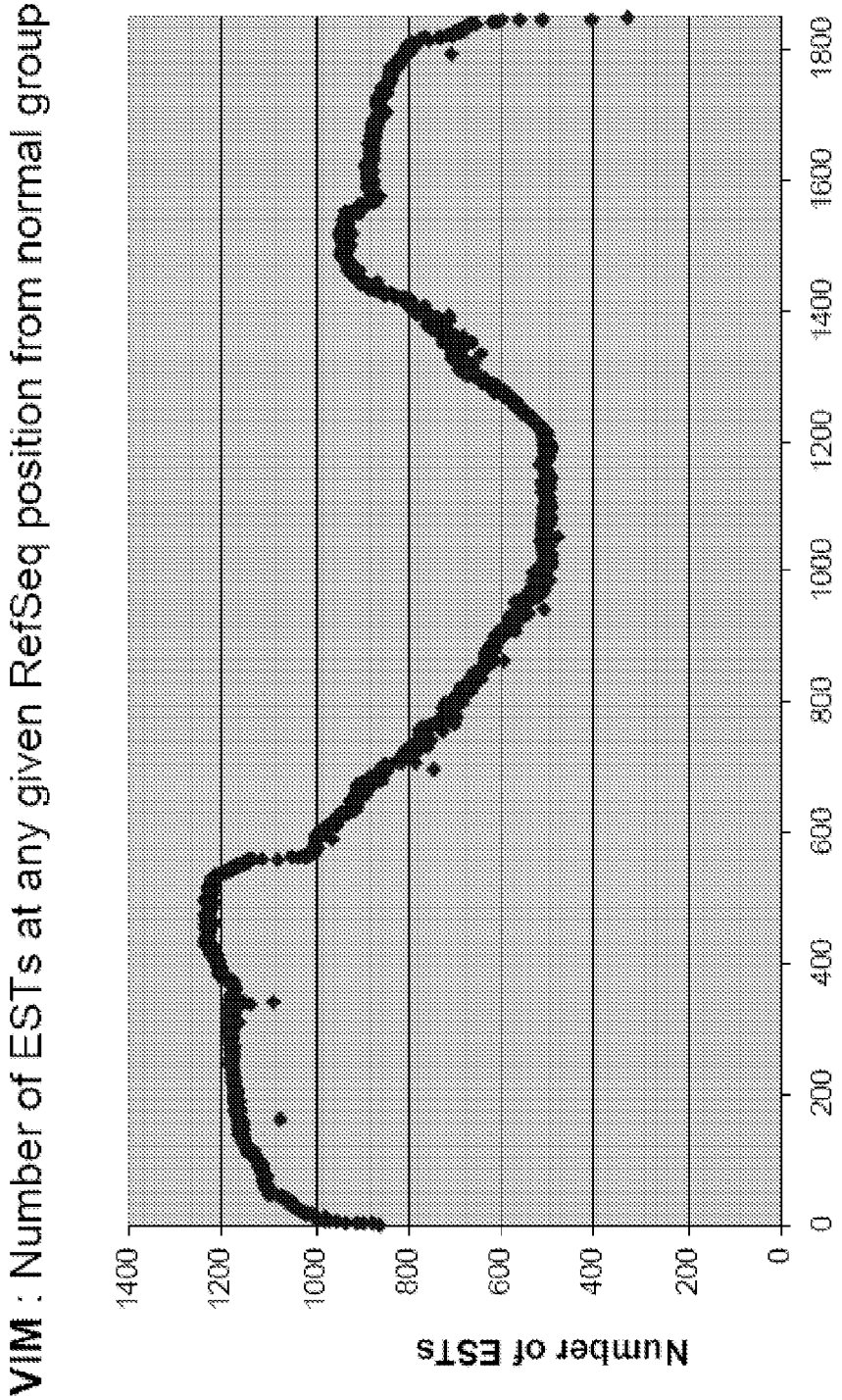

VIM : Number of ESTs at any given RefSeq position from normal group
Number of ESTs
0
200
400
600
800
1000
1200
1400
bp
0
200
400
600
800
1000
1200
1400
1600
1800

Figure 5f

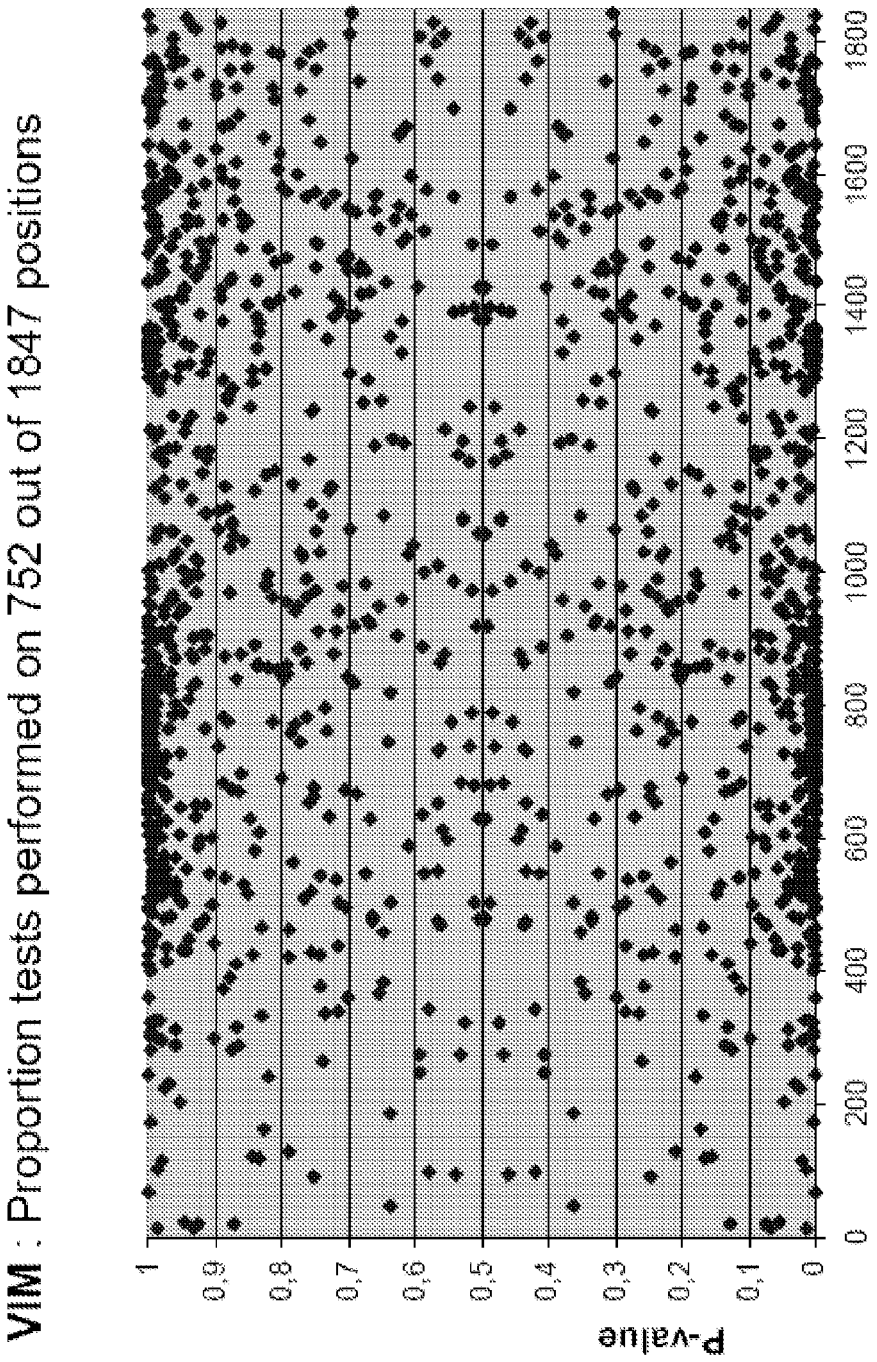

VIM : Proportion tests performed on 752 out of 1847 positions
P-value
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
200
400
600
800
1000
1200
1400
1600
1800
bp

Figure 5g

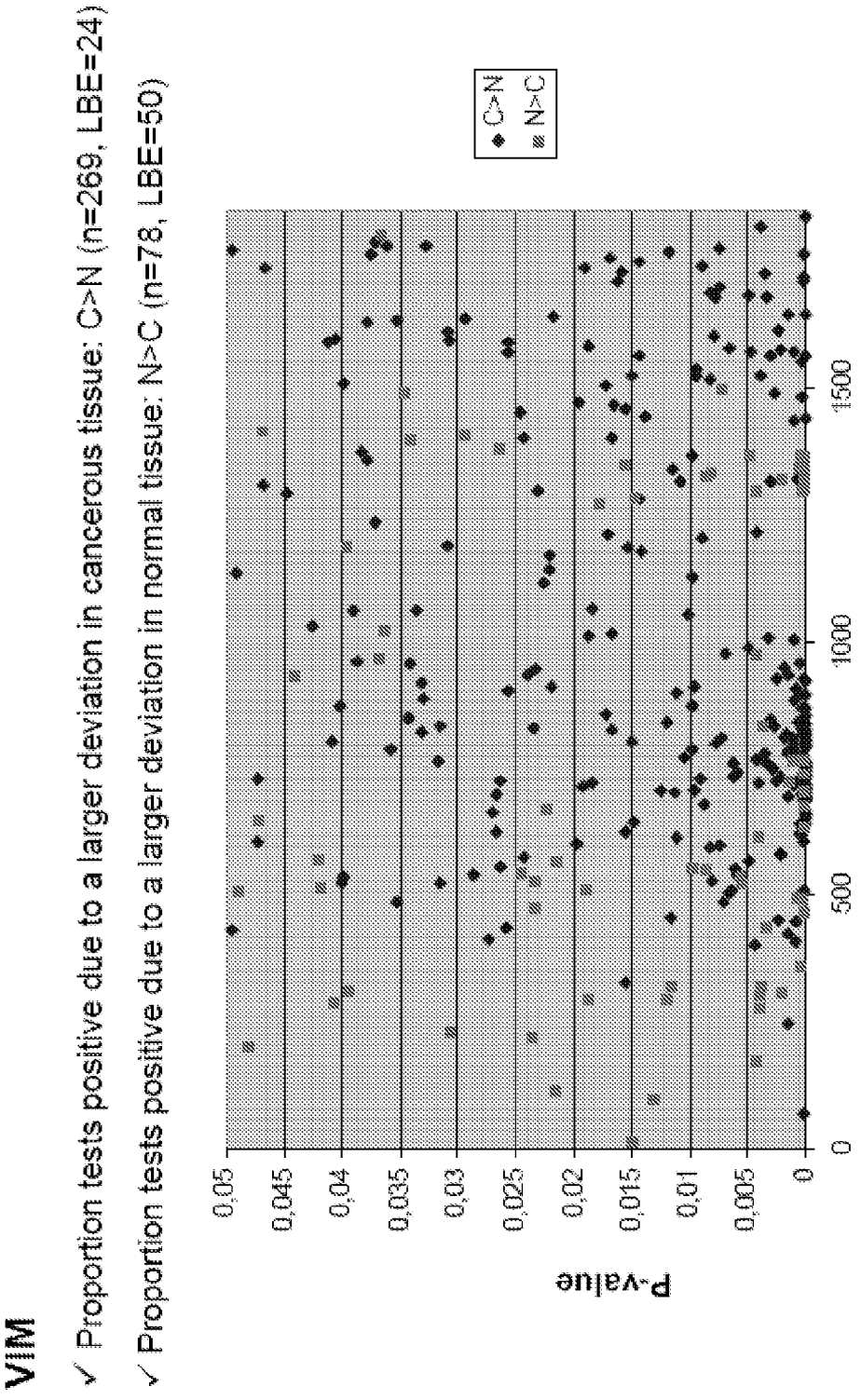

VIM
✓ Proportion tests positive due to a larger deviation in cancerous tissue: C>N (n=269, LBE=24)
✓ Proportion tests positive due to a larger deviation in normal tissue: N>C (n=78, LBE=50)
C>N
N>C
P-value
0,05
0,045
0,04
0,035
0,03
0,025
0,02
0,015
0,01
0,005
0
0
500
1000
1500
bp

Figure 5h

Proportion test analysis : results

Gene	Number of proportion tests	% coverage	C>N	LBE	N>C	LBE
ALB	194	9	**38**	10	**56**	8
ALDOA transcript1	336	30	**81**	11	**35**	21
ATP5A1 transcript1	238	38	**123**	3	**6**	20
CALM2	374	23	**69**	16	**40**	21
ENO1	614	27	**186**	18	**31**	42
FTH1	592	71	**226**	20	**57**	38
FTL	678	56	**118**	31	**86**	36
GAPDH	812	91	**311**	23	**61**	57
HSPA8 transcript1	513	59	**163**	13	**18**	37
LDHA	181	9	**70**	4	**10**	13
RPL7A	337	35	**103**	11	**33**	22
RPS4X	371	45	**124**	11	**27**	25
RPS6	432	19	**86**	17	**40**	25
TMSB4X	352	19	**70**	18	**74**	17
TPI1	379	31	**99**	14	**47**	23
TPT1	489	26	**145**	15	**26**	33
VIM	752	57	**269**	24	**78**	50

Figure 5i

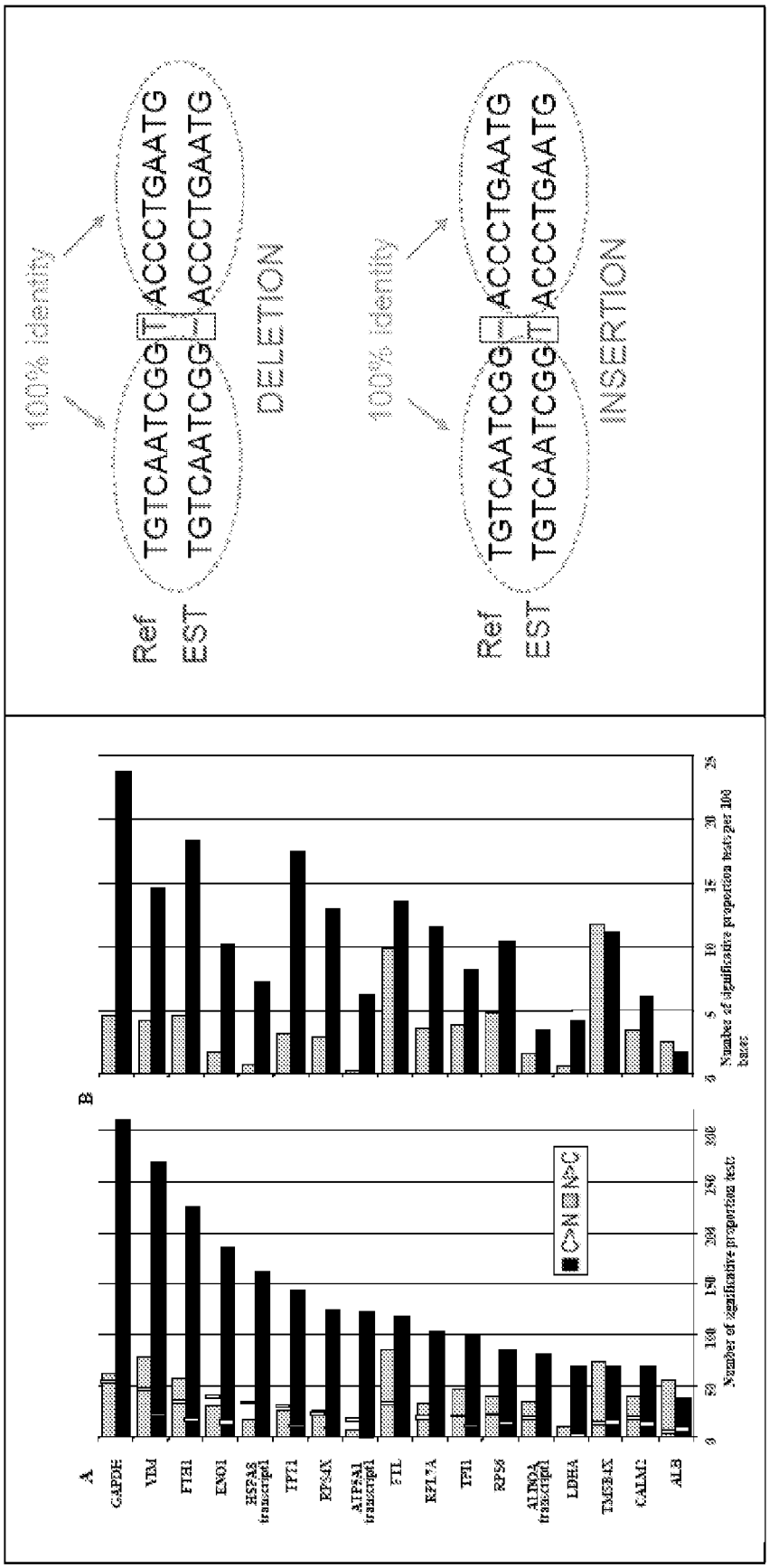

100% identity
Ref TGTCAATCGG T ACCCTGAATG
EST TGTCAATCGG - ACCCTGAATG
DELETION
100% identity
Ref TGTCAATCGG - ACCCTGAATG
EST TGTCAATCGG T ACCCTGAATG
INSERTION
A
B
GAPDH
VIM
FTH1
ENO1
HSPA8 transcript1
TPT1
RPS4X
ATP5A1 transcript1
FTL
KPYA
TPI1
RPS9
ATP5O transcript1
LDHA
TMSB4X
CALM2
ALB
C>N N>C
Number of significative proportion test
Number of significative proportion test per 100 tests

Figure 5j

Proportion test analysis : results *(deletions and insertions)*

DELETIONS						
Genes	Number of proportion tests	C>N	LBE	N>C	LBE	C>N / N>C
ALB	17	1	1	13	0	0,1
ALDOA	34	8	1	7	1	1,1
ATP5A1	24	16	0	0	2	
CALM2	50	14	2	4	2	3,5
ENO1	41	29	0	5	3	5,8
FTH1	67	46	0	4	5	11,5
FTL	100	42	2	9	7	4,7
GAPDH	99	59	2	17	7	3,5
HSPA8	30	27	0	0	2	
LDHA	14	6	0	3	0	2,0
RPL7A	45	26	0	1	3	26,0
RPS4X	35	29	0	1	3	29,0
RPS6	45	24	0	1	3	24,0
TMSB4X	28	16	0	4	1	4,0
TPI1	30	17	0	1	2	17,0
TPT1	43	27	0	4	3	6,8
VIM	26	11	1	9	1	1,2
Total	728	396	9	83	45	4,8

INSERTIONS						
Genes	Number of proportion tests	C>N	LBE	N>C	LBE	C>N / N>C
ALB	9	0	0	9	0	0,0
ALDOA	5	1	0	1	0	1,0
ATP5A1	9	6	0	0	0	
CALM2	72	10	3	10	3	1,0
ENO1	40	14	0	2	3	7,0
FTH1	114	54	1	3	9	18,0
FTL	188	26	9	33	9	0,8
GAPDH	137	46	4	18	9	2,6
HSPA8	14	7	0	2	1	3,5
LDHA	2	0	0	0	0	
RPL7A	56	12	1	1	3	12,0
RPS4X	29	9	1	2	1	4,5
RPS6	44	12	1	1	3	12,0
TMSB4X	22	5	1	6	0	0,8
TPI1	26	6	1	5	1	1,2
TPT1	41	7	1	4	2	1,8
VIM	27	10	1	3	1	3,3
Total	835	225	24	100	45	2,3

Figure 5k

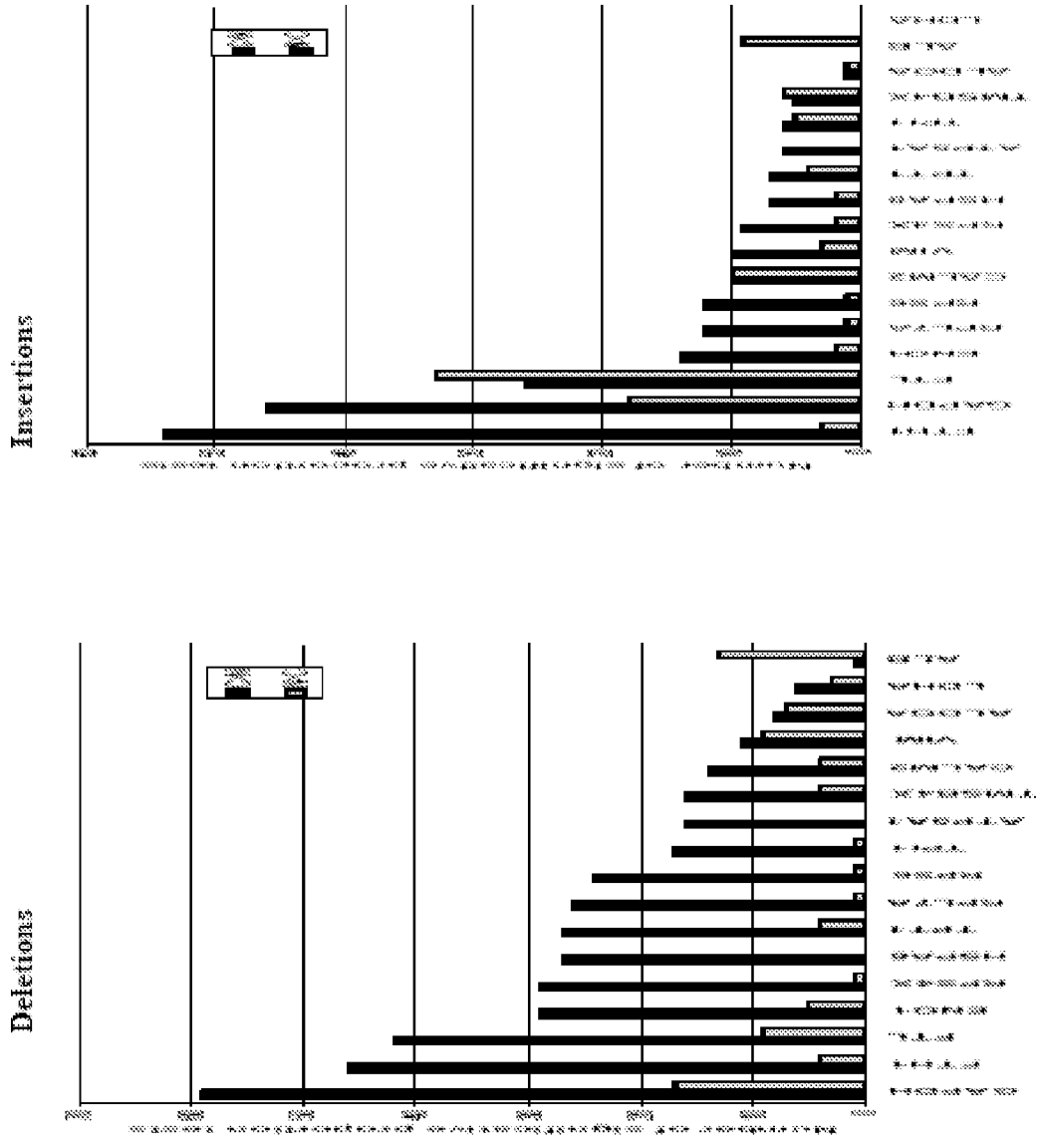


Figure 51

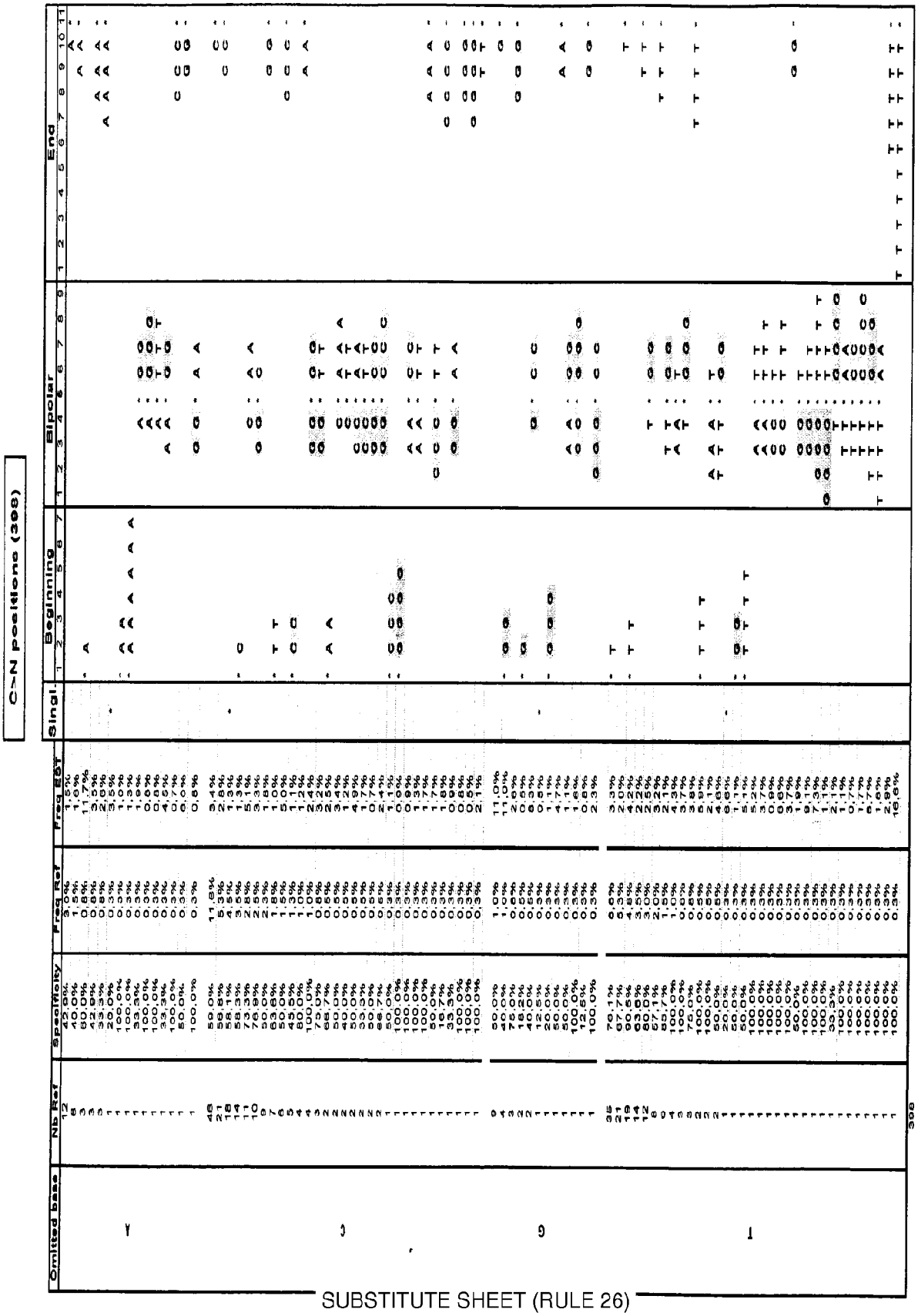

C->N positions (398)
Omitted base
Nb Ref
Specificity
Freq Ref
Freq Est
Singl.
Beginning
Bipolar
End
A
C
G
T
398

Figure 5m

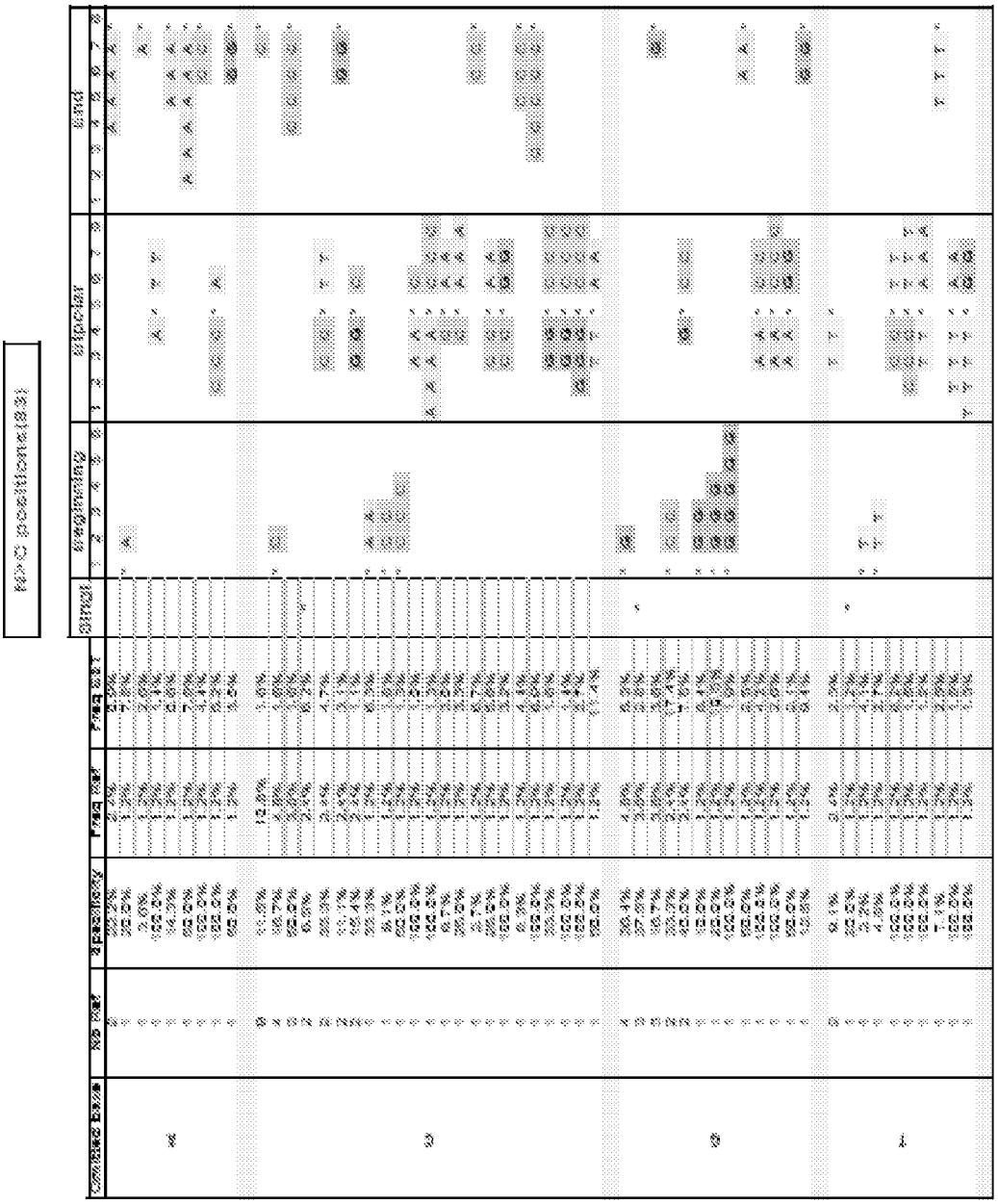


Figure 5n

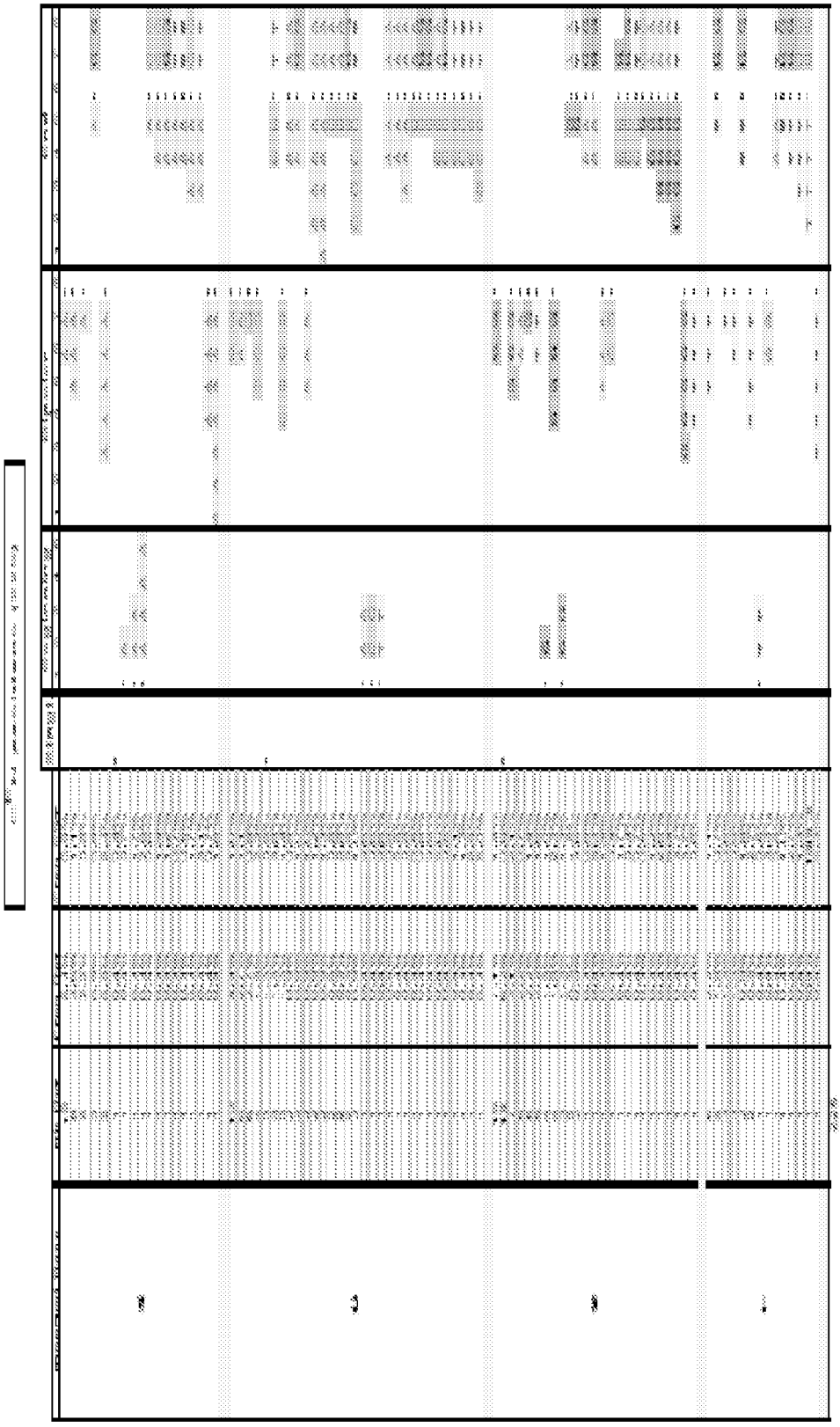


Figure 5o

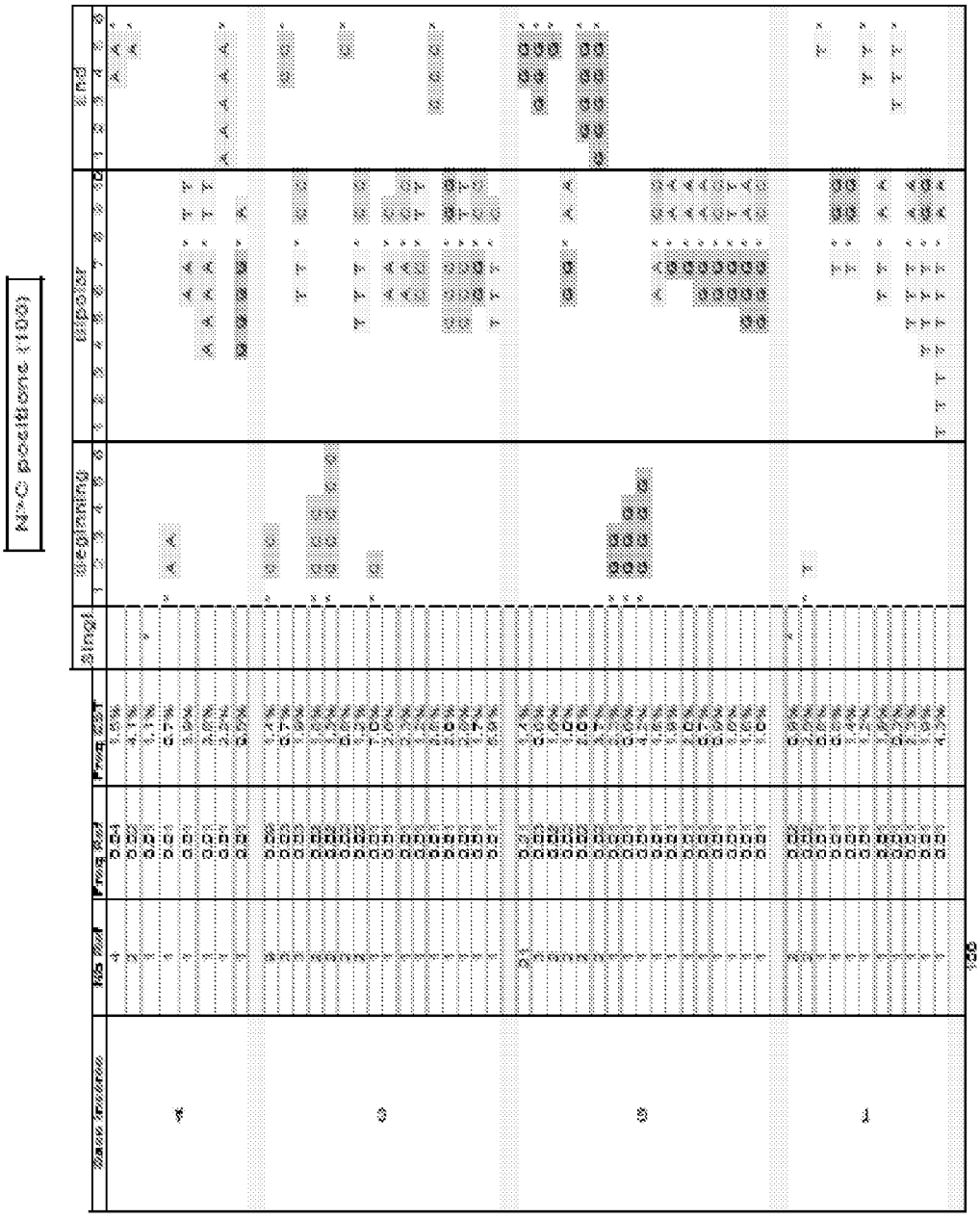
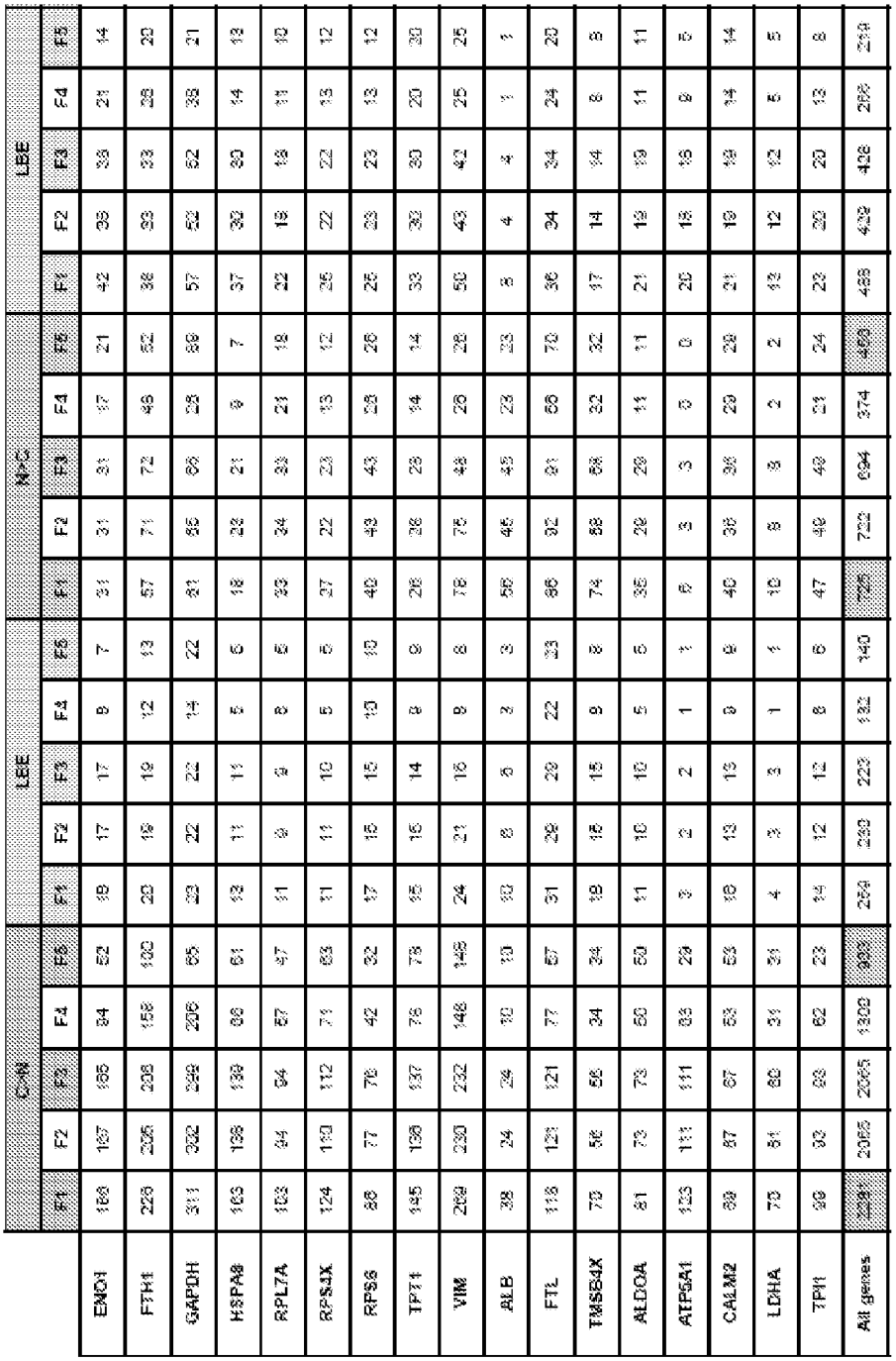

SNP positions (100)
Base position
KB Ref
Freq Ref
Freq Est
Singl
Segments
Ellipses
Rnd

Figure 5p

	Test Numbers					CAN					LBE				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
ENO1	814	586	557	331	220	217	183	163	111	73	40	37	37	18	14
FTH1	522	535	533	320	335	283	278	278	226	182	30	26	25	18	18
GAPDH	812	752	746	534	440	372	367	365	232	184	42	37	37	25	29
HSPA8	513	423	415	187	190	181	161	160	75	68	32	28	26	11	10
RPLP0	337	234	231	178	184	138	128	127	78	65	16	14	14	9	9
RPS4X	371	340	338	193	179	161	132	135	94	75	21	19	19	10	10
RPS6	432	387	397	238	226	129	120	119	68	58	31	26	29	17	16
TPT1	438	452	455	302	300	171	164	165	92	82	31	28	28	20	20
VIM	752	644	598	337	337	347	365	260	174	174	36	31	30	15	15
ALB	194	117	117	49	49	94	89	86	38	33	9	4	4	1	1
FTL	678	644	644	494	444	294	213	212	133	127	49	45	44	35	31
TMSB4X	352	309	302	183	169	144	124	124	88	86	20	19	19	8	9
ALDOA	336	297	297	174	174	118	102	102	81	81	22	19	19	11	11
ATP5A1	238	211	211	109	88	129	114	114	82	29	9	8	9	4	3
CALM2	374	336	335	245	245	109	103	103	82	82	26	23	23	16	16
LDHA	181	160	160	75	73	83	89	88	33	33	9	7	7	4	4
TPI1	379	331	330	225	150	143	142	142	83	47	20	18	18	10	10
All genes	7644	6782	6701	4148	3759	3038	2797	2759	1874	1369	448	381	378	236	223

F1 : EST alignment >= 500
F2 : >=70% EST aligned (once)
F3 : removal of paralogues and pseudogenes
F4 : 500 clustered at each alignment extremity
F5 : longer contig/gene

Figure 6a



	C->N					LBE					N->C					LBE				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
ENO1	164	167	165	84	62	19	17	17	9	7	31	31	31	17	21	42	39	39	21	14
FTH1	229	235	208	158	100	20	19	19	12	13	57	71	72	46	52	38	93	33	28	20
GAPDH	311	302	249	206	65	23	22	22	14	22	81	85	66	28	69	57	52	52	38	21
HSPA8	163	168	139	86	61	13	11	11	5	6	18	33	21	9	7	37	30	30	14	13
RPL7A	103	94	94	67	47	11	9	9	8	6	33	34	33	21	18	22	18	18	11	10
RPS3X	124	110	112	71	63	11	11	10	5	6	27	22	23	13	12	25	22	22	13	12
RPS6	88	77	76	42	32	17	15	15	10	10	40	43	43	23	26	25	23	23	13	12
TPT1	145	136	137	76	78	15	15	14	8	8	26	38	28	14	14	33	32	30	20	30
VIM	269	230	232	148	148	24	21	16	9	8	78	75	48	28	28	50	43	42	25	25
ALB	38	24	24	10	10	10	8	8	2	3	56	48	45	23	23	8	4	4	1	1
FTL	118	121	121	77	67	31	29	29	22	33	86	92	91	66	70	36	34	34	24	20
TMSB4X	70	58	58	34	34	19	16	15	9	8	74	68	68	32	32	17	14	14	8	8
ALDOA	81	73	73	60	60	11	10	10	5	6	36	29	29	11	11	21	19	19	11	11
ATP5A1	123	111	111	83	29	9	2	2	1	1	6	3	3	0	0	25	18	15	8	5
CALM2	89	87	87	63	63	19	13	13	8	6	40	36	36	29	29	21	19	19	14	14
LDHA	73	81	80	31	31	4	3	3	1	1	10	8	8	2	2	12	12	12	5	5
TPI1	96	93	93	62	23	14	12	12	8	6	47	46	46	21	24	23	20	20	13	8
All genes	229	2065	2065	1309	903	264	233	223	132	140	725	722	694	374	453	469	429	428	266	230

(C->N)/(N->C)

F1	F2	F3	F4	F5
3,15	2,86	2,98	3,48	2,05

Figure 6b

	C>N-LBE					N>C-LBE				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
ENO1	108	150	148	86	45	0	0	0	0	7
FTH1	208	156	157	146	87	19	38	39	22	32
GAPDH	288	260	277	192	43	4	13	14	0	68
HSPA8	150	127	128	61	56	0	0	0	0	0
RPLTA	92	85	85	51	42	11	16	15	10	6
RPS4X	113	99	102	68	68	2	0	1	0	0
RPS6	89	82	81	32	22	15	20	20	13	14
TPT1	130	121	123	69	69	0	0	0	0	0
VIM	245	209	216	140	140	23	32	6	1	1
ALB	28	16	16	7	7	48	41	41	22	22
FTL	87	92	92	55	34	50	58	57	32	50
TMSB4X	52	41	41	28	26	67	54	54	24	24
ALDOA	70	63	62	45	45	14	10	10	0	0
ATP5A1	120	109	109	62	23	0	0	0	0	0
CALM2	53	54	54	44	44	19	17	17	15	15
LDHA	88	58	67	30	30	0	0	0	0	0
TPI1	95	81	81	56	17	24	29	29	8	16
All genes	2022	1835	1842	1188	783	291	329	303	147	357

(C>N)-LBE/(N>C)-LBE				
F1	F2	F3	F4	F5
6,95	5,59	6,08	7,95	3,09

Figure 6c-1

	C>N	LBE	N>C	LBE	(C>N)/(N>C)	[(C>N)-LBE]/[(N>C)-LBE]
F1	2281	259	725	488	3,15	6,95
F2	2065	230	722	429	2,86	5,59
F3	2065	223	694	428	2,98	6,08
F4	1300	132	374	266	3,48	7,95
F5	933	140	456	219	2,05	3,09

Figure 6c-2

	Test Numbers		C>N		LBE		C>N		LBE		N>C		LBE		C>N-LBE		N>C-LBE	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
ENO1	614	74	217	28	40	4	192	18	18	3	31	12	42	3	166	12	0	0
FTH1	692	263	263	76	80	18	226	64	20	10	57	22	39	13	206	44	19	0
GAPDH	812	333	372	118	42	20	311	76	23	14	61	43	57	19	288	62	4	24
HSPA8	513	88	181	44	32	3	183	37	13	2	19	7	37	6	150	35	0	0
RPL7A	337	51	136	21	19	3	103	18	11	1	33	5	22	3	92	16	11	2
RPS4X	371	104	151	36	21	8	124	24	11	3	27	2	25	6	113	21	2	0
RPS6	432	140	126	32	31	12	88	24	17	6	40	8	26	8	69	18	15	0
TPT1	469	302	171	80	31	21	145	63	15	12	28	27	33	19	133	51	0	0
VIM	752	119	347	36	33	7	269	31	24	3	79	4	50	9	245	29	23	0
ALB	194	55	94	37	9	1	38	19	10	2	58	18	8	2	28	17	48	15
FTL	679	364	224	102	49	27	118	29	31	21	83	78	36	14	87	2	60	65
TMSB4X	352	177	144	82	23	11	70	18	16	10	74	45	17	7	52	8	57	38
ALDOA	338	53	116	28	23	2	91	22	11	1	35	5	21	3	70	22	14	2
ATP5A1	238	9	128	4	8	0	123	4	3	0	6	5	20	0	120	4	0	0
CALM2	374	58	136	16	28	3	89	31	18	2	46	5	21	3	58	6	19	2
LDHA	181	0	83	7	0	0	70	6	4	0	10	2	12	0	66	6	0	0
TPH1	378	65	146	19	20	4	89	21	14	2	47	5	23	3	85	9	24	5
All genes	7644	2268	3009	747	448	145	2251	498	259	92	724	292	488	119	2077	493	291	158

F1 EST alignment >= 80b
F2 Clip 400

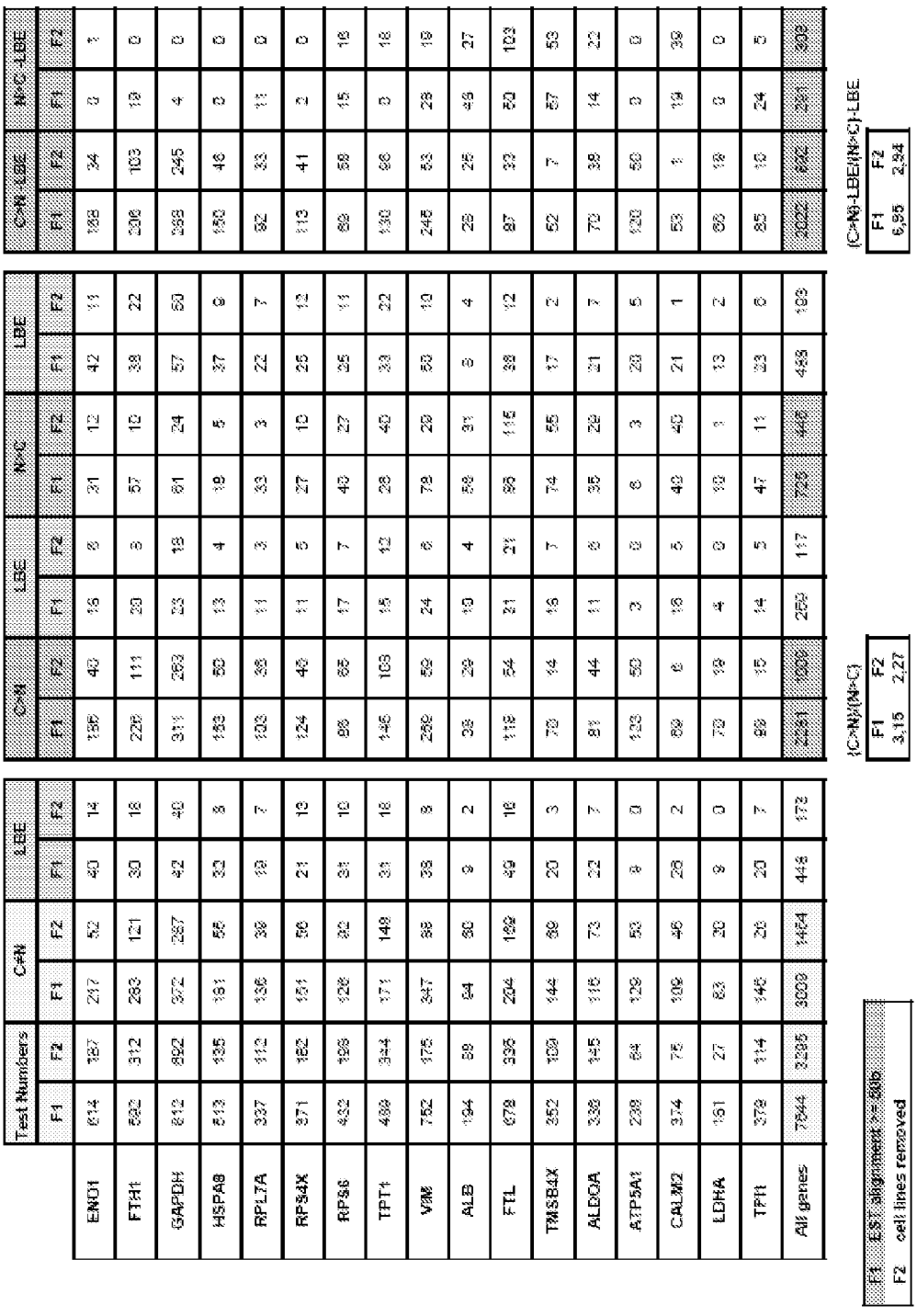
(C>N)/(N>C)

F1	F2
3,15	1,56

(C>N)-LBE/(N>C)-LBE

F1	F2
6,35	2,15

Figure 6d



	Test Numbers		C=N		LBE	
	F1	F2	F1	F2	F1	F2
ENO1	614	187	217	52	40	14
FTH1	592	312	283	121	30	18
GAPDH	612	892	372	287	42	40
HSPA8	613	135	191	56	32	8
RPL7A	337	112	136	36	18	7
RPS4X	371	162	151	56	21	10
RPS6	432	198	128	82	31	10
TPT1	499	344	171	148	31	18
VIM	752	175	347	88	33	8
ALB	194	88	64	80	9	2
FTL	679	336	284	102	49	16
TMSB4X	352	109	144	89	20	3
ALDOA	338	145	116	73	22	7
ATP5A1	239	64	129	53	8	0
CALM2	374	75	109	46	28	2
LDHA	161	27	83	30	9	0
TPI1	379	114	146	26	20	7
All genes	7544	3295	3009	1464	448	172

	C=N		LBE		N>C		LBE		C=N-LBE		N>C-LBE	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
ENO1	186	40	18	8	31	12	42	11	188	34	8	1
FTH1	229	111	29	8	57	10	38	22	306	103	19	0
GAPDH	311	263	53	18	61	24	57	60	368	245	4	0
HSPA8	183	50	13	4	19	5	37	0	150	46	8	0
RPL7A	103	36	11	3	33	3	22	7	92	33	11	0
RPS4X	124	46	11	5	27	10	25	12	113	41	2	0
RPS6	86	65	17	7	40	27	25	11	89	58	15	16
TPT1	146	103	15	12	28	40	33	22	130	66	3	16
VIM	260	92	24	6	79	29	53	10	246	83	29	19
ALB	38	29	10	4	58	31	8	4	28	25	49	27
FTL	118	54	21	21	95	115	38	12	97	33	20	103
TMSB4X	78	14	16	7	74	55	17	2	62	7	57	53
ALDOA	81	44	11	6	35	29	21	7	70	38	14	22
ATP5A1	133	50	3	0	8	3	20	5	120	50	3	0
CALM2	69	6	16	5	40	40	21	1	53	1	19	39
LDHA	70	19	4	0	10	1	13	2	66	18	0	0
TPI1	98	15	14	5	47	11	33	6	85	10	24	5
All genes	2241	1006	252	117	725	448	493	199	2022	682	261	309

(C>N)/(N>C)

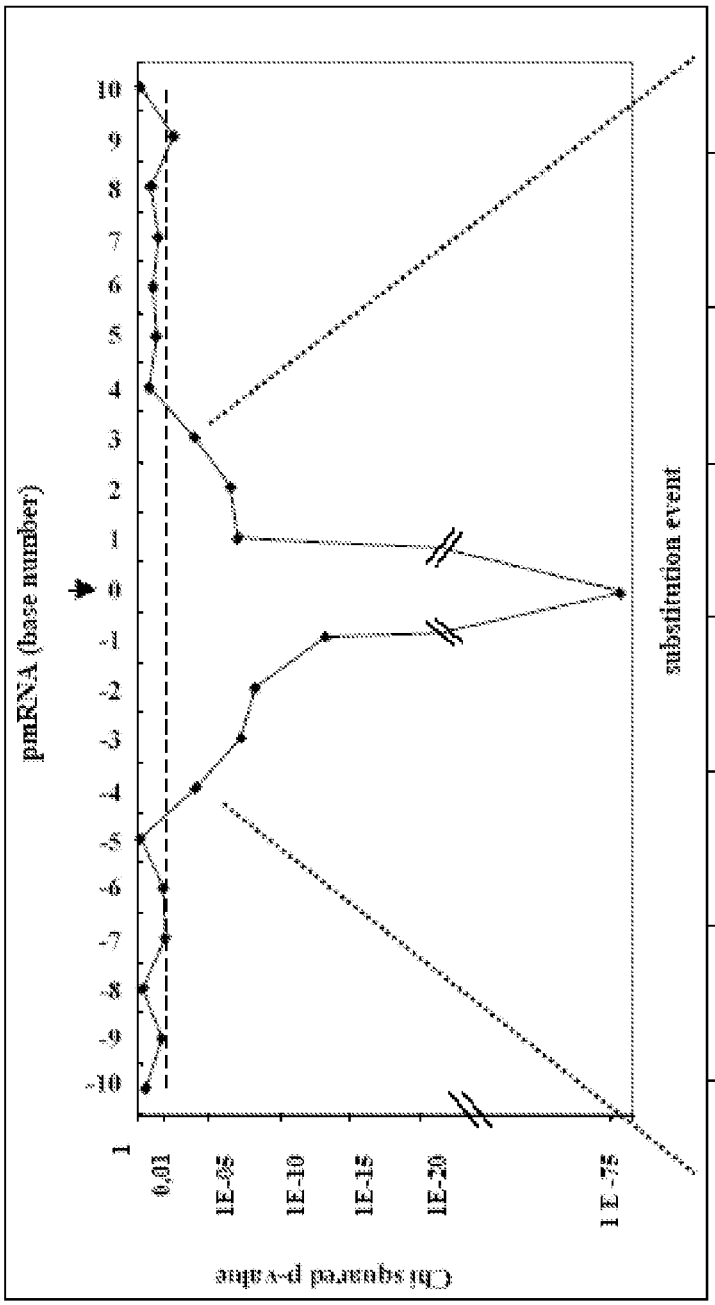
F1	F2
3,15	2,27

(C>N)-LBE/(N>C)-LBE

F1	F2
6,95	2,94

F1 EST alignment >= 50b
F2 cell lines removed

Figure 6e



	b-4		b-3		b-2		b-1		b0		b+1		b+2		b+3	
	H	NH	H	NH	H	NH	H	NH	H	NH	H	NH	H	NH	H	NH
A	28	24	24	27	25	27	30	22	33	23	29	25	27	26	28	24
T	23	25	28	23	26	23	22	27	32	22	22	24	24	26	24	24
C	24	25	25	24	21	26	23	26	21	26	21	25	20	24	22	26
G	25	26	23	26	28	24	25	25	14	29	28	26	29	24	26	26

Figure 7a

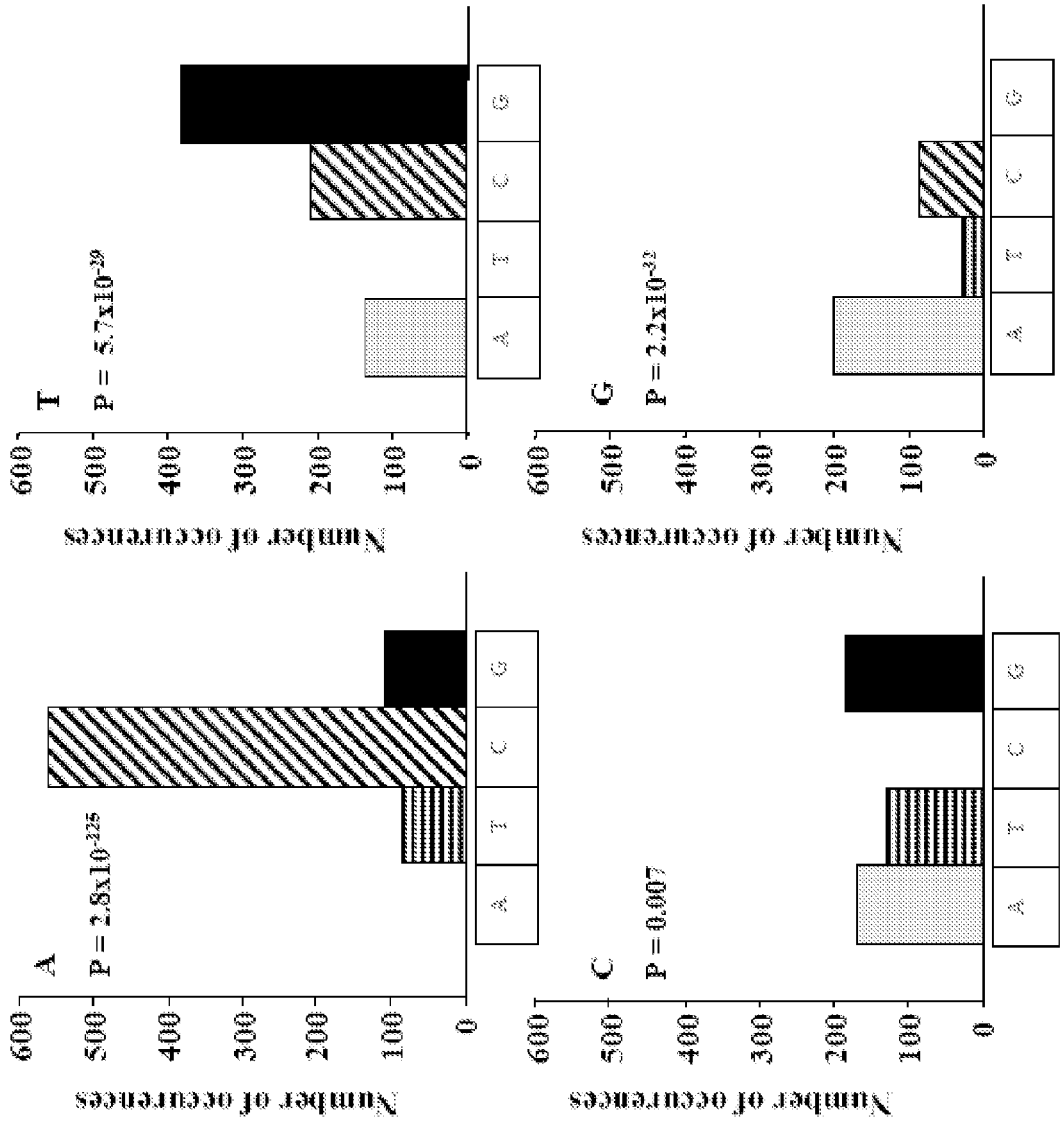

T
P = 5.7x10-39
G
P = 2.2x10-33
A
P = 2.8x10-225
C
P = 0.007
Number of occurrences
0
100
200
300
400
500
600
A
T
C
G

Figure 7b

A. All deviations (n = 2281)

affected \ replacement	A	T	C	G	%
A		85	560	107	33
T	139		210	384	32
C	169	130		185	21
G	199	28	85		14
%	22	11	37	30	

B. Deviations towards b-1 (n = 799)

affected \ replacement	A	T	C	G	%
A		41	201	42	36
T	28		56	83	21
C	77	41		126	30
G	75	9	20		13
%	23	11	35	31	

C. Deviations towards b+1 (n = 530)

affected \ replacement	A	T	C	G	%
A		24	52	37	21
T	27		45	177	47
C	38	48		15	19
G	40	11	16		13
%	20	16	21	43	

D. Deviations towards neither b-1 nor b+1 (n = 347)

affected \ replacement	A	T	C	G	%
A		7	148	8	47
T	42		50	9	29
C	17	10		23	14
G	7	2	24		10
%	19	5	64	12	

E. Deviations towards b-1 = b+1 (n = 339)

affected \ replacement	A	T	C	G	%
A		5	52	12	20
T	11		12	95	35
C	25	30		17	21
G	71	1	8		24
%	31	11	21	37	

F. Deviations within Triplet (n = 266)

affected \ replacement	A	T	C	G	%
A		8	107	8	46
T	31		47	20	37
C	12	1		4	6
G	6	5	17		11
%	19	5	64	12	

Figure 7c

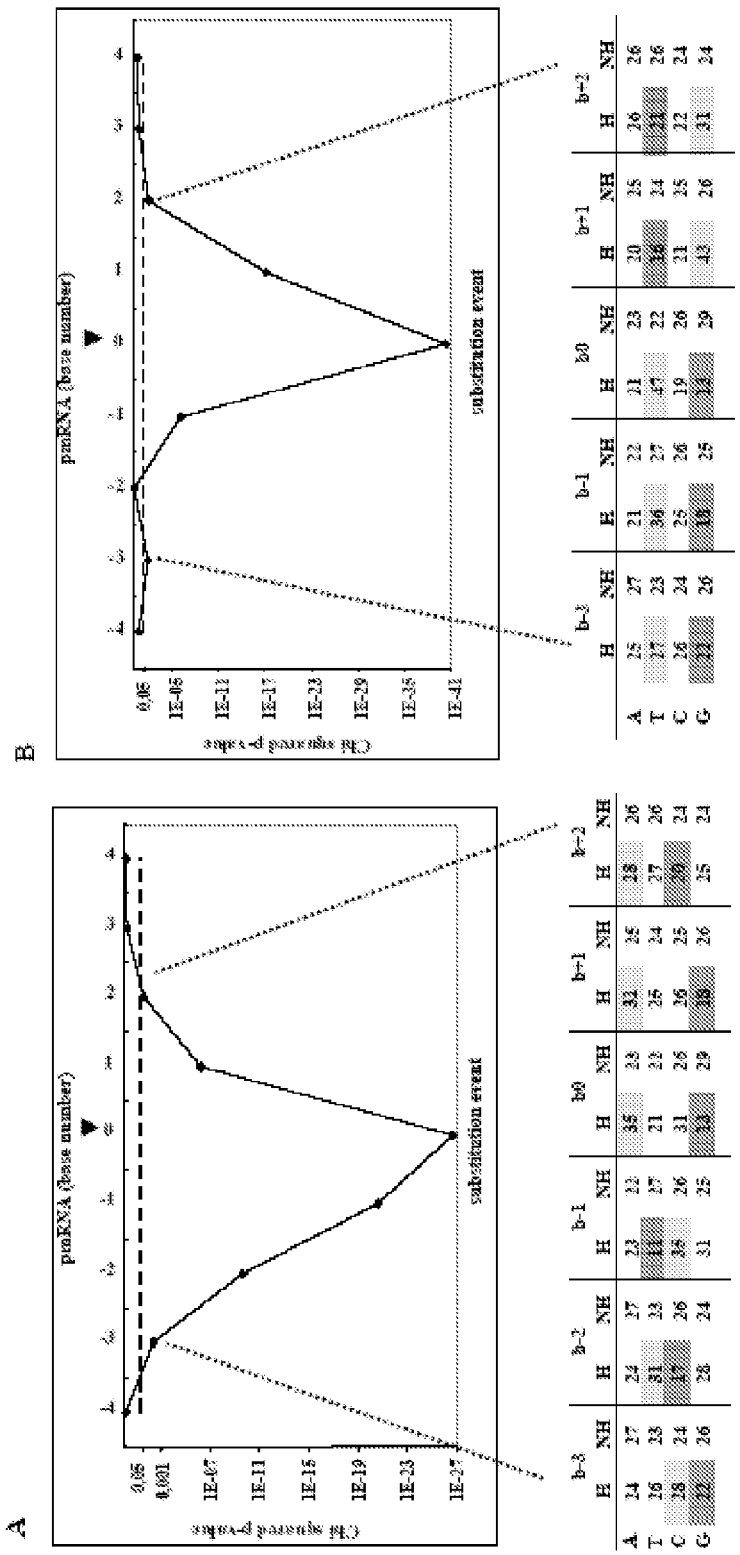

A
B
peaKNA (base number)
Chi squared p-value
subclonation event
b-3
b-2
b-1
b0
b+1
b+2
H
NH
A
T
C
G

Figure 7d

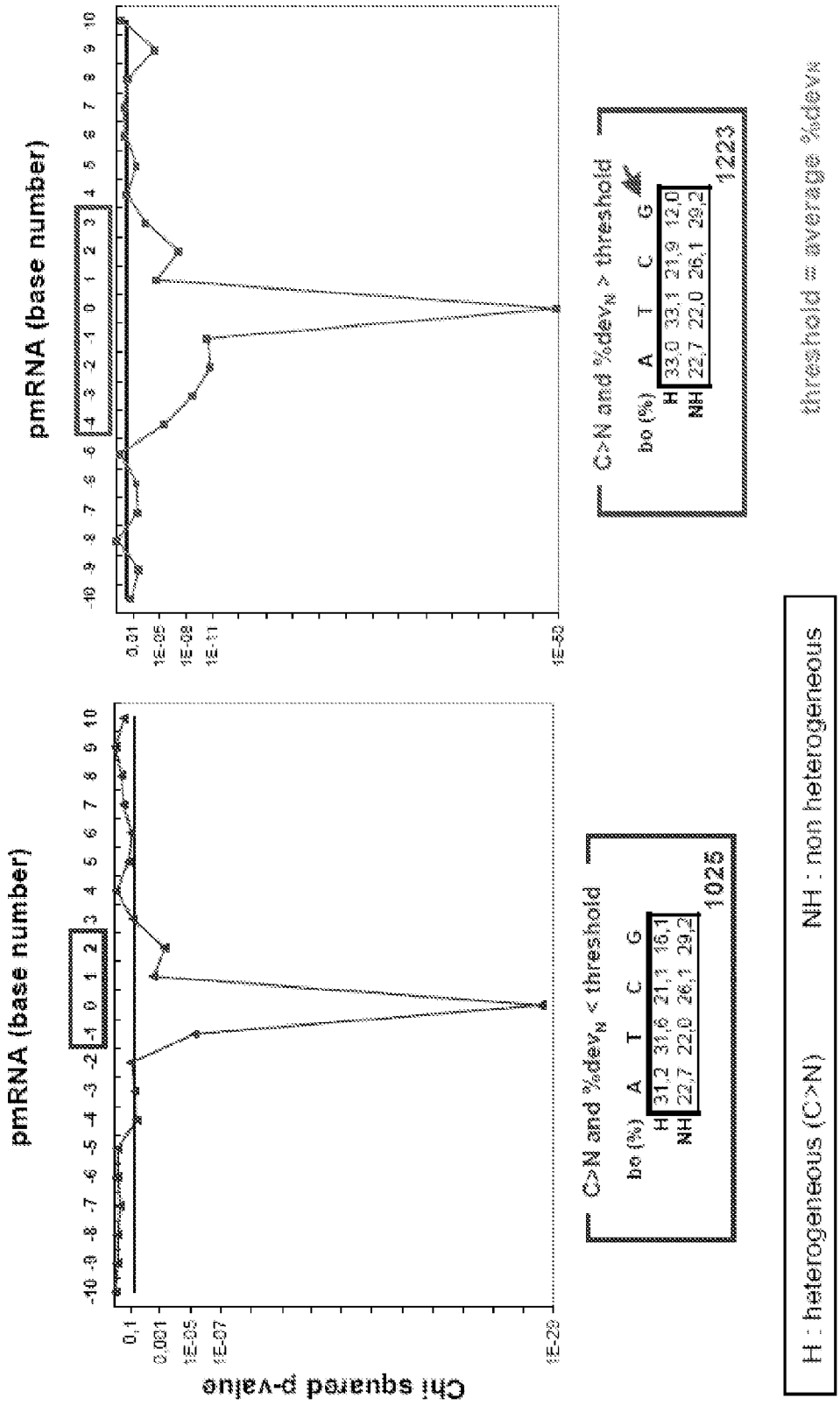

pmRNA (base number)
Chi squared p-value
0,1
0,001
1E-05
1E-07
1E-29
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
C>N and %dev_N < threshold
bo (%) A T C G
H 31,2 31,6 21,1 16,1
NH 22,7 22,0 26,1 29,2
1025
pmRNA (base number)
0,01
1E-05
1E-08
1E-11
1E-50
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
C>N and %dev_N > threshold
bo (%) A T C G
H 33,0 33,1 21,9 12,0
NH 22,7 22,0 26,1 29,2
1223
H : heterogeneous (C>N)
NH : non heterogeneous
threshold = average %dev_N

Figure 7e

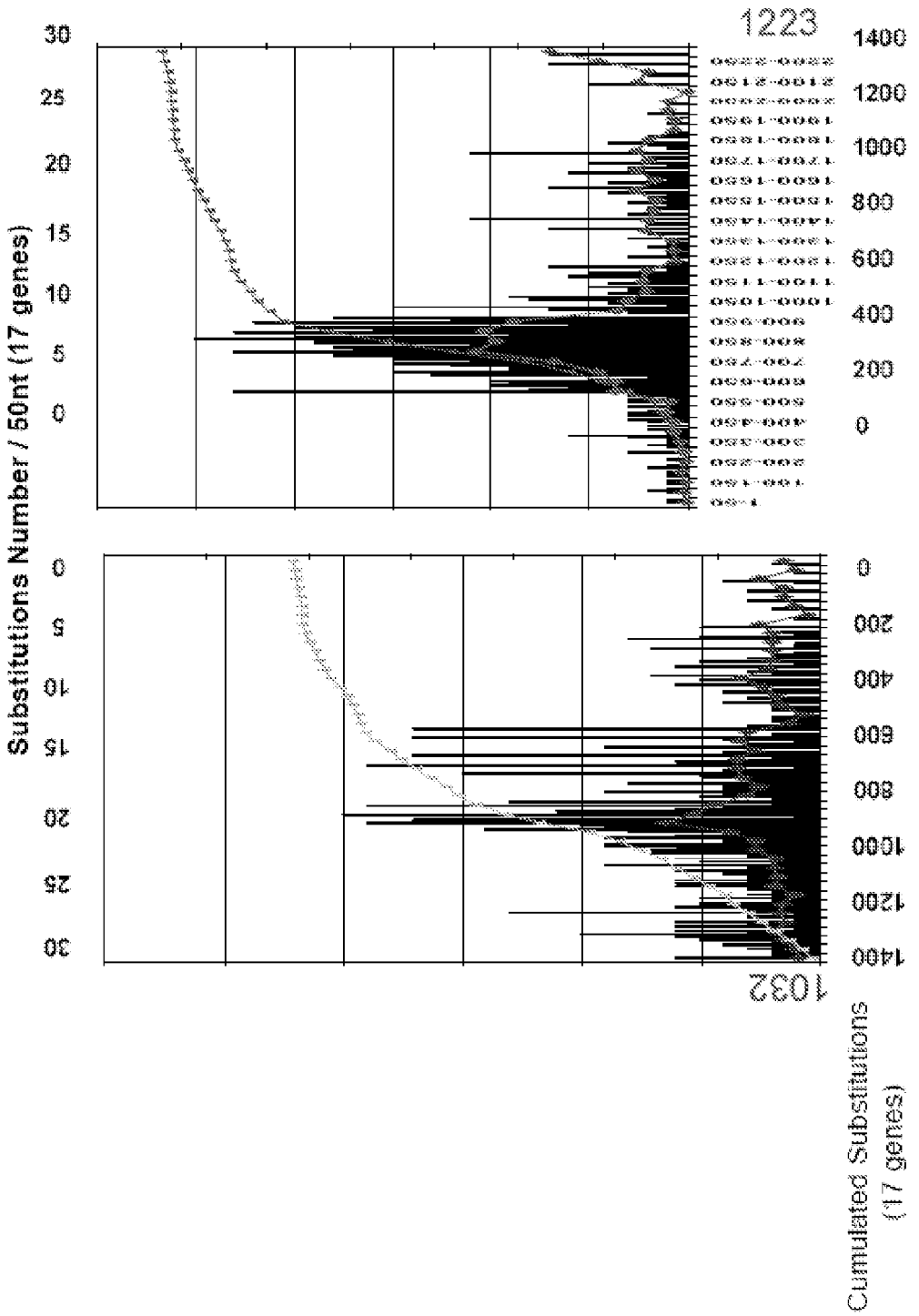

Substitutions Number / 50nt (17 genes)
Cumulated Substitutions (17 genes)
1223
1032

Figure 7f

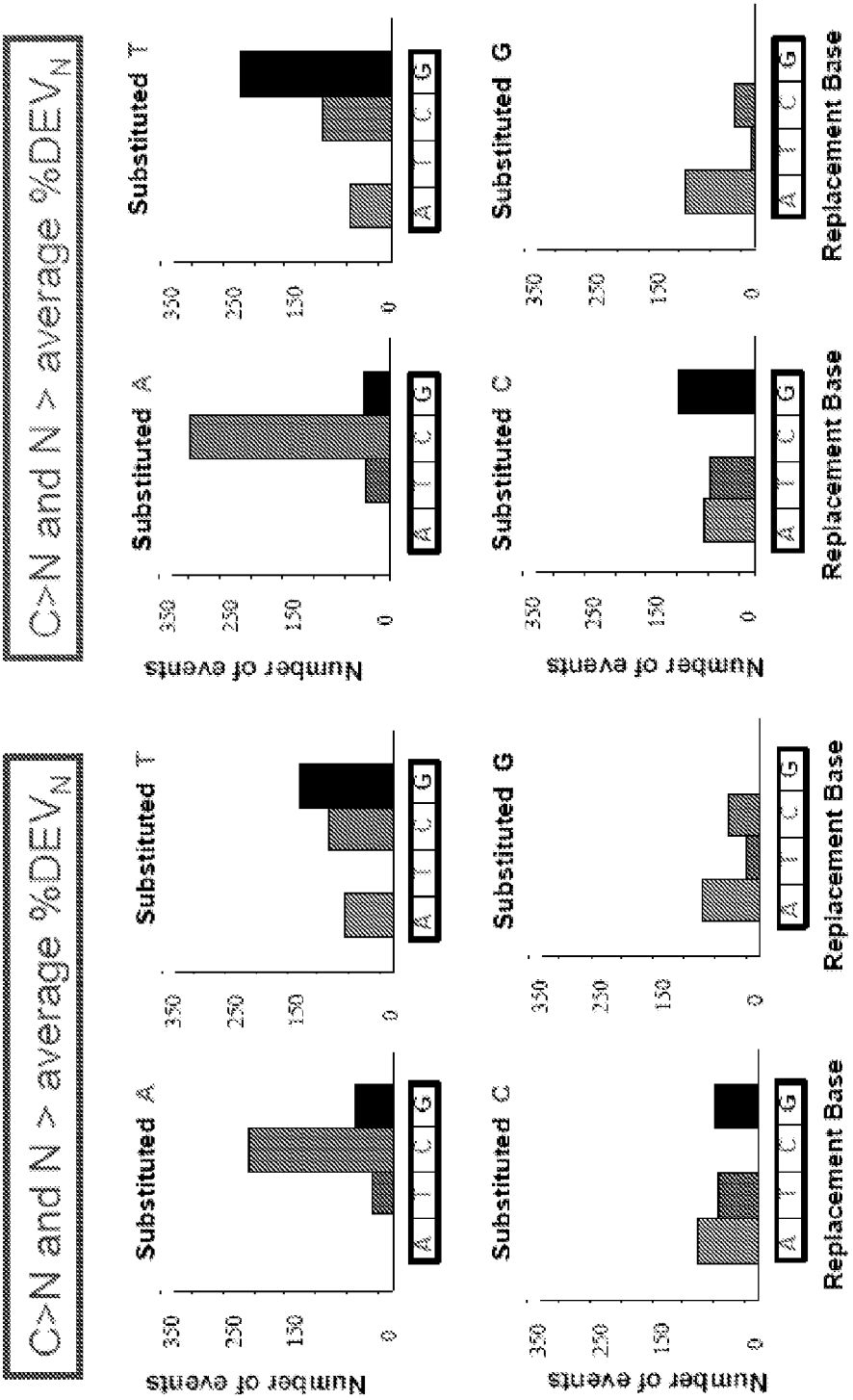

C>N and N > average %DEV$_N$
C>N and N > average %DEV$_N$
Substituted A
Substituted T
Substituted C
Substituted G
Number of events
Replacement Base
A T C G
0
150
250
350

Figure 7g

>vim nat, 1847 bases, 3DE7CB32

>vim variant, 1847 bases, 3DE7CB32

Figure 8a

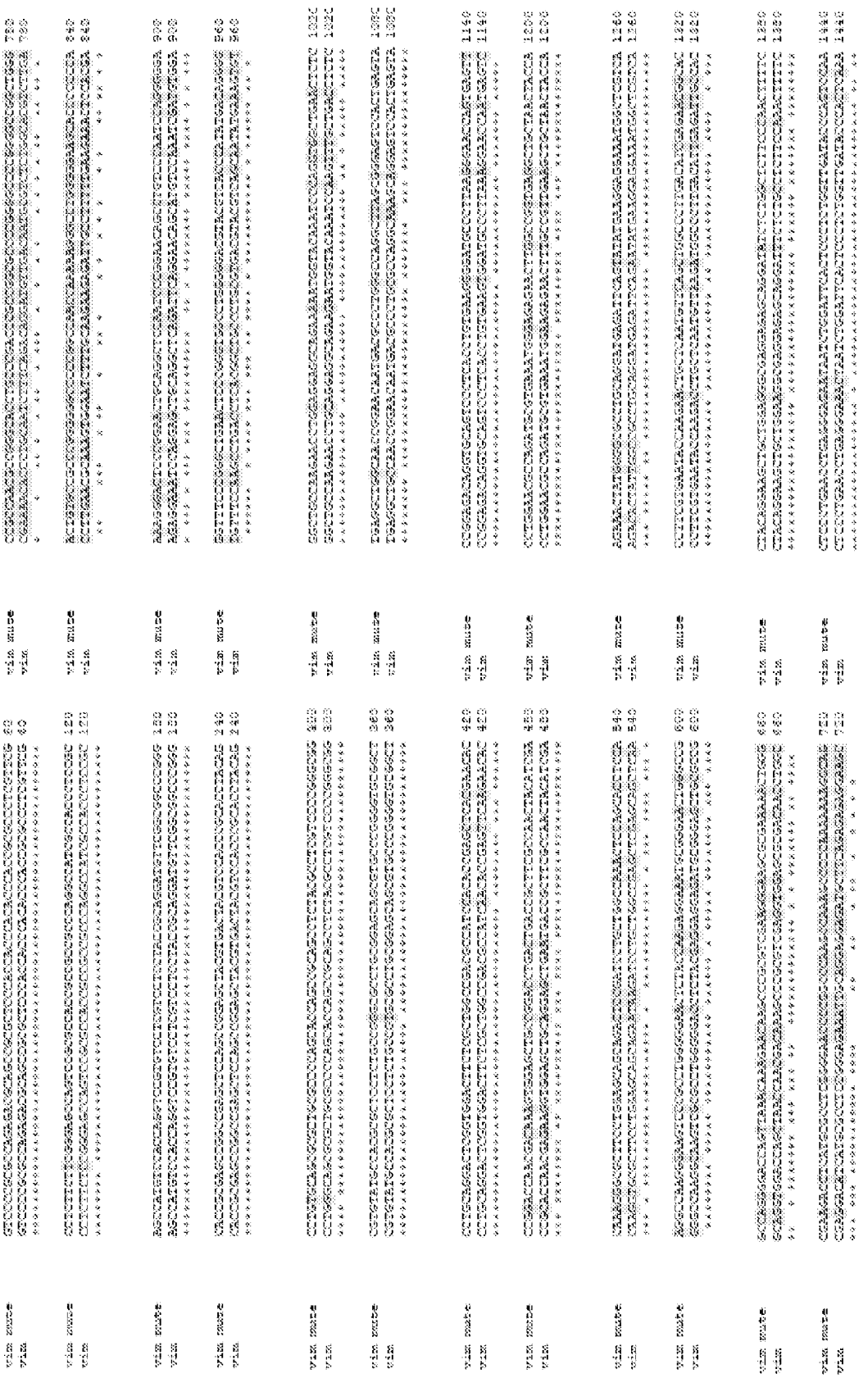


Figure 8b

```
vim mutB  GAGGACACTTTGATTAGACGGTGGACACTAGAATGGCCAGGTTATCCAACGAAACTTC 1500
vim       AAGGACACTTCTGATTAGACGGTGGAAACTAGAATGGCAGGTTATCCAACGAAACTTC 1500
           ********** ****** ******** ** ********** ********** ********** 

vim mutB  TTAGCGTCACGATGAACTTGATTCAAATTGCCACACTCAGTGCAGCGTATATTACCC 1560
vim       TCAGCATCACGATGACCTTGATAAAATTGCACACACTCAGTGCAGCGATATTACCA 1560
          * **** ********** ***** * ********** ********** ********** 

vim mutB  GGCAGAAACCAAAAGATCCCTATCTTTGAAAACCAGGTTTCAAGGGCCTTCTGCAGT 1620
vim       GCAAGATAAAAAGAAAATCCATATCTTTAAAGAACCAGCTTTCAAGTGCCTTCTGCAGT 1620
          * **** * *** * **** ****** * ** *** ********** *** ********** 

vim mutB  TTTTCAACACCCAGATAAATCCGGAATAGGATTAGCTCTAGTTCTTAACAACCCAC 1680
vim       TTTTCAGGAGCCAAGATAGATTTGGATAGGATTAGCTCTAGTTCTTAACAACCCACA 1680
          ****** ** *** ****** ** ********** ********** ********** 

vim mutB  CTCCTCCAAGATTTTGAAAAAGTTTCCCACATTATTAGTTTCCCGAAAAATTTGGTGT 1740
vim       CTCCTACAAGATTTAGAAAAAGCTTTACAACATTAATCTAGTTTACAGAAAAATTTGTGC 1740
          ***** ********** ********** * ***** ********** * ******** * *** 

vim mutB  TAGAAAACTTTTACAAGGTATCCGGAAACCCTTACCCTGCTCTTTTTTTCTAGCAA 1800
vim       TAGAATACTTTTTAAAGGTATTTTGATACCATTAACTGCTCTTTTTTTCCAGCAA 1800
          ****** ******** ***** *** *** * ***** ********** ********** 

vim mutB  GTATCCAACCAACTGGTTCTGCTTCAATAAATCTTTGGGAAAACTC 1847
vim       GTATCCAACCAACTGGTTCTGCTTCAATAAATCTTTGGAAAACTC 1847
          ********** ********** ********** ********** 
```

Figure 8c

>VIM 1 a 1481 Phase 1 (sens direct) 467 aa
MSTRSVSSSSYRRMFGGPGTASRPSSSRSYVTISTRTYSLGSALRPSTSRSLYASSPGGVYATRSSAVRLRSSVPGVRLLQDSVDFSLADAINTEFKNTRTNEKVELQEL
NDRFANYIDKVRFLEQQNKILLAELEQLKGQGKSRLGDLYEEEMRELRRQVDQLTNDKARVEVERDNLAEDIMRLREKLQEEMLQREEAENTLQSFRQDVDNASLARLDL
ERKVESLQEEIAFLKKLHEEEIQELQAQIQEQHVQIDVDVSKPDLTAALRDVRQQYESVAAKNLQEAEEWYKSKFADLSEAANRNNDALRQAKQESTEYRRQVQSLTCEV
DALKGTNESLERQMREMEENFAVEAANYQDTIGRLQDEIQNMKEEMARHLREYQDLLNVKMALDIEIATYRKLLEGEESRISLPLPNFSSLNLRETNLDSLPLVDTHSKR
TLLIKTVETRDGQVINETSQHEDDLE*

>VIM variant 1 a 1723 Phase 1 (sens direct) 574 aa
MSTRSVSSSSYRRMFGGPGTASRPSSSRSYVTTSTRTYSLCSALRPSTSRSLYASSPGGVYATRSSAGRLRSSVPGVRLLQDSVDFSLADAIHIELINTRTNEKVELPDL
IDRFANYIDKGRFLEQQTPILLAKLQHLQGQGKSRLGELYQEEMRELGRPGDQLNNEQARVEGKREKLGEDLMRLGEPRPKPKPFKKGSSRQRRAPARPAGAPGPRGRLGL
CRPGGPRPTKKGLGEAPPQRGLPELQAPIPEQPVPIQGEVSPAELPGGLGIVRHPYDRGAAANNLEEAEKWYKSQVAELSEAGNRNNDALGQA*BESTEYRRQVQSLICEG
DALNGTSEFLERQMREMGENLAGEAANYQETMGRLQDEIQYMKEEMARHLPEVQELLNVQLALDIENGTYRKLLEGEESRISLALPNFSSLNLRENNLDSLPLVETQSKR
TLLI*TVDTRDGQVIHETS*RHDELDFKLPTLSAAIYYPAENQKDFYL*KPGEKGPSAVFQERQINPE*E*ALVLNWRTSSKILKKVSHII*FPEKFGGRKLFTRYPEMP
YTCSFFSSKYPTNWVLLQ*ILGKT

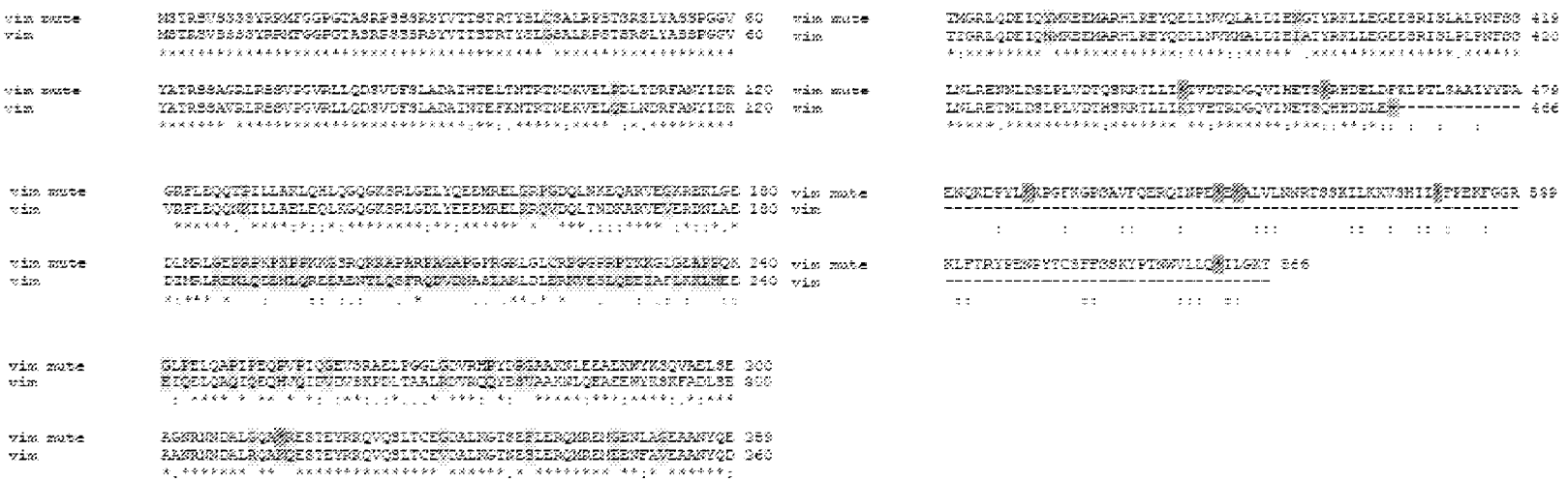


Figure 8d

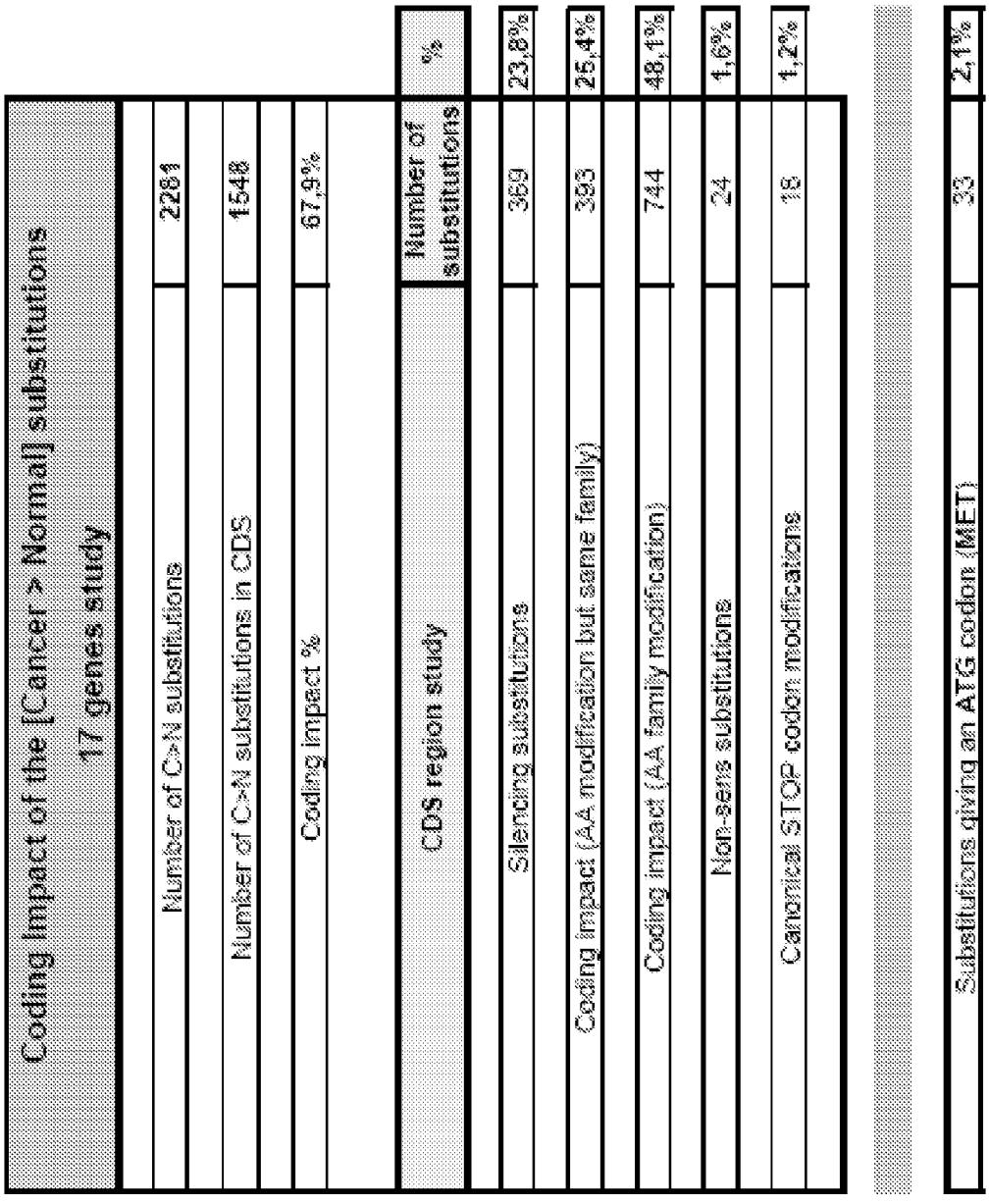

Coding Impact of the [Cancer > Normal] substitutions
17 genes study
Number of C>N substitutions
2281
Number of C>N substitutions in CDS
1548
Coding impact %
67,9%
CDS region study
Number of substitutions
%
Silencing substitutions
369
23,8%
Coding impact (AA modification but same family)
393
25,4%
Coding impact (AA family modification)
744
48,1%
Non-sens substitutions
24
1,6%
Canonical STOP codon modifications
18
1,2%
Substitutions giving an ATG codon (MET)
33
2,1%

Figure 9

normal proteins
additional peptide of proteins whose STOP codon is affected

Figure 10a

normal proteins
additional peptide of proteins whose STOP codon is affected

Figure 10b

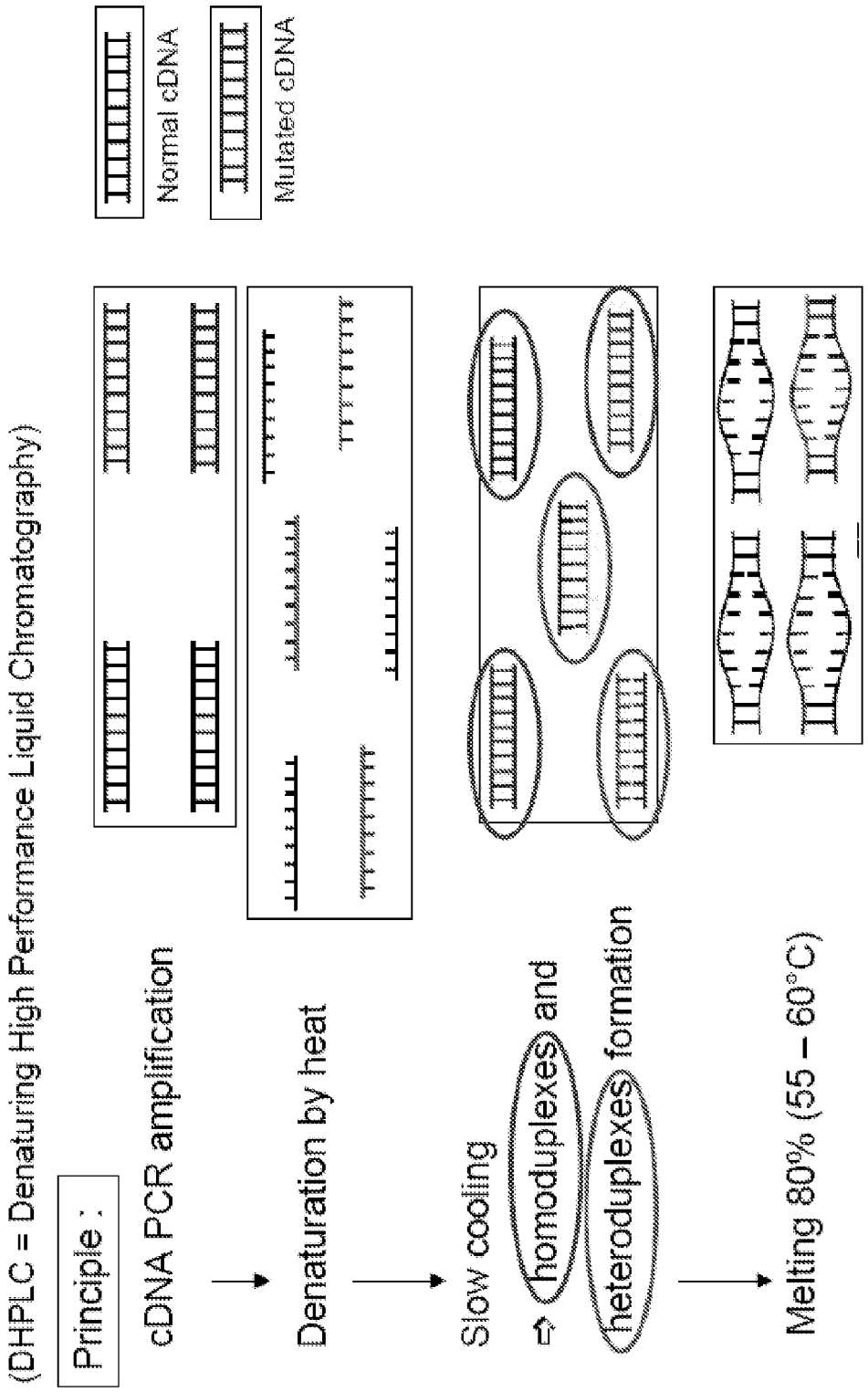

(DHPLC = Denaturing High Performance Liquid Chromatography)
Principle :
cDNA PCR amplification
Denaturation by heat
Slow cooling
homoduplexes and
heteroduplexes formation
Melting 80% (55 – 60°C)
Normal cDNA
Mutated cDNA

Figure 11a

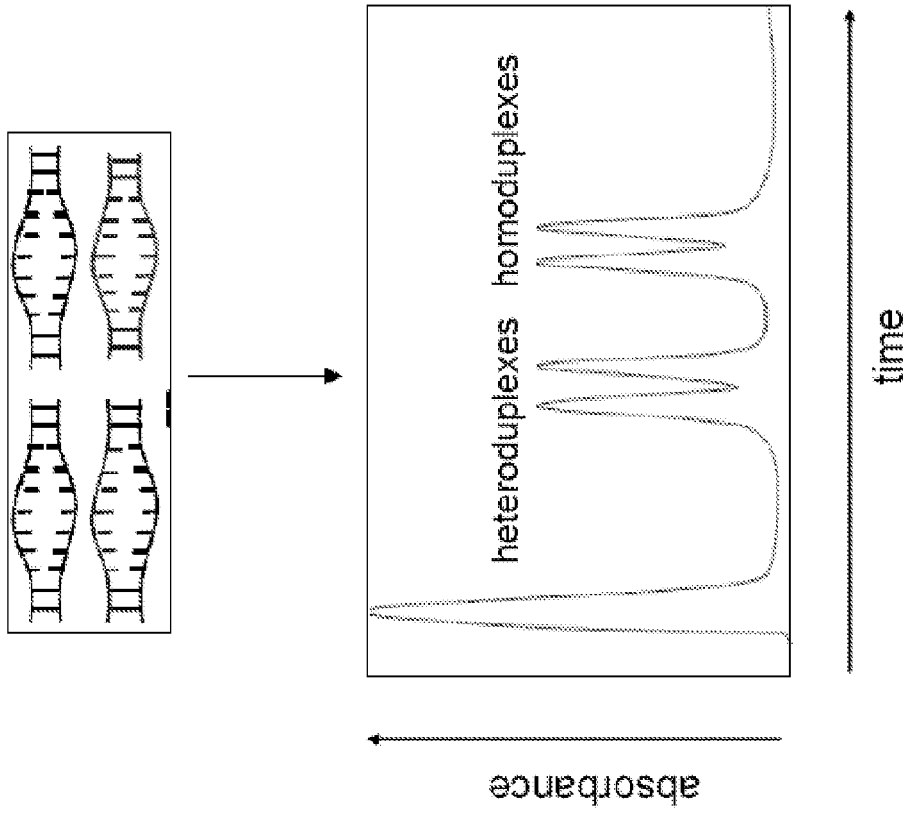

Injection of the sample into the DHPLC column :
⇨ heteroduplexes elute before homoduplexes
heteroduplexes
homoduplexes
absorbance
time

Figure 11b

Figure 2. Principle of mutation detection by DHPLC. *a*, DHPLC compares typically two chromosomes as a mixture of PCR products denatured at 95°C for 3 min and re-annealed over 30 min by gradual cooling from 95 to 65°C prior to analysis. In the presence of mismatch, not only are the original homoduplices formed, but also the sense and antisense strands of either homoduplex form heteroduplices. *b*, Heteroduplices denature more extensively at the analysis temperature (ranges from 50 to 70°C) and are eluted earlier than the homoduplices in the DNA Sep column; *c*, Corresponding chromatographic pattern shows four different peaks belonging to four different species of DNA

Source : Theru A. Sivakumaran et al., Current Science, 84 : 291 – 296, 2003

Figure 11c

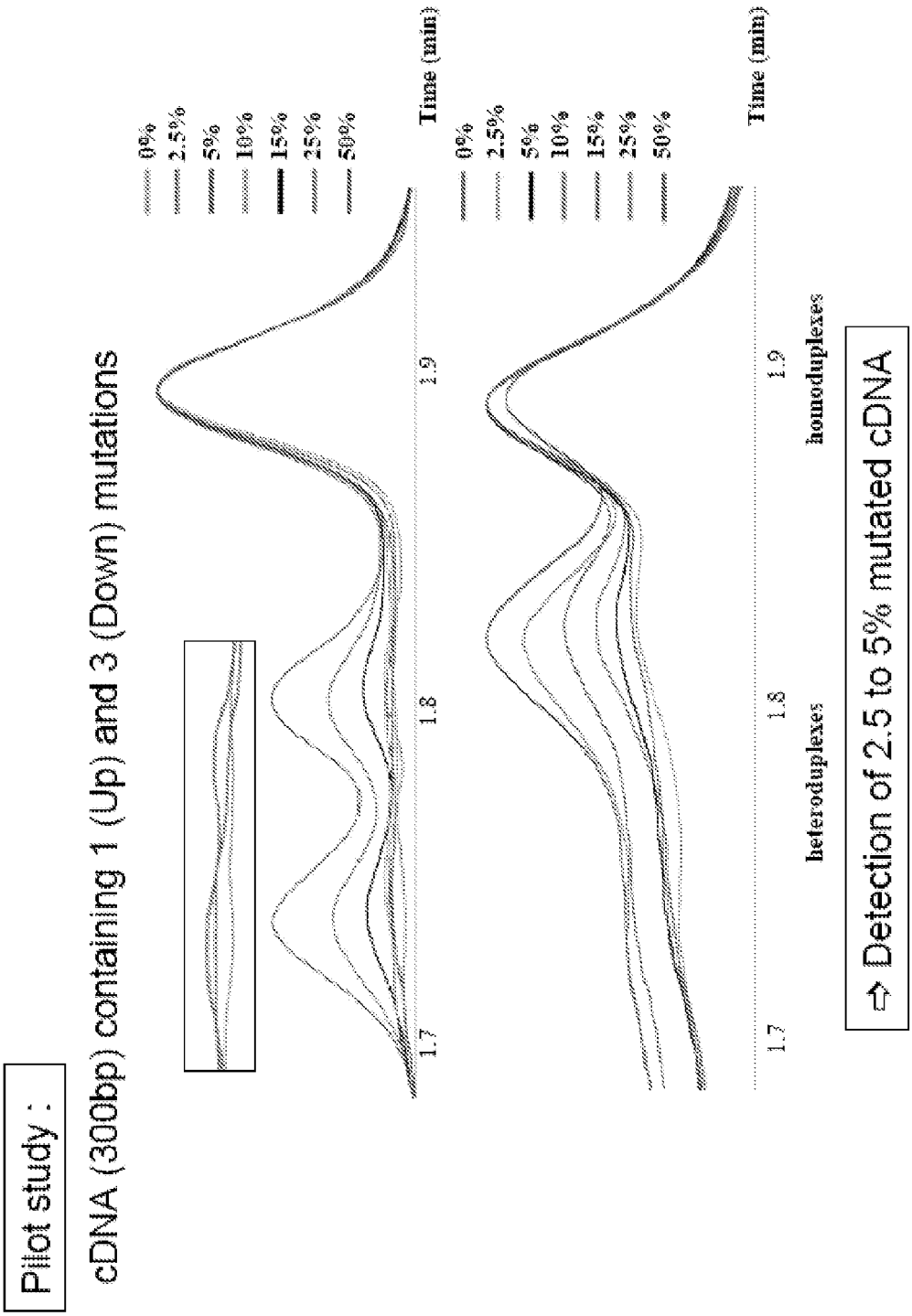

Pilot study :
cDNA (300bp) containing 1 (Up) and 3 (Down) mutations
0%
2.5%
5%
10%
15%
25%
50%
Time (min)
1.7
1.8
1.9
heteroduplexes
homoduplexes
⇨ Detection of 2.5 to 5% mutated cDNA

Figure 11d

```
5'CTGCCAAGGCTGTGGGCAAG 3'    713-732
5'GAAATGAGCTTGACAAAGTGGTCGT 3'    986-1010
Amplification product length : 298    713-1010
GC% : 58,0%
Suggested hybridization temperature (TA) : 58,9
Primer strand + :
Primer strand - :
Primer :  :    5' CTGCCAAGGCTGTGGGCAAG 3'
                  ||||||||||||||||||||
Target : 713 : 5' CTGCCAAGGCTGTGGGCAAG 3'
score : 174
TM : 74,4 / 52,8 / 60,9
Primer :  :    3' ACGACCACTTTGTCAAGCTCATTTC 5' (complement)
                  |||||||||||||||||||||||||
Target : 986 : 5' ACGACCACTTTGTCAAGCTCATTTC 3'
score : 181
TM : 72,5 / 50,9 / 58,314
```

>gi|83641890|ref|NM_002046.3| Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), mRNA¶

¶

CTCTCTGCTCCTCCTGTTCGACAGTCAGCCGCATCTTCTTTTGCGTCGCCAGCCGAGCCACATCGCTCAGACACCATGGGGAAGGTGAAGGTCGGAGTCAACGGATTTGGTCGTATTG¶

GGCGCCTGGTCACCAGGGCTGCTTTTAACTCTGGTAAAGTGGATATTGTTGCCATCAATGACCCCTTCATTGACCTCAACTACATGGTTTACATGTTCCAATATGATTCCACCCATGGG¶

CAAATTCCATGGCACCGTCAAGGCTGAGAACGGGAAGCTTGTCATCAATGGAAATCCCATCACCATCTTCCAGGAGCGAGATCCCTCCAAAATCAAGTGGGGCGATGCTGGCGCTGAG¶

TACGTCGTGGAGTCCACTGGCGTCTTCACCACCATGGAGAAGGCTGGGGCTCATTTGCAGGGGGGAGCCAAAAGGGTCATCATCTCTGCCCCCTCTGCTGATGCCCCCATGTTCGTCA¶

TGGGTGTGAACCATGAGAAGTATGACAACAGCCTCAAGATCATCAGCAATGCCTCCTGCACCACCAACTGCTTAGCACCCCTGGCCAAGGTCATCCATGACAACTTTGGTATCGTGGAA¶

AGGACTCATGACCACAGTCCATGCCATCACTGCCACCCAGAAGACTGTGGATGGCCCCTCCGGGAAACTGTGGCGTGATGGCCGCGGGGCTCTCCAGAACATCATCCCTGCCTCTACT¶

GGCGCTGCCAAGGCTGTGGGCAAGGTCATCCCTGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCCGTGTCCCCACTGCCAACGTGTCAGTGGTGGACCTGACCTGCCGTCTAGAAA¶

AACCTGCCAAATATGATGACATCAAGAAGGTGGTGAAGCAGGCGTCCGAGGGCCCCCTCAAGGGCATCCTGGGCTACACTGAGCACCAGGTGGTCTCCTCTGACTTCAACAGCGACACC¶

CACTCCTCCACCTTTGACGCTGGGGCTGGCATTGCCCTCAACGACCACTTTGTCAAGCTCATTTCCTGGTATGACAACGAATTTGGCTACAGCAACAGGGTGGTGGACCTCATGGCC¶

CACATGGCCTCCAAGGAGTAAGACCCCTGGACCACCAGCCCCAGCAAGAGCACAAGAGGAAGAGAGAGACCCTCACTGCTGGGGAGTCCCTGCCACACTCAGTCCCCCACCACACTGA¶

ATGTCCCCTCCTCACAGTTGCCATGTAGACCCCTTGAAGAGGGGAGGGGCCTAGGGAGCCGCACCTTGTCATGTACCATCAATAAAGTACCCTGTGCTCAACC¶

Figure 12a

```
5' TCCCCTCGTGTCAGCTCAGG 3' 1505-1524
5' CTCATGGGTCACTGAGGCTTTTTAT 3' 1788-1812
Amplification product length : 308    1505-1812
GC% : 56,0%
Suggested hybridization temperature(TA) : 58,1
Primer strand + :
Primer strand - :
Primer :  :     5' TCCCCTCGTGTCAGCTCAGG 3'
                   ||||||||||||||||||||
Target : 1505 : 5' TCCCCTCGTGTCAGCTCAGG 3'
score : 174             TM : 74,4 / 52,8 / 60,
Primer :  :     3' ATAAAAAGCCTCAGTGACCCATGAG 5' (complement)
                   |||||||||||||||||||||||||
Target : 1788 : 5' ATAAAAAGCCTCAGTGACCCATGAG 3'
score : 181             TM : 72,5 / 50,9 / 57,1
```

>gi|16507965|ref|NM_001428.2| Homo sapiens enolase 1, (alpha) (ENO1), mRNA¶

¶

TAGCTAGGCAGGAAGTCGGCGCGGGCGGCGCGGACAGTATCTGTGGGTACCCGGAGCACGGAGATCTCGCCGGCTTTACGTTCACCTCGGTGTCTGCAGCACCCTCCGCTTCCTCTCCTA
GGCGACGAGACCCAGTGGCTAGAAGTTCACCATGTCTATTCTCAAGATCCATGCCAGGGAGATCTTTGACTCTCGCGGGAATCCCACTGTTGAGGTTGATCTCTTCACCTCAAAGGTCT
CTTCAGAGCTGCTGTGCCCAGTGGTGCTTCAACTGGTATCTATGAGGCCCTAGAGCTCCGGGACAATGATAAGACTCGCTATATGGCGAAGGGTGTCTCAAAGGCTGTTGAGCACATCAA
TAAAACTATTGCGCCTGCCCTGGTTAGCAAGAAACTGAACGTCACAGAACAAGAGAAGATTGACAAACTGATGATCGAGATGGATGGAACAGAAAATAAATCTAAGTTTGGTGCGAACGC
CATTCTGGGGGGTGTCCCTTGCCGTCTGCAAAGCTGGTGCCGTTGAGAAGGGGGTCCCCCTGTACCGCCACATCGCTGACTTGGCTGGCAACTCTGAAGTCATCCTGCCAGTCCCGGCGTT
CAATGTCATCAATGGCGGTTCTCATGCTGGCAACAAGCTGGCCATGCAGGAGTTCATGATCCTCCCAGTCGGTGCAGCAAAGTTCAGGGAAGCCATCCGCATTGGAGCAGAGGTTTACCA
CAACCTGAAGAATGTCATCAAGGAGAAATATGGGAAAGATGCCACCAATGTGGGGGATGAAGGCGGGTTTGCTCCCAACATCCTGGAGAATAAAGAGGCCTGGAGCTGCTGAAGACTGC
TATTGGGAAAGCTGGCTACACTGATAAGGTGGTCATCGGCATGGACGTAGCGGCCTCCGAGTTCTTCAGGTCTGGGAAGTATGACCTGGACTTCAAGTCTCCCGATGACCCCAGCAGGTA
CATCTCGCCTGACCAGCTGGCTGACCTGTACAAGTCCTTCATCAAGGACTACCCAGTGGTGTCTATCGAAGATCCCTTTGACCAGGATGACTGGGGAGCTTGGCAGAAGTTCACAGCCAG
TGCAGGAATCCAGGTAGTGGGGGATGATCTCACAGTGACCAACCCAAAGAGGATCGCCAAGGCCGTGAACGAGAAGTCCTGCAACTGCCTCCTGCTCAAAGTCAACCAGATTGGCTCCGT
GACCGAGTCTCTTCAGGCGTGCAAGCTGGCCCAGGCCAATGGTTGGGGCGTCATGGTGTCTCATCGTTCGGGGGAGACTGAAGATACCTTCATCGCTGACCTGGTTGTGGGGCTGTGCAC
TGGGCAGATCAAGACTGGTGCCCCTTGCCGATCTGAGCGCTTGGCCAAGTACAACCAGCTCCTCAGAATTGAAGAGGAGCTGGGCAGCAAGGCTAAGTTTGCCGGCAGGAACTTCAGAAA
CCCCTTGGCCAAGTAAGCTGTGGGCAGGCAAGCCCTTCGGTCACCTGTTGGCTACACAGACCCCTCCCCTCGTCAGCTCAGGCAGCTCGAGGCCCCTGACCAACTCTTGCAGGGTTC
CTGCTAGTTAACGGCCGACGGCGTGGAGTTCGTACCGCTTGCTTAGAACTTCTACAGAAGCAAGCTCCCTGGAGCCCTGTTGGCAGCTCTAGCTTTGCAGTCGTCTAATTGGGCCAAG
TCATTGTTTTCTGGCCTCAGTTTCCAGCAAGTGTCTACAGCATCTTAGCCTCGTGTCATGTCGGGTGGCCACAGGCTAGATCCCCGTGGTTTTGTGTCAATATAAAAGCCTCA
GTGACCCATGAG¶

Figure 12b

```
5' GGATCAAGTTTAAATGACTGTGCTG 3'    311-335
5' TTTTATTTGCATATTAAAAAAATTGTGC 3'    590-614
Amplification product length : 304
GC% : 42,0%
Suggested hybridization temperature(TA) : 53,3
Primer strand + :
Primer strand - :
Primer :  :    5' GGATCAAGTTTAAATGACTGTGCTG 3'
                  |||||||||||||||||||||||||
Target : 311 : 5' GGATCAAGTTTAAATGACTGTGCTG 3'
score : 181
TM : 70,9 / 48,3 / 55,
Primer :  :    3' GCACAATTTTTTTAATATGCAAATAAAA 5' (complement)
                  ||||||||||||||||||||||||||||
Target : 587 : 5' GCACAATTTTTTTAATATGCAAATAAAA 3'
score : 184
TM : 64,4 / 42,8 / 50,5
```

>gi|34328943|ref|NM_021109.2| Homo sapiens thymosin, beta 4, X-linked (TMSB4X), mRNA¶

¶

ACAACTCGGTGGTGGCCACTGCGCAGACCAGACTTCGCTCGTACTCGTGCGCCTCGCTTCGCTTTTCCTCCGCAACCATGTCTGACAAACCCGATATGGCTGAGATCGAGAAATTCGATA
AGTCGAAACTGAAGAAGACAGAGACGCAAGAGAAAAATCCACTGCCTTCCAAAGAAACGATTGAACAGGAGAAGCAAGCAGGCGAATCGTAATGAGGCGTGCGCCGCCAATATGCACTGT
ACATTCCACAAGCATTGCCTTCTTATTTTACTTCTTTTAGCTGTTTAACTTTGTAAGATGCAAAGAGCTTGCATCAACTTTAAATGACTGTGCTGCCCCTTTCACATCAAAGAACTACTG
ACAACGAAGGCCGCTCTTCCTTTCCTCTCGTCTATCTATCTGCCTCGCAGGGAAGCAAAGAACTTGCACTGTTGGTCAAAGCAGAGACTGGCTGGAAGCAATGTAGGTGCACGACAACTG
AAATCTAGAGTAAACTAAGCTGGCCTAAGTGTCCTCAGGCTCTAATGCAGCTTAATTGAGTGCCATTTTTTTTTTCTTCAAATGATTTAATTATTGGAATGCACAATTTTTTTA
ATATGCAAATAAAAAAGTTTAAAAACTT¶

Figure 12c

Symbol	Name	Expasy	Plasma Proteome database	
		protein		mRNA
		SP accession	PPD accession	PPD accession
A2M	alpha-2-macroglobulin	P01023	NP_000005	NM_000014
AHSG	alpha-2-HS-glycoprotein	P02765	NP_001613	NM_001622
ALB	albumin	P02768	AAH39235	BC039235
APCS	amyloid P component	P02743	NP_001630	NM_001639
APOA1	apolipoprotein A-I	P02647	NP_000030	NM_000039
APOA2	apolipoprotein A-II	P02652	NP_001634	NM_001643
APOC1	apolipoprotein C-I	P02654	NP_001636	NM_001645
APOC2	apolipoprotein C-II	P02655	NP_000463	NM_000474
APOC3	apolipoprotein C-III	P02656	NP_000031	NM_000040
APOD	apolipoprotein D	P05090	NP_001638	NM_001647
APOE	apolipoprotein E	P02649	NP_000032	NM_000041
AZGP1	Zn-alpha2-glycoprotein	Q5XKQ4	CAA42438	X59766
B2M	beta-2-microglobulin	P61769	NP_004039	NM_004048
C1S	complement component 1, s subcomponent	P09871	NP_958850	NM_201442
C3	complement component 3	Q6LDJ0	NP_000055	NM_000064
C7	complement component 7	P10643	NP_000578	NM_000587
CDH1	E-cadherin	P12830	BAA86957	AB025106
CFB	complement factor B	P00751	NP_001701	NM_001710
CHI3L1	chitinase 3-like 1	P36222	NP_001267	NM_001276

Figure 13a

		Expasy	Plasma Proteome database	
		proteins		mRNA
Symbol	Name	SP accession	PPD accession	PPD accession
CLEC3B	C-type lectin domain family 3, member B	P05452	NP_003269	NM_003278
CLU	clusterin	P10909	NP_001822	NM_001831
CP	ceruloplasmin	P00450	NP_000087	NM_000096
CRP	C-reactive protein	P02741	AAH20766	BC020766
EDN1	endothelin 1	P05305	NP_001946	NM_001955
EGFR	epidermal growth factor receptor	P00533	NP_958441	NM_201284
F13A1	coagulation factor XIII, A1 polypeptide	P00488	NP_000120	NM_000129
F13B	coagulation factor XIII, B polypeptide	P05160	NP_001985	NM_001994
FGA	fibrinogen alpha chain	P02671	NP_000499	NM_000508
FGB	fibrinogen beta chain	Q32P65	NP_005132	NM_005141
FGG	fibrinogen gamma chain	P02679	NP_068656	NM_021870
GPI	glucose phosphate isomerase	P06744	NP_000166	NM_000175
GSTP1	glutathione S-transferase pi	P09211	NP_000843	NM_000852
HDLBP	cDNA FLJ45936 fis,highly similar to Vigilin	Q00341	BAC87153	AK127833
HP	haptoglobin	P00738	NP_005134	NM_005143
HPX	hemopexin	P02790	NP_000604	NM_000613
HRG	histidine-rich glycoprotein	P04196	NP_000403	NM_000412
IGFBP3	insulin-like growth factor binding protein 3	P17936	NP_000589	NM_000598
IGJ	immunoglobulin J polypeptide	P01591	NP_653247	NM_144646

Figure 13b

Symbol	Name	Expasy	Plasma Proteome database	
		proteins	proteins	mRNA
		SP accession	PPD accession	PPD accession
INHA	inhibin, alpha	P05111	NP_002182	NM_002191
KLK11	kallikrein-related peptidase 11	Q9UBX7	NP_659196	NM_144947
LDHA	lactate dehydrogenase A	P00338	NP_005557	NM_005566
LGALS3BP	lectin, galactoside-binding, soluble, 3 binding protein	Q08380	NP_005558	NM_005567
LRG1	leucine-rich alpha-2-glycoprotein 1	P02750	NP_443204	NM_052972
ORM1	Homo sapiens orosomucoid 1	P02763	NP_000598	NM_000607
PKM2	pyruvate kinase, muscle	P14618	NP_872270	NM_182470
PLG	plasminogen	P00747	NP_000292	NM_000301
PON1	paraoxonase 1	P27169	NP_000437	NM_000446
PROC	protein C	P04070	NP_000303	NM_000312
PZP	pregnancy-zone protein	P20742	NP_002855	NM_002864
RBP4	retinol binding protein 4	P02753	P02753	BC020633
SAA1	serum amyloid A1	P02735	NP_000322	NM_000331
SERPINA1	serpin peptidase inhibitor, clade A	P01009	NP_000286	NM_000295
SERPINA3	serpin peptidase inhibitor, clade A, member 3	P01011	NP_001076	NM_001085
SERPINA5	serpin peptidase inhibitor, clade B, member 5	P05154	NP_000965	NM_002974
TF	transferrin	P02787	NP_001054	NM_001063
TFRC	transferrin receptor	P02786	NP_003225	NM_003234
TGFB1	transforming growth factor, beta 1	P01137	NP_000651	NM_000660

Figure 13c

Symbol	Name	Expasy	Plasma Proteome database	
		proteins	proteins	mRNA
		SP accession	PPD accession	PPD accession
TIMP1	TIMP metallopeptidase inhibitor 1	P01033	NP_003245	NM_003254
TTR	transthyretin	P02766	NP_000362	NM_000371
VTN	vitronectin	P04004	NP_000629	NM_000638

Figure 13d

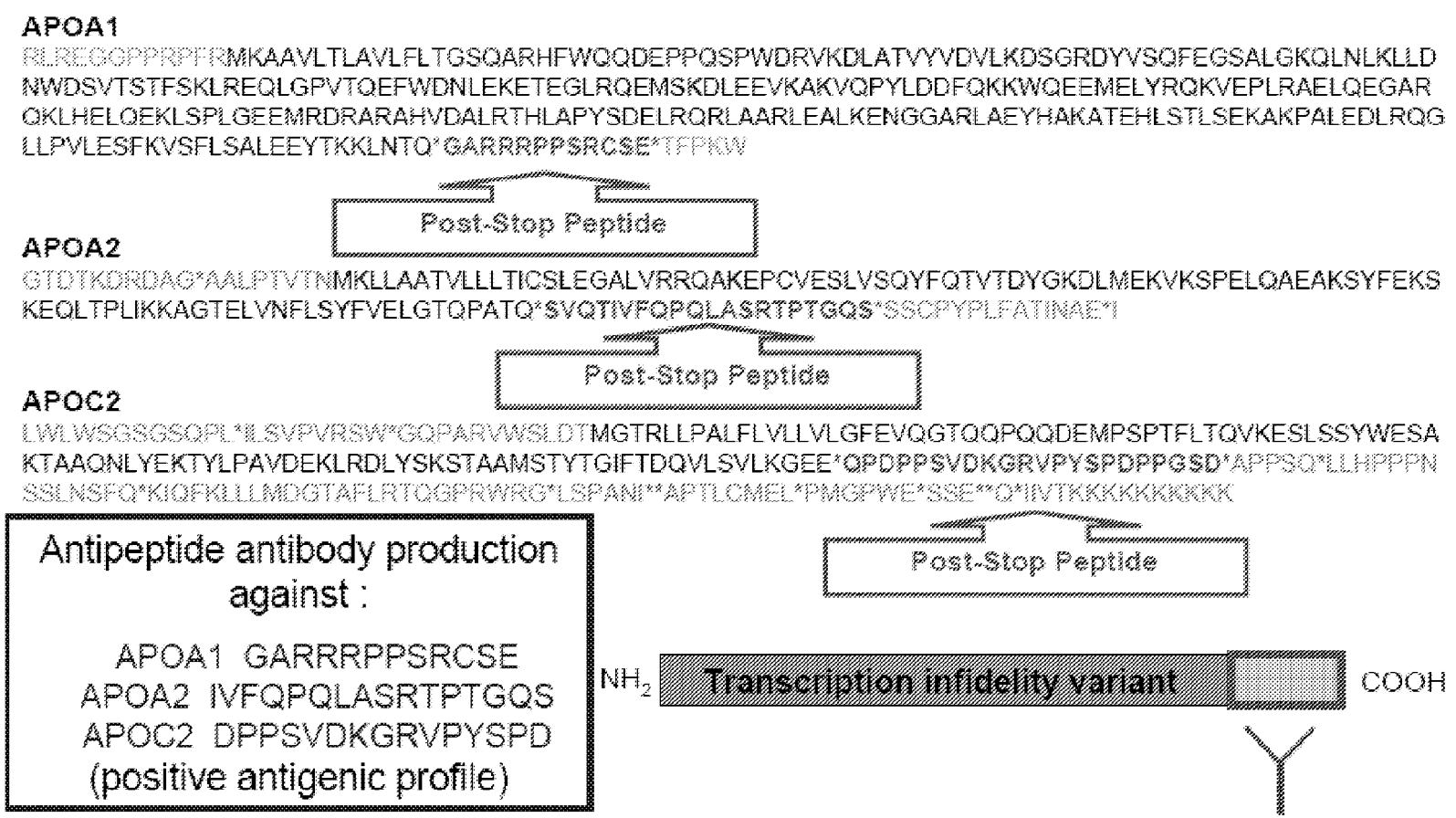
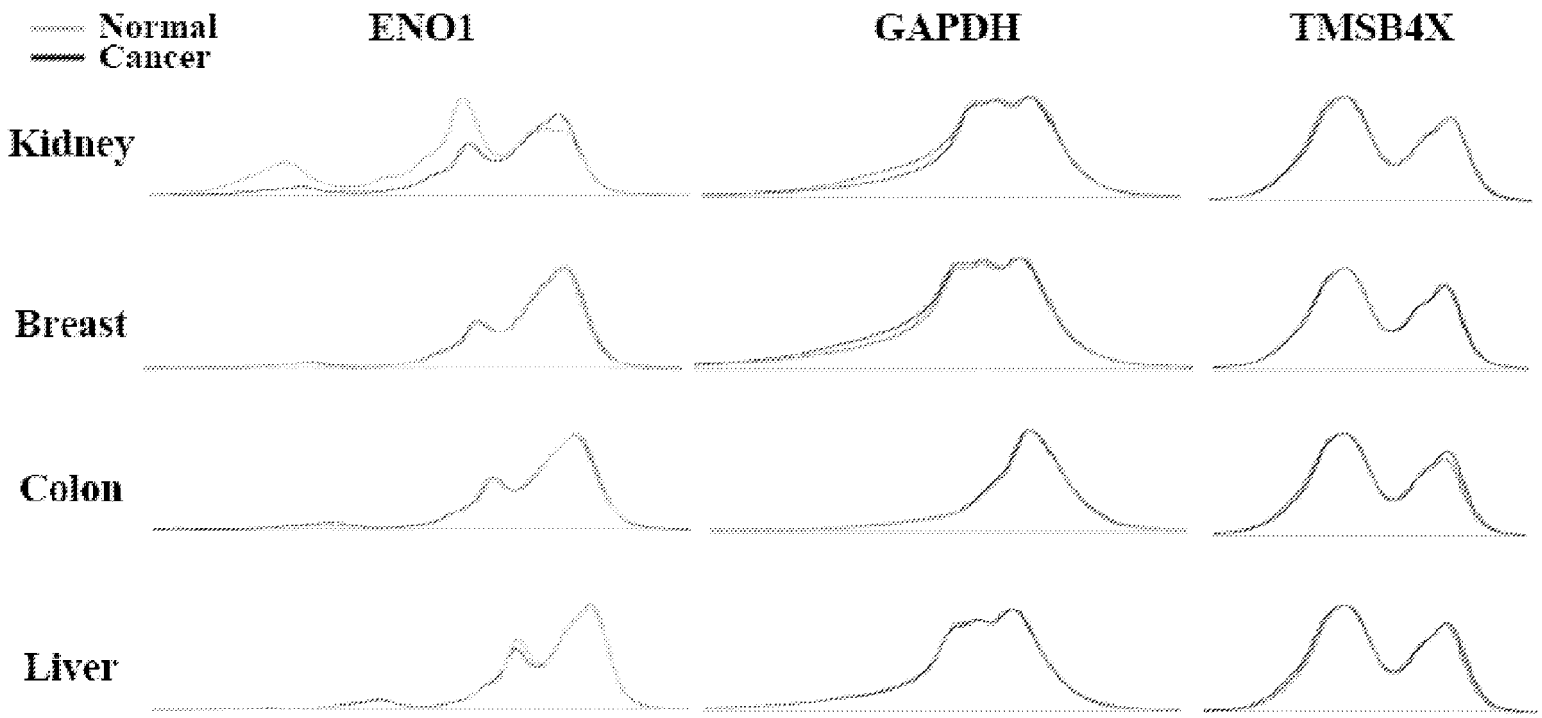

Selection of 3 plasma proteins :
APOA1
RLREGQPPRPFRMKAAVLTLAVLFLTGSQARHFWQQDEPPQSPWDRVKDLATVYVDVLKDSGRDYVSQFEGSALGKQLNLKLLD
NWDSVTSTFSKLREQLGPVTQEFWDNLEKETEGLRQEMSKDLEEVKAKVQPYLDDFQKKWQEEMELYRQKVEPLRAELQEGAR
QKLHELQEKLSPLGEEMRDRARAHVDALRTHLAPYSDELRQRLAARLEALKENGGARLAEYHAKATEHLSTLSEKAKPALEDLRQG
LLPVLESFKVSFLSALEEYTKKLNTQ*GARRRPPSRCSE*TFPKW
Post-Stop Peptide
APOA2
GTDTKDRDAG*AALPTVTNMKLLAATVLLLTICSLEGALVRRQAKEPCVESLVSQYFQTVTDYGKDLMEKVKSPELQAEAKSYFEKS
KEQLTPLIKKAGTELVNFLSYFVELGTQPATQ*SVQTIVFQPQLASRTPTGQS*SSCPYPLFATINAE*I
Post-Stop Peptide
APOC2
LWLWSGSGSQPL*ILSVPVRSW*GQPARVWSLDTMGTRLLPALFLVLLVLGFEVQGTQQPQQDEMPSPTFLTQVKESLSSYWESA
KTAAQNLYEKTYLPAVDEKLRDLYSKSTAAMSTYTGIFTDQVLSVLKGEE*QPDPPSVDKGRVPYSPDPPGSD*APPSQ*LLHPPPN
SSLNSPQ*KIQFKLLLMDGTAFLRTQGPRWRG*LSPANI**APTLQMEL*PMGPWE*SSE**Q*IVTKKKKKKKKKK
Post-Stop Peptide
Antipeptide antibody production against :
APOA1 GARRRPPSRCSE
APOA2 IVFQPQLASRTPTGQS
APOC2 DPPSVDKGRVPYSPD
(positive antigenic profile)
NH2
Transcription infidelity variant
COOH

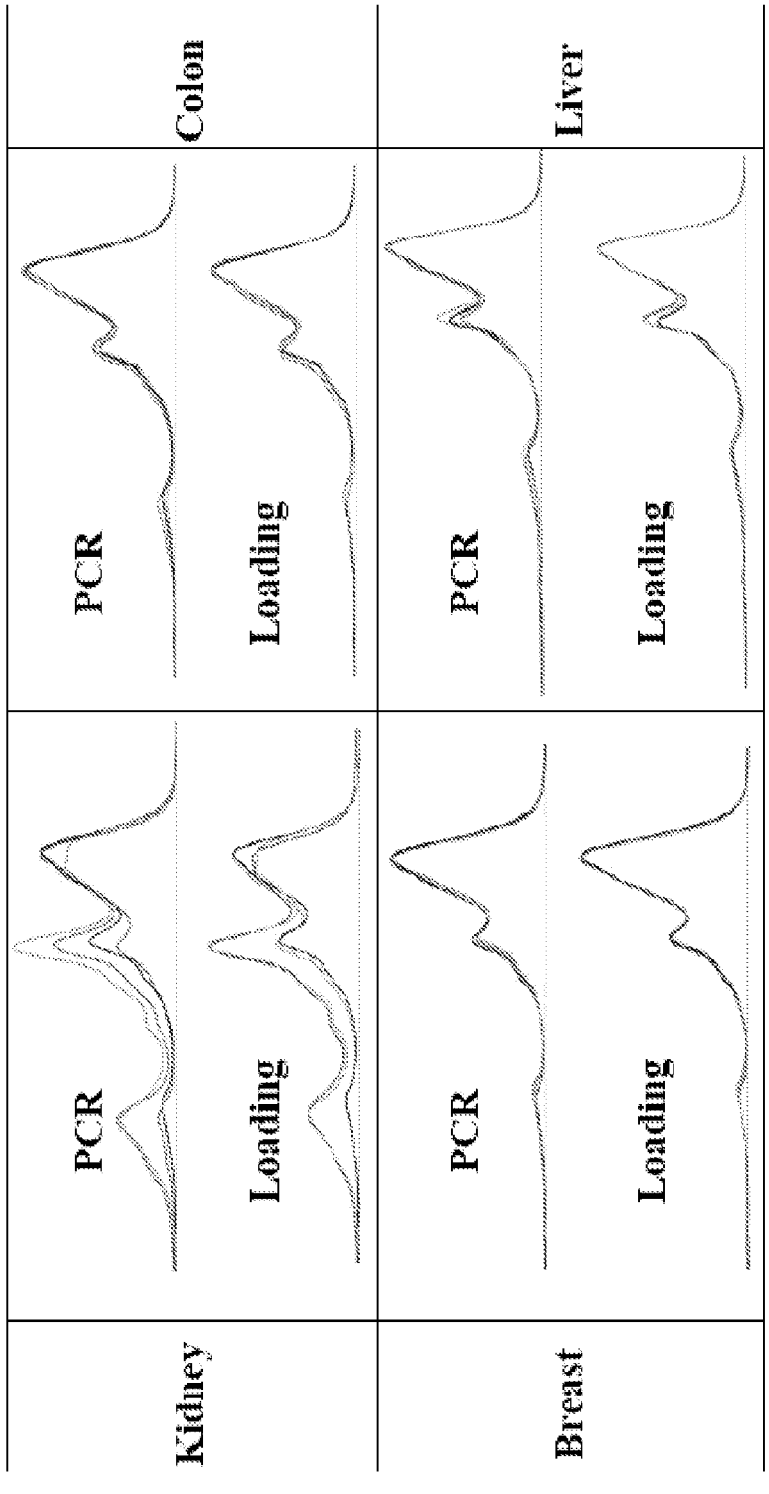
Figure 14



DHPLC profiles of tumor tissue versus normal adjacent tissue for the ENO1, GAPDH and TMSB4X genes.

cDNA from cancerous patients (Biochain Inc.) were amplified by PCR using oligonucucleotides complementary to the ENO1, GAPDH and TMSB4X genes and the *pfx* polymerase. cDNA were obtained from Kidney, Breast, colon and Liver tissues. Normal DHPLC profile is represented in Green and the cancerous profile is represented in Black. The double stranded DNA samples were injected on a DHPLC system (Transgenomic). The temperature of the oven are 62.5°C for the GAPDH gene, 61.5°C for the ENO1 gene and 55°C for the TMSB4X gene. Curves were visualized and normalized with the Navigator software (Transgenomic).

Figure 15a



Different PCR and Loading DHPLC profiles of tumor tissue versus normal adjacent tissue for the ENO1 gene.

Two different PCR products were prepared for the 4 tissues and each PCR product was injected twice on the DHPLC system. Normal DHPLC profile is represented in Green and the cancerous profile is represented in Black. The DHPLC profile obtained for the PCR duplicate is represented in : Red for Normal and Blue for cancerous. For the loading duplicate, profiles are represented in : Purple for Normal and Light Blue for Cancerous. Curves were visualized and normalized with the Navigator software (Transgenomic).

Figure 15b

Different PCR and Loading DHPLC profiles of tumor tissue versus normal adjacent tissue for the GAPDH gene.

Two different PCR products were prepared for the 4 tissues and each PCR product was injected twice on the DHPLC system. Normal DHPLC profile is represented in Green and the cancerous profile is represented in Black. The DHPLC profile obtained for the PCR duplicate is represented in : Red for Normal and Blue for cancerous. For the loading duplicate, profiles are represented in : Purple for Normal and Light Blue for Cancerous. Curves were visualized and normalized with the Navigator software (Transgenomic).

Figure 15c

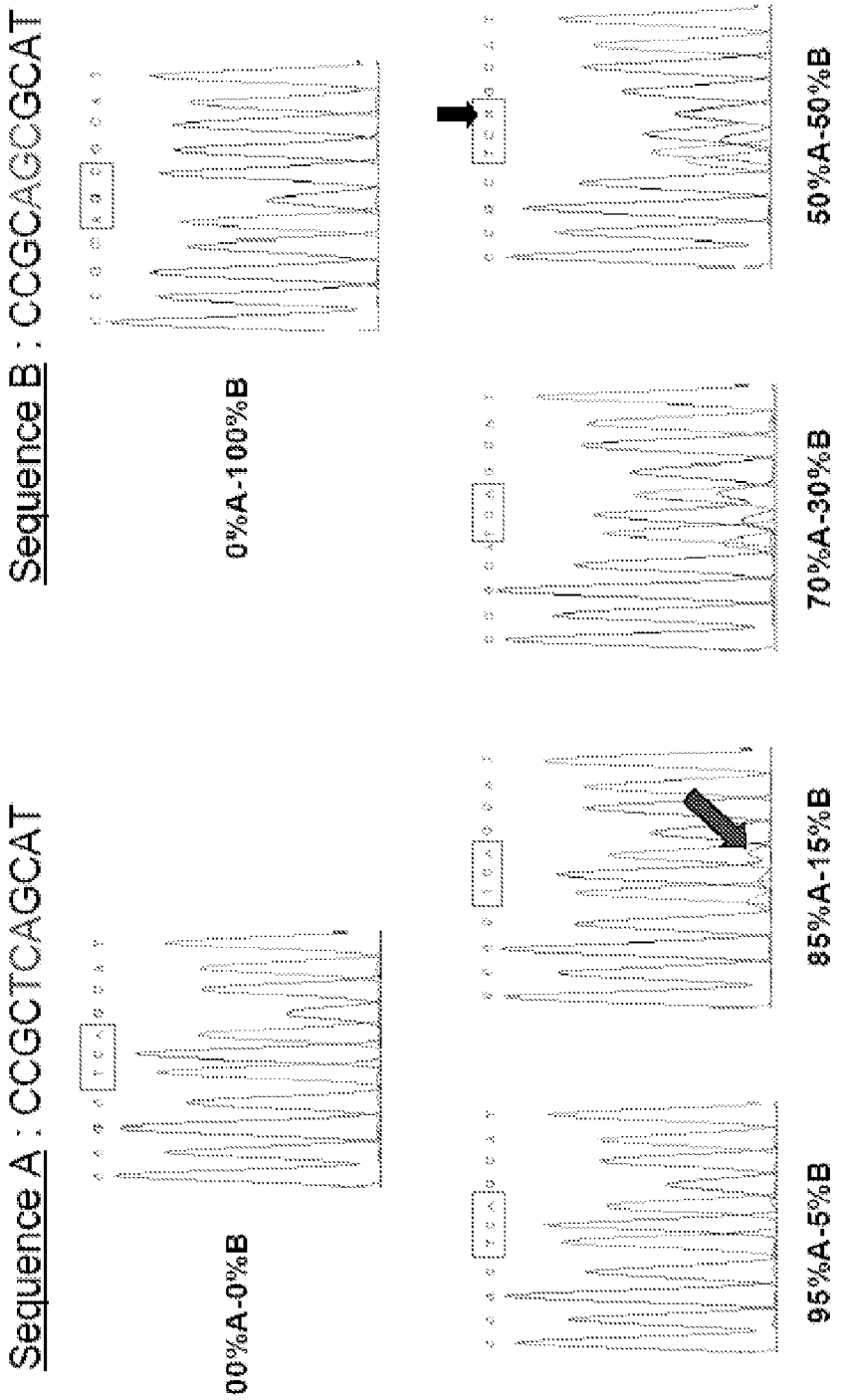

Sequence A : CCGCTCAGCAT
Sequence B : CCGCAGCGCAT
100%A-0%B
95%A-5%B
85%A-15%B
70%A-30%B
50%A-50%B
0%A-100%B

Figure 16

Figure 17a

Figure 17b

Plasma from patients with prostate cancer (a) or with cancer in the metastatic phase (b) were applied to 4-12% SDS polyacrylamide gels under reducing conditions. After transfer onto PVDF membranes, and blocking in TBST containing 5% nonfat dry milk (1h, room temperature), Westerns were performed using rabbit pre-immune serum (dilution 1/100, left panel), commercially available goat anti-human apoAII antibody (1/250, middle panel), and rabbit polyclonal anti-PSP apoAII antibody (1/100, right panel). After 1 h incubation at room temperature, the membranes were washed and then incubated with the appropriate conjugated second antibody. Bands were then revealed by chemiluminescence and exposure to X-ray films.

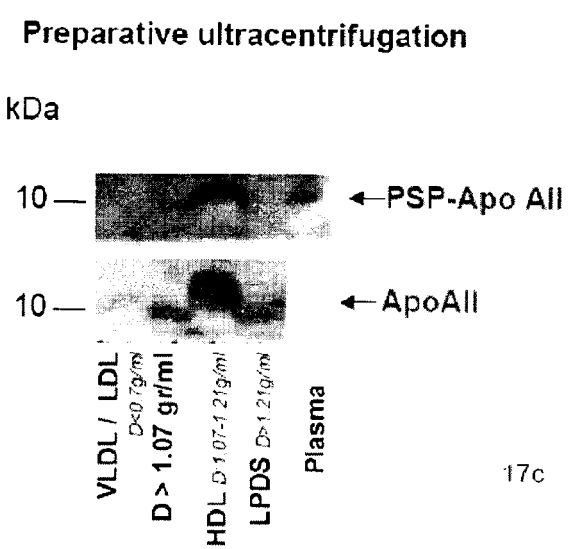


Figure 17c

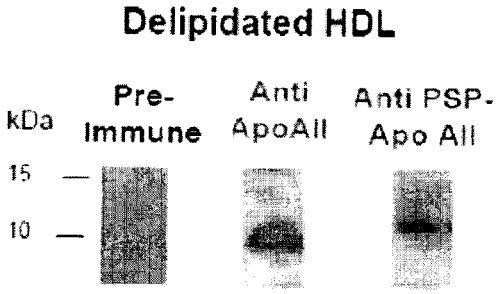


Figure 17d

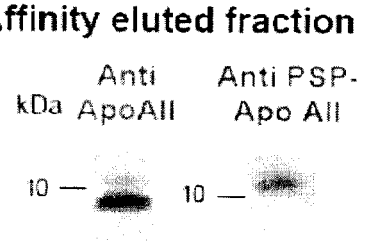


Figure 17e

c. HDL (d1.07-1.21 g/ml) was prepared by sequential ultracentrifugation of pooled plasma from 10 cancer patients in the metastatic phase. The different fractions recovered after each step were dialyzed against 0.15 M NaCl, pH 7.4 containing 0.01% EDTA. Western blots were performed on the indicated fractions with either commercial anti-apoAII antibody (lower panel) or anti-PSP apoAII antibody (upper panel) using the same dilutions as indicated in Figures 17a et 17b. Gels were overloaded, which explains the distortion and the large bands. Western blot was also performed on the original pooled plasma or a corresponding volume of the lipoprotein-deficient serum (d>1.21 g/ml). **d.** Purified HDL was dialyzed against 0.01%EDTA, pH 7.4, frozen at -80°C and then lyophilized. Delipidation on the lyophilized HDL was then performed using ice-cold chloroform-methanol (2:1, v/v; 5 ml per 20 mg HDL protein). A second delipidation using diethyl ether was then followed by centrifugation (1000 rpm, 5 min, 4°C) to pellet the delipidated protein. This was repeated, and the final pellet dried under N2(g). The pellet was redissolved in 10 mM Tris containing 6 M urea, pH 8. Western blotting was performed on 10 ug delipidated HDL using preimmune serum, commercial anti-apoAII, or anti-PSP apoAII as described above. **e.** Affinity chromatography experiments were carried out to isolate the PSP form of apoAII using the anti-PSP antibody immobilized on matrix beads. The eluted fraction is analysed by Western blotting using both the commercial anti-apoAII and anti-PSPapoAII antibodies.

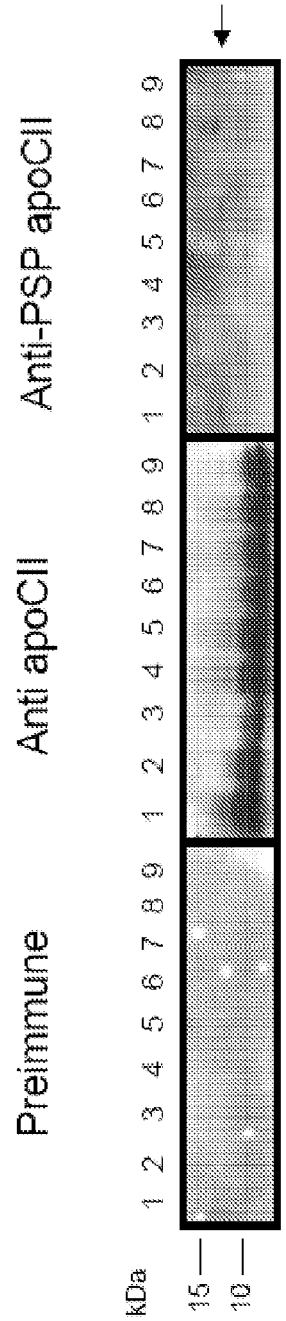


Figure 17f

Similar experiment was performed as for Fig a. with the exception that rabbit polyclonal anti-PSP apoCII antibody was used.

Name	mRNA id	PSP	Decision
AHSG	NM_001622.1	ARHGRDEEVYHFHVDAWAVYSSLVGWPAVCHMRSTLISLESLLPVPHRTEAEWVYKMPDRRH	YES
ALB	BC039235	RSKAYSSVFLFRVDCAKTLSKVHKFL	YES
APCS	NM_001639.2	GLDSTRALENEMTV	YES
APOA1	NM_000039.1	GARRSPPSKCSE	YES
APOA2	NM_001643.1	SVQTIVFGPQLASRTFTGQS	YES
APOC2	NM_000483.3	QPDFPSVDKGRVPYSPFDPASD	YES
APOC3	NM_000040.1	DLNTPSPPAYPSDELLGSGNLQGQPCRLLDRDSLSALLPHLMPGPFPGKLASQ	YES
APOD	NM_001647.2	PGSTGRLHPLHVTSASLSPTPPPPHKDKPWHDKGS	YES
APOE	NM_000041.2	TPWPAAMRPHATPCLLPPRSLQRETLSPPQPSSWGGP	YES
AZGP1	X56786.1	EARVGGNWSSGTQ	YES
CHI3L1	NM_001276.3	PSVLHTARGPRMPRPFLAPAGREPDHLPC	YES
CLU	NM_001831.2	DVDVAFAPTGASESSGPQDELGPPRESSARHQVTRPQPGPQLRPASPRSGSETLTLDSAAHGKNRIAPACN	YES
HRG	NM_000412.2	NVFLARKMNYTLN	YES
IGFBP3	NM_000598.4	TPAAFLMWSKMPYFAGNTAKDMTSSWLQPRFPLFVVH	YES
ITHA	NM_002181.2	GWGVFLLMPNAGGHAPTIIEWEERQSWEDGSHSSLLGLLCLWATLPTPLLSQ	YES
KLK11	NM_144947.3	TGPTHHSPSPSSTWCLVPVHSVWMKP	YES
PKM2	NM_182470.1	WTPEPLLQPLSHPLPPAHPLGQQPL	YES
PLG	NM_000301.1	LDGRQSDALTHLEAGTWVGI	YES
SERPINA3	NM_001085.4	SLPSSSGALSKELGMMAAGCLGLWAGPGPCAPSGHGMGGPWKLSLEGDSDSLCSSHMHRGPWTLGSGSSWAS	YES
TF	NM_001063.2	NLRGRAATKVKMGTQMHEFALVSLAQVVCANMVCLHGSVLPCVLNKW	YES
TGFB1	NM_000660.3	GPAPPRPAPAGPAPPRPAPAALPMGAVPKDTSAPSPPGAPLKMERGLRSYSLGAKCLGSPSLTFPHSHSLSLPLGLL PVCTRPLPGHAAQGTSSEHYCS	YES
TTR	NM_000371.1	GTSPPVDLKDEGWDFM	YES

Figure 18a

>P02765|AHSG Alpha-2-HS-glycoprotein
>P02768| Albumin
>P02743|APCS Serum amyloid P-component
>P02647|APOA1 Apolipoprotein A-I
>P02652|APOA2 Apolipoprotein A-II
>P02655|APOC2 Apolipoprotein C-II
>P02656|APOC3 Apolipoprotein C-III
>P05090|APOD Apolipoprotein D
normal proteins
additional peptide of proteins whose STOP codon is affected

Figure 18b

>P02649|APOE Apolipoprotein E

MKVLWAALLVTFLAGCQAKVEQAVETEPEPELRQQTEWQSGQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELRALMDETMKELKAY
KSELEEQLTPVAEETRARLSKELQAAQARLGADMEDVCGRLVQYRGEVQAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQKRLAVYQAGAR
EGAERGLSAIRERLGPLVEQGRVRAATVGSLAGQPLQERAQAWGERLRARMEEMGSRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQAR
LKSWFEPLVEDMQRQWAGLVEKVQAAVGTSAAPVPSDNH*TPKPAAMRPHATPCLLFFRSLQRETLSPPQPSSWGGP*

>Q5XKQ4|AZGP1 Alpha-2-glycoprotein 1, zinc-binding

MWASMSRMLPVLLSLLLLLGPAVPQEWQDGRYSLTYYYTGLSKHVEDVPAFQALGSLNDLQFFRYNSKDRKSQPMGLWRQVEGMEDWKQDSQ
LQKAREDIFMETLKDIVEYYNDSNGSHVLQGRFGCEIENNRSSGAFWKYYYDGKDYIEFNKEIPAWVPFDPAAQITKQKWEAEPVYVQRAKAYLE
EECPATLRKYLKYSKNILDRQDPPSVVVTSHQAPGEKKKLKCLAYDFYPGKIDVHWTRAGEVQEPELRGDVLHNGNGTYQSWVVVAVPPQDTAP
YSCHVQHSSLAQPLVVPWEAS*EARVGGNCGGTQ*

>P36222|CHI3L1 Chitinase-3-like protein 1

MGVKASQTGFVVLVLLQCCSAYKLVCYYTSWSQYREGDGSCFPDALDRFLCTHIIYSFANISNDHIDTWEWNDVTLYGMLNTLKNRNPNLKTLLS
VGGWNFGSQRFSKIASNTQSRRTFIKSVPPFLRTHGFDGLDLAWLYPGRRDKQHFTTLIKEMKAEFIKEAQPGKKQLLLSAALSAGKVTIDSSYDIA
KISQHLDFISIMTYDFHGAWRGTTGHHSPLFRGQEDASPDRFSNTDYAVGYMLRLGAPASKLVMGIPTFGRSFTLASSETGVGAPISGPGIPGRFT
KEAGTLAYYEICDFLRGATVHRTLGQQVPYATKGNQWVGYDDQESVKSKVQYLKDRQLAGAMVWALDLDDFQGSFCGQDLRFPLTNAIKDALA
AT*PSVLHTAAREPTNMFTPPLAPAGREPDHLPC*

>P10909|CLU Clusterin

MQVCSQPQRGCVREQSAINTAPPSAHNAASPGGARGHRVPLTEACKDSRIGGMMKTLLLFVGLLLTWESGQVLGDQTVSDNELQEMSNQGSK
YVNKEIQNAVNGVKQIKTLIEKTNEERKTLLSNLEEAKKKKEDALNETRESETKLKELPGVCNETMMALWEECKPCLKQTCMKFYARVCRSGSGL
VGRQLEEFLNQSSPFYFWMNGDRIDSLLENDRQQTHMLDVMQDHFSRASSIIDELFQDRFFTREPQDTYHYLPFSLPHRRPHFFFPKSRIVRSLM
PFSPYEPLNFHAMFQPFLEMIHEAQQAMDIHFHSPAFQHPPTEFIREGDDDRTVCREIRHNSTGCLRMKDQCDKCREILSVDCSTNNPSQAKLRR
ELDESLQVAERLTRKYNELLKSYQWKMLNTSSLLEQLNEQFNWVSRLANLTQGEDQYYLRVTTVASHTSDSDVPSGVTEVVVKLFDSDPITVTVP
VEVSRKNPKFMETVAEKALQEYRKKHREE*DVDVAFAPTGASESSSPQQELDDPRESSARHQVTRPQFPGPQLRPASPRGSCTLTLDSAMHG
KNRIMPACN*

>P04196|HRG Histidine-rich glycoprotein

MKALIAALLLTLQYSCAVSPTDCSAVEPEAEKALDLINKRRRDGYLFQLLRIADAHLDRVENTTVYYLVLDWQESDCSVLSRKYWNDCEPPDSRRP
SEIVIGQCKVIATRHSHESQDLRVIDFNCTTSSVSSALANTKDSPVLIDFFEDTERYRKQANKALEKYKEENDDFASFRVDRIERVARVRGGEGTGY
FVDFSVRNCPRHHFPRHPWWFGFCRADLFYDVEALDLESPKNLVINCEVFDPQEHENINGVAPHLGHPFHWGGHERSSTTKPPFKPHIGSRDHHH
PHKPHEHGPPPPPDERDHSHGPPLPQGPPPLLPMSCSSCQHATFGTNGAQRHSHNMNSSDLHPHKHHSHEQHPHGHHPHAHHPHEHDTHRQ
HPHGHHPHGHHPHGHHPHGHHPHCHDFQDYGPCDPPPHNQGHCCHGHGPPPGHLRRRGPGKGPRPFHCRQIGSVYRLPPLRKGEV
LPLPEAMFPSFPLPHHKHPLKPDNQPFPQSVSESCPGKFKSGFPQVSMFFTHTFPK*MVIPLARKMMMTLN*

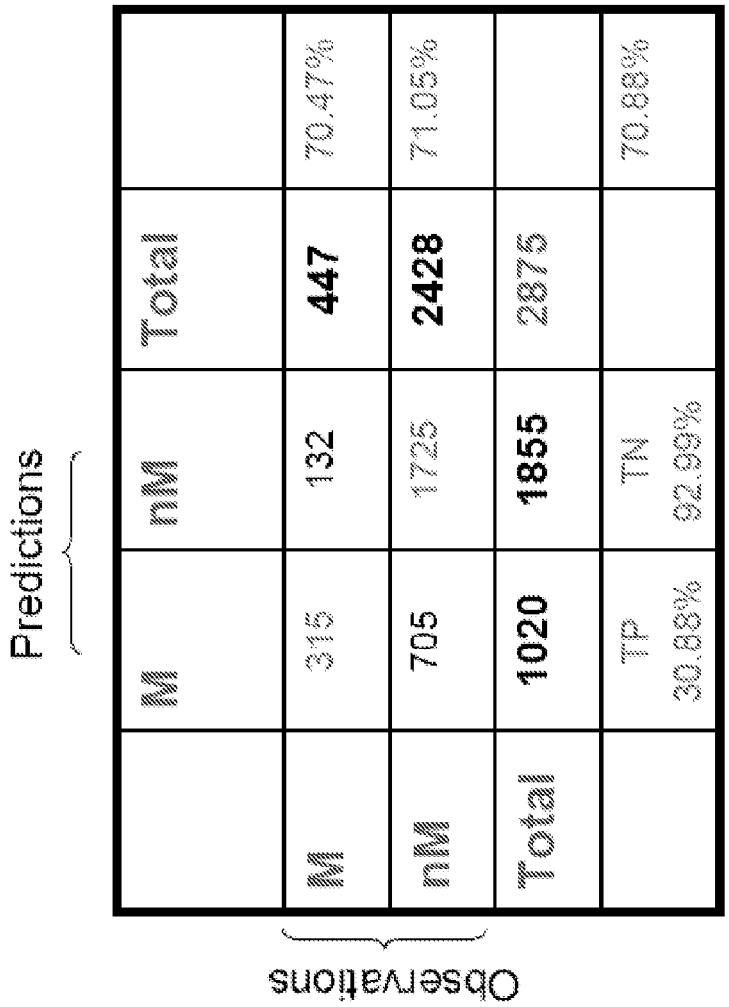
Figure 18c

>P17936|IGFBP3 Insulin-like growth factor-binding protein 3
MQRARPTLWAAALTLLVLLRGPPVARAGASSAGLGPVVRCEPCDARALAQCAPPPAVCAELVREPGCGCCLTCALSEGQPCGIYTERCGSGLR
CQPSPDEARPLQALLDGRGLCVNASAVSRLRAYLLPAPPAPGNASESEEDRSAGSVESPSVSSTHRVSDPKFHPLHSKIIIIKKGHAKDSQRYKVD
YESQSTDTQNFSSESKRETEYGPCRREMEDTLNHLKFLNVLSPRGVHIPNCDKKGFYKKKQCRPSKGRKRGFCWCVDKYGQPLPGYTTKGKED
VHCYSMQSK*TPAARLMWSSNMPYFAQKTAKDMTSSWLQPRFIFLFVVN*
>P05111|INHA Inhibin alpha chain
MVLHLLLFLLLTPQGGHSCQGLELARELVLAKVRALFLDALGPPAVTREGGDPGVRRLPRRHALGGFTHRGSEPEEEEDVSQAILFPATDASCED
KSAARGLAQEAEEGLFRYMFRPSQHTRSRQVTSAQLWFHTGLDRQGTAASNSSEPLLGLLALSPGGPVAVPMSLGHAPPHWAVLHLATSALSLL
THPVLVLLLRCPLCTCSARPEATPFLVAHTRTRPPSGGERARRSTPLMSWPWSPSALRLLQRPPEEPAAHANCHRVALNISFQELGWERWIVYP
PSFIFHYCHGGCGLHIPPNLSLPVPGAPPTPAQPYSLLPGAQPCCAALPGTMRPLHVRTTSDGGYSFKYETVPNLLTQHCACI*GWGVFLLNPMA
GGHAPTIISWEERQSWEIDGSHSSLLSLLCLWATLPTPLLSQ*
>Q9UBX7|KLK11 Kallikrein-11
MQRLRWLRDWKSSGRGLTAAKEPGARSSPLQAMRILQLILLALATGLVGGETRIIKGFECKPHSQPWQAALFEKTRLLCGATLIAPRWLLTAAHCL
KPRYIVHLGQHNLQKEEGCEQTRTATESFPHPGFNNSLPNKDHRNDIMLVKMASPVSITWAVRPLTLSSRCVTAGTSCLISGWGSTSSPQLRLPH
TLRCANITIIEHQKCENAYPGNITDTMVCASVQEGGKDSCQGDSGGPLVCNQSLQGIISWGQDPCAITRKPGVYTKVCKYVDWIQETMKNN*TGP
THHSPSPSISTWCLVPVHSVNKKP*
>P14618|PKM2 Pyruvate kinase isozymes M1/M2
MSKPHSEAGTAFIQTQQLHAAMADTFLEHMCRLDIDSPPITARNTGIICTIGPASRSVETLKEMIKSGMNVARLNFSHGTHEYHAETIKNVRTATES
FASDPILYRPVAVALDTKGPEIRTGLIKGSGTAEVELKKGATLKITLDNAYMEKCDENILWLDYKNICKVVEVGSKIYVDDGLISLQVKQKGADFLVTE
VENGGSLGSKKGVNLPGAAVDLPAVSEKDIQDLKFGVEQDVDMVFASFIRKASDVHEVRKVLGEKGKNIKIISKIENHEGVRRFDEILEASDGIMVA
RGDLGIEIPAEKVFLAQKMMIGRCNRAGKPVICATQMLESMIKKPRPTRAEGSDVANAVLDGADCIMLSGETAKGDYPLEAVRMQHLIAREAEAA
MFHRKLFEELVRASSHSTDLMEAMAMGSVEASYKCLAAALIVLTESGRSAHQVARYRPRAPIIAVTRNPQTARQAHLYRGIFPVLCKDPVQEAWA
EDVDLRVNFAMNVGKARGFFKKGDVVIVLTGWRPGSGFTNTMRVVPVP*WTPEPLLQPLSHPLPPAHPLCQQRL*
>P00747|PLG Plasminogen
MEHKEVVLLLLLFLKSGQGEPLDDYVNTQGASLFSVTKKQLGAGSIEECAAKCEEDEEFTCRAFQYHSKEQQCVIMAENRKSSIIRMRDVVLFEK
KVYLSECKTGNGKNYRGTMSKTKNGITCQKWSSTSPHRPRFSPATHPSEGLEENYCRNPDNDPQGPWCYTTDPEKRYDYCDILECEEECMHC
SGENYDGKISKTMSGLECQAWDSQSPHAHGYIPSKFPNKNLKKNYCRNPDRELRPWCFTTDPNKRWELCDIPRCTTPPPSSGPTYQCLKGTGE
NYRGNVAVTVSGHTCQHWSAQTPHTHNRTPENFPCKNLDENYCRNPDGKRAPWCHTTNSQVRWEYCKIPSCDSSPVSTEQLAPTAPPELTPV
VQDCYHGDGQSYRGTSSTTTTGKKCQSWSSMTPHRHQKTPENYPNAGLTMNYCRNPDADKGPWCFTTDPSVRWEYCNLKKCSGTEASVVAP
PPVVLLPDVETPSEEDCMFGNGKGYRGKRATTVTGTPCQDWAAQEPHRHSIFTPETNPRAGLEKNYCRNPDGDVGGPWCYTTNPRKLYDYCD
VPQCAAPSFDCGKPQVEPKKCPGRVVGGCVAHPHSWPWQVSLRTRFGMHFCGGTLISPEWVLTAAHCLEKSPRPSSYKVILGAHQEVNLEPHV
QEIEVSRLFLEPTRKDIALLKLSSPAVITDKVIPACLPSPNYVVADRTECFITGWGETQGTFGAGLLKEAQLPVIENKVCNRYEFLNGRVQSTELCAG
HLAGGTDSCQGDSGGPLVCFEKDKYILQGVTSWGLGCARPNKPGVYVRVSRFVTWIEGVMRNN*LDGRQSDALTTWVGI*

Figure 18d

>P01011|SERPINA3 Alpha-1-antichymotrypsin
MERMLPLLALGLLAAGFCPAVLCHPNSPLDEENLTQENQDRGTHVDLGLASANVDFAFSLYKQLVLKAPDKNVIFSPLSISTALAFLSLGAHNTTLT
EILKGLKFNLTETSEAEIHQSFQHLLRTLNQSSDELQLSMGNAMFVKEQLSLLDRFTEDAKRLYGSEAFATDFQDSAAAKKLINDYVKNGTRGKITD
LIKDLDSQTMMVLVNYIFFKAKWEMPFDPQDTHQSRFYLSKKKWVMVPMMSLHHLTIPYFRDEELSCTVVELKYTGNASALFILPDQDKMEEVEA
MLLPETLKRWRDSLEFREIGELYLPKFSISRDYNLNDILLQLGIEEAFTSKADLSGITGARNLAVSQVVHKAVLDVFEEGTEASAATAVKITLLSALVE
TRTIVRFNRPFLMIIVPTDTQNIFFMSKVTNPKQA*SLPSSSGALSKELGMQAGCLGLWAQPGPCAPSGHGMCGPVCLSLEGDSDSLCSSHMHR
GPWTLQSGGSWAS*
>P02787|TF Serotransferrin
MRLAVGALLVCAVLGLCLAVPDKTVRWCAVSEHEATKCQSFRDHMKSVIPSDGPSVACVKKASYLDCIRAIAANEADAVTLDAGLVYDAYLAPNN
LKPVVAEFYGSKEDPQTFYYAVAVVKKDSGFQMNQLRGKKSCHTGLGRSAGWNIPIGLLYCDLPEPRKPLEKAVANFFSGSCAPCADGTDFPQL
CQLCPGCGCSTLNQYFGYSGAFKCLKDGAGDVAFVKHSTIFENLANKADRDQYELLCLDNTRKPVDEYKDCHLAQVPSHTVVARSMGGKEDLIW
ELLNQAQEHFGKDKSKEFQLFSSPHGKDLLFKDSAHGFLKVPPRMDAKMYLGYEYVTAIRNLREGTCPEAPTDECKPVKWCALSHHERLKCDE
WSVNSVGKIECVSAETTEDCIAKIMNGEADAMSLDGGFVYIAGKCGLVPVLAENYNKSDNCEDTPEAGYFAVAVVKKSASDLTWDNLKGKKSCH
TAVGRTAGWNIPMGLLYNKINHCRFDEFFSEGCAPGSKKDSSLCKLCMGSGLNLCEPNNKEGYYGYTGAFRCLVEKGDVAFVKHQTVPQNTGG
KNPDPWAKNLNEKDYELLCLDGTRKPVEEYANCHLARAPNHAVVTRKDKEACVHKILRQQQHLFGSNVTDCSGNFCLFRSETKDLLFRDDTVCL
AKLHDRNTYEKYLGEEYVKAVGNLRKCSTSSLLEACTFRRP*NLRGRAATKVKMGTQMIHEFALVSLAQVVCANFVCLHSSVLPCVLNKK*
>P01137|TGFB1 Transforming growth factor beta-1
MPPSGLRLLPLLLPLLWLLVLTPGRPAAGLSTCKTIDMELVKRKRIEAIRGQILSKLRLASPPSQGEVPPGPLPEAVLALYNSTRDRVAGESAEPEP
EPEADYYAKEVTRVLMVETHNEIYDKFKQSTHSIYMFFNTSELREAVPEPVLLSRAELRLLRLKLKVEQHVELYQKYSNNSWRYLSNRLLAPSDSP
EWLSFDVTGVVRQWLSRGGEIEGFRLSAHCSCDSRDNTLQVDINGFTTGRRGDLATIHGMNRPFLLLMATPLERAQHLQSSRHRRALDTNYCFS
STEKNCCVRQLYIDFRKDLGWKWIHEPKGYHANFCLGPCPYIWSLDTQYSKVLALYNQHNPGASAAPCCVPQALEPLPIVYYVGRKPKVEQLSN
MIVRSCKCS*GPAPPRPAPAGPAPPRPAPAALPMGAVFKDTRAPSPPGAPLKMERGLRISVSLGACLGSPSLTFPHSHSLSLPLCLLLPVCTIPLP
GIKAQGTSGEHYCS*
>P02766|TTR Transthyretin
MASHRLLLLCLAGLVFVSEAGPTGTGESKCPLMVKVLDAVRGSPAINVAVHVFRKAADDTWEPFASGKTSESGELHGLTTEEEFVEGIYKVEIDTK
SYWKALGISPFHEHAEVVFTANDSGPRRYTIAALLSPYSYSTTAVVTNPKE*GTSPPVDLKDEGWDFM*

Figure 18e



***TP** : True Positives ; **TN** : True Negatives*

Figure 19