GENE-BASED ALGORITHMIC CANCER PROGNOSIS AND CLINICAL OUTCOME OF A PATIENT

Field of the Invention

[0001] The present invention is related to new method and tools for improving cancer prognosis and clinical outcome of a patient.

Background of the Invention

[0002] Micro-array profiling, or the assessment of the mRNA expression levels of hundreds and thousands of genes, has shown that cancer can be divided into distinct molecular subgroups by the expression levels of certain genes. These subgroups seem to have distinct clinical outcomes and also may respond differently to different therapeutic agents used in cancer treatment. But the current understanding of the underlying biology does not permit "individualization" of a particular cancer patients' care. As a result for breast cancer, for example, many women today are given systemic treatments such as chemotherapy or endocrine therapy in an attempt to reduce her risk of the breast cancer recurring after initial diagnosis. Unfortunately, this systemic treatment only benefits a minority of women who will relapse, hence exposing many women to unnecessary and potentially toxic treatment. New prognostic tools developed using micro-array technology show potential in allowing us to facilitate tailored treatment of breast cancer patients. These genomic

tools may be a much needed improvement over currently used clinical methods.

[0003] Histological grading of breast carcinomas has long been recognised to provide significant clinical prognostic information. However, despite recommendations by the College of American Pathologists for use of tumor grade as a prognostic factor in breast cancer, the latest Breast Task Force serving the American Joint Committee on Cancer (AJCC) did not include it in its staging criteria, citing insurmountable inconsistencies between institutions and lack of data. This may be in part related to inter-observer variability and the various grading approaches used, resulting in poor reproducibility across institutions. With the advent of standardized methods such as those developed by Elston and Ellis (Histopathology 19 (5) p. 403-410 (1991)), concordance between institutions has been improved. Nevertheless, whilst grade 1 (low risk) and 3 (high risk) are clearly associated with different prognoses, tumors classified as intermediate grade present a difficulty in clinical decision making for treatment because their survival profile is not different from that of the total (non-graded) population and their proportion is large (40%-50%). A more accurate grading system would allow for better prognostication and improved selection of women for further breast cancer treatment.

[0004] The majority of breast cancers diagnosed today are hormone responsive. Tamoxifen is the most common anti-estrogen agent prescribed today in the adjuvant treatment of these patients. Yet up to 40% of these patients will relapse when given tamoxifen in this setting. At present, due to the positive results of several large trials evaluating the use of aromatase inhibitors instead of, or in combination or sequence with tamoxifen in the adjuvant setting, there are many options available for post

menopausal women with hormone responsive breast cancer. Furthermore, it is unclear which treatment option is the best especially given that the long term health costs of aromatase inhibitor use are unknown. The ability to identify a group at high risk of relapse when given tamoxifen could aid in identifying patients for whom tamoxifen is probably not the best option. These patients could then be specifically targeted for alternative treatment strategies.

[0005] Accordingly need exists for methods and systems that can accurately assess prognosis and hence help oncologists tailor their treatment decisions for the individual cancer patient. In particular, a need exists for methods and systems directed to breast cancer patients.

Aims of the Invention

[0006] The present invention aims to provide new methods and tools for obtaining cancer prognosis and/or for obtaining a clinical outcome, preferably a survival outcome of a cancer affecting a (human) subject especially a subject treated by (with) an antiestrogen compound or an aromatase inhibitor, that do not present the drawbacks of the methods and tools of the state of the art or which improves the methods and tools of the state of the art.

Summary of the Invention

[0007] The present invention is related to a gene set comprising at least one, two, three gene sequences preferably four, five, six, seven or eight (but not more than eight) gene sequences or specific portions thereof (probes or primer sequences) selected from the group consisting of gene sequences CCNB1, CCNA2, CDC2, CDC20, MCM2, MYBL2, KPNA2 and STK6 which are unexpectedly sufficient for performing a gene expression analysis to

obtain an efficient prognosis and diagnosis of cancer, especially breast cancer

[0008] The gene set of the invention is also unexpectedly sufficient for obtaining a clinical outcome, preferably a survival outcome of a subject (human patient) affected by a cancer, especially a patient treated by (with) an antiestrogen compound and/or an aromatase inhibitor (hormonotherapy).

[0009] The inventors have identified that this gene set is sufficient for obtaining by gene expression analysis a prediction upon the efficiency of an anti-tumoral treatment (and calculating the relapsing score after treatment with administration of a selective antitumoral compound).

[0010] This gene set of the invention could be used for gene expression analysis in a method for selecting the most adequate and effective treatment to be applied to this patient, for avoiding an hormonotheray upon this group of patients, where this hormonotheray is not efficient for the treatment of these patients, which presents ER+ type breast cancer or ER- type breast cancer.

[0011] The inventors have also discovered unexpectedly that others proliferation gene sequences could be used for obtaining a prediction of the efficiency (and correlating the relapsing score of the treatment with a selective antitumoral compound) especially with the sequences of the gene set and method described in the document WO2006/119593, especially by using a prognosis method based upon the gene expression grade index (GGI) analysis.

[0012] Preferably, in the present invention, the gene set comprising at least four (but less than eight) gene sequences selecting from the group consisting of

CCNB1, CDC2, CDC20, MCM2, MYBL2, CCNA2, STK6 and KPNA2 gene sequences.

[0013] Preferably, the gene set comprises these at least four genes, more preferably consists only of four gene sequences that are CCNB1, CDC2, CDC20 and MCM2 or more preferably only the four gene sequences CDC2, CDC20, MYBL2 and KPNA2 or specific portion of these sequences (probes, primers, etc.). Therefore, the set may comprise the gene sequences CDC2, CDC20, KPNA2 and MYBL2 (this last sequence being possibly replaced by CCNB1, MCM2, STK6 or CCNA2) or the gene sequences CDC2, CDC20, KPNA2, MYBL2 and one, two, three or four gene sequences selected from the group consisting of CCNB1, MCM2, STK6 and CCNA2 gene sequences.

[0014] The expression analysis of these gene sequences in a tumoral sample are sufficient for obtaining an efficient prognosis and diagnosis and prediction of the treatment to be applied to a subject (human patient) suffering from a cancer, especially breast cancer, more preferably ER+ type breast cancer

[0015] The expression analysis of these genes also presents a prediction of a treatment to be applied (or to be avoid) to this subject (human patient) suffering from ER- type breast cancer. The characteristics of the genes can be found in various databases, for instance upon the websites www.genecards.org www.genenames.org www.ncbi.nlm.nih.gov

[0016] These genes present the following characteristics:

[0017] MYBL2 (Refseq : NM_002466) : V-myb myeloblastosis viral encogene homolog (avian)-like 2 is a member of the MYB family of transcription factor genes, a nuclear protein involved in cell cycle progression. The encoded protein is phosphorylated by cyclin A/cyclin-

dependent kinase 2 during the S-phase of the cell cycle and possesses both activator and repressor activities. It has been shown to activate the cell division cycle 2, cyclin D1, and insulin-like growth factor-binding protein 5 genes. Transcript variants may exist for this gene, but their full-length natures have not been determined.

[0018] KPNA2 (Refseq : XM_001133253): Karyopherin alpha 2 is implicated in the import of protein to the nuclear envelope. KPNA2 is also a regulator of cell cycle checkpoint mediators and may play a role in V(D)J recombination.

[0019] CDC2 (also known as Cdk1) (Refseq : NM_001786): cell division cycle 2 protein is a member of the Ser/Thr protein kinase family. This protein is a catalytic subunit of the highly conserved protein kinase complex known as M-phase promoting factor (MPF), which is essential for G1/S and G2/M phase transitions of eukaryotic cell cycle. Mitotic cyclins stably associate with this protein and function as regulatory subunits. The kinase activity of this protein is controlled by cyclin accumulation and destruction through the cell cycle. The phosphorylation and dephosphorylation of this protein also play important regulatory roles in cell cycle control.

[0020] CDC20 (Refseq : NM_001255): cell division cycle 20 homolog (S. Cerevisiae) appears to act as a regulatory protein interacting with several other proteins at multiple points in the cell cycle. It is required for two microtubule-dependent processes, nuclear movement prior to anaphase and chromosome separation.

[0021] STK6 (also called Aurora kinase A) (Refseq : NM_003600) is a member of a family of mitotic serine/threonine kinases. It is implicated with important processes during mitosis and meiosis whose proper function is integral for healthy cell proliferation. Aurora A is activated by one or more phosphorylations and its activity peaks during the G2 phase to M phase transition in the cell cycle.

[0022] CCNB1 (Refseq : NM_031966) : Cyclin B1 is a regulatory protein involved in mitosis. The gene product complexes with p34 (cdc2) to form the maturation promoting factor (MPF). Two alternative transcripts have been found, a constitutively expressed transcript and a cell cycle-regulated transcript, that is expresses predominantly during G2/M phase. The different transcripts result from the use of alternate transcription initiation sites.

[0023] CCNA2 (Refseq : NM_001237) : Cyclin A2 that belongs to the highly conserved cyclin family, whose members are characterized by a dramatic periodicity in protein abundance through the cell cycle. Cyclins function as regulators of CDK kinases. Different cyclins exhibit distinct expression and degradation patterns which contribute to the temporal coordination of each mitotic event. In contrast to cyclin A1, which is present only in germ cells, this cyclin is expressed in all tissues tested. This cyclin binds and activates CDC2 or CDK2 kinases, and thus promotes both cell cycle G1/S and G2/M transitions.

[0024] MCM2 (NM_004526): Mini-chromosome maintenance deficient 2 (mitotin) is one of the highly conserved mini-chromosome maintenance proteins (MCM) that are involved in the initiation of eukaryote genome replication. The

hexameric proteic complex formed by MCM proteins is a key component of the pre-replication complex (pre-RC) and may be involved in the formation of replication forks and in the recruitment of other DNA replication related proteins. This protein forms a complex with MCM 4, MCM 6 and MCM 7 and has been shown to regulate the helicase activity of the complex. This protein is phosphorylated and thus regulated by protein kinases CDC2 and CDC7.

[0025] Advantageously, the gene sequences set could be included in a detection kit or device that may (further) comprise the following primer sequences SEQ ID NO 1 to SEQ ID NO 16 and possibly one or more elements (cycle, buffer, polymerase,…) used for a detection (amplification of these genes present in a biological sample).

[0026] The detection kit or device according to the invention or the gene sequences set according to the invention could also comprise additional normalization gene sequences used as reference genes. Preferably, these gene sequences are selected from the group consisting of the gene sequences TFRC, GUS, RPLPO and TBP. The characteristics of these genes can be found in various databases, for instance on the websites www.genecards.org www.genenames.org www.ncbi.nlm.nih.gov

Advantageously, the primer sequences for the amplification of these gene sequences is also present in the kit according to the invention, preferably they have the sequences SEQ ID NO 17 to SEQ ID NO 24.

[0027] The gene sequences (probes) of this gene sequences set can be bound to a solid support (micro-well plate, plates, beads of glass or plastic material, filters, membranes, etc.) surface as an array and be present in a detection kit or device, possibly including means and media for a genetic amplification, such as real time PCR means and media (preferably for qRT-PCR amplification).

[0028] The present invention is also related to one or more following primer sequences SEQ ID NO 1 to SEQ ID NO 16, for a specific amplification of one or more gene sequences that are preferably present in the kit or device of the invention.

[0029] The kit (or device) or the gene sequences set according to the invention could also comprise one or more additional normalization gene sequence(s) used as reference(s). Preferably, these references gene sequences are selected from the group consisting of the gene sequences TFRC, GUS, RPLPO and TBP or their specific portions. Advantageously, the primer sequences SEQ ID NO 17 to SEQ ID NO 24 for an amplification of these reference gene sequences are also present in the kit or device according to the invention.

[0030] This kit or device may further comprise a computerized system comprising the gene sequence(s) of this gene sequences set bound upon a solid support surface, as an array and a processor module, preferably configured to calculate gene expression grade index GGI or possibly relapse score (RS) based on the gene expression of complementary bounded sequences and possibly to generate a risk assessment for this tumor sample as described in WO 2006/119593.

[0031] The present invention is also related to a method for obtaining a gene expression of nucleotide sequences in a sample by a binding between nucleotide sequences obtained from a tumor sample, having one or more, preferably two, three, four, five, six, seven, eight gene sequences or specific portion thereof (probes, primers, …) selected from the eight or four genes above described, more preferably CCNB1, CCNA2, CDC2, CDC20, MCM2, MYBL2, KPNA2 and STK6, more particularly CCNB1, CDC2, CDC20, MCM2 or CDC2, CDC20, MYBL2 and KPNA2 gene sequences

or portions or one or more of the primer sequences SEQ.ID.NO 1 to SEQ.ID.NO 16,

possibly combined with one or more of the primer sequences SEQ.ID.NO 17 to SEQ.ID.NO 24 for an amplification of these reference gene sequence(s) that are preferably present in the kit according to the invention

for obtaining a prognosis, a diagnosis or a clinical outcome of tumor sample.

[0032] Therefore, the(se) detection step(s) can be combined with a selection step of the adequate treatment to be applied to a patient from which this sample has been obtained, according to the result (gene expression) of this detection obtained by binding between nucleotide sequences present in the sample and the sequences of the set of gene sequences according to the invention.

[0033] Preferably, the method according to the invention is based upon genetic amplification (preferably by PCR), preferably a qRT-PCR based upon the use of the primer sequence(s) above described which allows an amplification of the preferred nucleotide sequences (mRNA) or their complementary strands of the gene nucleotide(s) set of the invention.

[0034] Another aspect of the present invention is related to a method comprising the steps of

(a) measuring gene expression in a tumor sample submitted to an analysis and obtained from a mammal subject, preferably a human patient;

(b) calculating the gene-expression grade index (or genomic grade)(GGI) of the tumor sample using the formula:

$$\sum_{j \in G_3} x_j - \sum_{j \in G_1} x_j$$

wherein: x is the gene expression level of mRNA, G_1 and G_3 are sets of genes up-regulated in histological grade 1

(HG1) and histological grade 3 (HG3) preferably set of genes of the invention, respectively, and j refers to a probe or probe set wherein the gene set comprises or correspond (consist of) the gene (sequences) set of the invention,

(c) possibly selecting the adequate treatment to be applied to a human patient from which this sample has been obtained according to the gene expression grade index (GGI) obtained from this tumor sample. This tumor sample could be ER+ type breast cancer or ER- type breast cancer.

[0035] In the method, kit or device according to the invention, the tumor (cancer) sample submitted to a diagnosis is (obtained) from a tissue affected by a cancer selected from the group consisting of breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, or brain cancer. Preferably, this tumor sample is a breast tumor sample (more preferably a histological breast tumor sample identified as grade HG2, HG1 or HG3). The sample could be also frozen (FS) or dried tumor sample (paraffin-embedded tumor samples (FFPE)) of an (early breast cancer (BC)) patient.

[0036] The method, kit or device may further comprise designating the tumor sample as low risk (GG1) or high risk (GG3) based on the gene expression grade index (GGI). This embodiment may further comprise providing a breast cancer treatment regimen for a patient consistent with the low risk or high risk designation of the breast tumor sample submitted to the analysis.

[0037] The gene expression grade index GGI may include cutoff and scale values chosen so that the mean GGI of the HG1 cases is about -1 and the mean GGI of the HG3

cases is about +1. The cutoff value is required for calibration of the data obtained from different platforms applying different scales:

$$GGI = scale\left[\sum_{j \in G_3} x_j - \sum_{j \in G_1} x_j - cutoff \right]$$

[0038] The G_1 gene set may comprise at least one gene selected from the genes designated as "Up-regulated in grade 1 tumors" (see WO 2006/119593). The G_3 gene set may comprise at least one preferably at least two, three, four, five, six, seven or eight gene(s) selected from the genes of the gene set of the invention.

[0039] In another aspect of the invention, the method according to the invention comprises the steps of (a) measuring gene expression in a tumor sample;

(b) calculating a relapse score (RS) or clinical outcome of the tumor sample especially the survival outcome of a human patient from which this sample has been obtained and which has been possibly treated with addition of an antiestrogen compound or an aromatase inhibitor using the formula:

$$\sum_{i \in G} w_i \sum_{j \in P_i} \frac{x_{ij}}{n_i}$$

wherein: G is a gene set that is associated with distant recurrence of cancer, P_i is the probe or probe set, i identifies the specific cluster or group of genes, w_i is the weight of the cluster i, j is the specific probe set value, x_{ij} is the intensity of the probe set j in cluster i, and n_i is the number of probe sets in cluster i.

[0040] The method, kit or device may further comprises the step of classifying the said tumor sample based on the relapse score as low risk or high risk for cancer relapse. The cutoff for distinguishing low risk from high risk may

be a relapse score (RS) of from -100 to +100 or a relapse score (RS) of from -10 to +10.

[0041] The detected relapse may be relapse after treatment with tamoxifen, aromatase inhibitor, a chemotherapy, an endocrine therapy, an (antibody) immunotherapy or any other treatment method used by the person skilled in the art. Preferably, the relapse is after treatment with tamoxifen. This method could be applied to suppress a non efficient endocrine treatment upon ER+ or ER- breast type cancer.

[0042] The patient's treatment regimen may be adjusted based on the tumor sample's cancer relapse risk status. For example (a) if the patient is classified as low risk, treating the low risk patient sequentially with an antiestrogen compound (selective estrogen receptor modulator or down regulator (SERM or SERD), such as tamoxifen, raloxifen, flaslodex and (sequential) aromatase inhibitors (AIs), or (b) if the patient is classified as high risk, possibly treating the high risk patient with an alternative endocrine treatment other than tamoxifen. For a patient classified as high risk, the patient's treatment regimen may be adjusted to chemotherapy treatment (administration of paclitaxel, taxanes, anthracyclins, 5-fluoruracil, cyclophosphamide, etc.) or specific molecularly targeted anti-cancer therapies or possibly immunotherapy.

[0043] The gene set may be generated and the sample may be collected from an estrogen receptor (or another marker specific of the cancer tissue sample) positive population. The gene set may be generated by a variety of methods and the component genes may vary depending on the patient population and the specific disorder.

[0044] Another aspect of the invention provides a computerized system or diagnostic device (or kit),

comprising: (a) a bioassay module, preferably a bioarray, configured for detecting gene expression for a tumor sample based on the gene set of the invention; and (b) a processor module configured to calculate GGI, possibly RS or possibly the clinical outcome of the tumor sample based on the gene expression and to generate a risk assessment for the breast tumor sample and for selecting the treatment to be applied.

[0045] The inventors have also observed unexpectedly that it is possible to use the primer(s) according to the invention for obtaining an efficient qRT-PCR assay upon a tumor sample obtained directly from a mammal (including a human patient) or upon conserved sample especially frozen (FS) and dried tumor sample (paraffin-embedded tumor samples (FFPE))from early breast cancer (BC) patient.

[0046] The inventors have tested such qRT-PCR assay accuracy and concordance with original micro-array derives GGI(Genomic Grade Index)using breast cancer population from which frozen and paraffin-embedded tumor samples tissues were collected and the inventors have obtained a statistical significant correlation between the Genomic Grade Index (GGI) generated by micro-array and these qRT-PCR assay using frozen (FS material) as well as paraffin-embedded samples (FFPE material) and between the Genomic Grade Index (GGI)using qRT-PCR derived from frozen (FS) and paraffin-embedded tumors samples (FFPE).

[0047] The inventors have tested the prognostic value on an independent ER-positive tamoxifen only treated frozen breast cancer population and on an independent population of paraffin-embedded breast cancer samples consecutively diagnosed at Jules Bordet Institute (Brussels, Belgium).

[0048] The inventors have observed unexpectedly that a high Genomic Grade Index (GGI) level assessed by qRT-PCR associated with a higher risk of recurrence in the global

breast cancer population and particularly in the ER-positive patients. This was in accordance with the present micro-array result. In multivariate analyses, the GGI assessed by qRT-PCR remained significant. Therefore, qRT-PCR based on a limited number of genes, preferably the gene selected in the gene set according to the invention, recapitulate in an accurate and reproducible manner the prognostic power of Genomic Grade Index derived from micro-array using both frozen and paraffin-embedded tumor samples (FFPE).

[0049] Another aspect of the present invention concerns a method for an efficient screening and/or testing of active compound(s) (or treatment method based upon an administration of active compounds) upon cancer that comprises the method and tools according to the invention especially that comprises the step of testing and monitoring and modulating the effects of this compound upon a tumor sample of a mammal subject including human patients by testing the risk of a cancer in these subjects with the method and tools of the invention before and after this compound is applied to the patient.

[0050] Therefore, this method comprises a selection of one or more active compound(s) used in endocrine (anti estrogen) therapy, such as selective estrogen receptor modulator or down regulator (SERM or SERD) i.e. tamoxifen, raloxifen, flaslodex, aromatase inhibitor (AI), in chemotherapy such as anthracyclins, taxanes, 5-fluoruracil, paclitaxel, cyclophosphamide, in molecular targeted anti cancer therapy, in immunotherapy, etc. which could be administrated (separately or simultaneously) to a mammal subject for treating and/or preventing a cancer, and for testing the efficacy of said active compound(s), by collecting from the treated mammal a tumor sample (biopsy) before and after the administration of said compound(s) to

the mammal, submitting said tumor sample to a diagnosis with the method and tools according to the invention (by detecting gene expression in said tumor sample with the genes set according to the invention or the kit or device according to the invention), possibly generating a risk assessment relapsing score or clinical outcome (or genetic profile or pattern) of this tumor sample before or after the administration of the tested compounds and possibly identifying if the compound(s) may have an effect upon a cancer or may present a risk of developing a cancer. Consequently, this method represents a screening testing or monitoring method of new antitumoral or possibly tumoral and toxic compounds.

[0051] The method according to the invention could be applied upon a mammal presenting a predisposition to a cancer or subject, including a human patient suffering from cancer for the monitoring of the effect of the administrated therapeutical active compound(s). The method of the invention may also comprise the step of administrating the active compound to a group of human patient population which presents the same tumoral genetic profile identified by this method.

Brief Description of the Drawings

[0052] Figure 1 represents Kaplan Meyer survival curves for distant metastasis free survival for GGI (high vs. low).

[0053] Figure 2 represents survival analyses of patient ER+ (frozen samples) **(A)** by histological grade (HG1 HG2 and HG3), **(B)** by gene expression grade index (GGI) (GG low and GG high), **(C)** by qRT-PCR performed with 4 selected genes according to the invention (GG(RT-PCR) low and GG(RT-PCR) high). **(D)** Analyse of the node negative patients by qRT-PCR grading (GG(RT-PCR) low and GG(RT-PCR) high). **(E)**

Cross-tab for RT-PCR grading (GG(RT-PCR)), gene expression grade index (GGI) and histological grade (HG). Difference in relapse-free survival between two groups is summarized by the hazard ration (HR) for recurrence with its 95% CI. All statistical test were two-sided.

[0054] Figure 3 represents survival analyses of ER+ and ER- patients (paraffin-embedded samples) in function of the index defined by qRT-PCR performed with 4 selected genes according to the invention. **(A)** Analyse of the whole population (GG(rt-PCR) low and GG(rt-PCR) high) **(B)** Analyse of the Estrogens positive samples (GG(rt-PCR) low and GG(rt-PCR) high). **(C)** Analyse of the Estrogens positive node negative samples (GG(rt-PCR) low and GG(rt-PCR) high). **(D)** Analysis of the patient of histological grade 2 (HG2) tumors by RT-PCR grading. The 90 patients with HG2 tumors were separated into low-and high-risk subsets by this signature as GG(rt-PCR) low and high. Difference in relapse-free survival between two groups is summarized by the hazard ration (HR) for recurrence with its 95% CI. All statistical test were two-sided.

[0055] Figure 4 represents progression free survival (PFS) analyses for ER+ advanced breast cancer tamoxifen only treated patients (N=279) by RT-PCR grading. The low-risk patients recurring 7.5 month later is compared to high-risk patients (difference observed at 50% survival). Difference in relapse-free survival between two groups is summarized by the hazard ration (HR) for recurrence with its 95% CI. All statistical test were two-sided.

Detailed Description of the Invention

Definitions

[0056] Most terms scientific, medical and technical terms are commonly understood to one skilled in the art.

[0057] The term "micro-array" refers to an ordered arrangement of hybridizable array elements, preferably polynucleotide probes, on a substrate (an insoluble solid support surface).

[0058] The terms "differentially expressed gene", "differential gene expression" and their synonyms, which are used interchangeably, refer to a gene whose expression is activated to a higher or lower level in a subject suffering from a disease, specifically cancer, such as breast cancer, relative to its expression in a normal or control subject. The terms also include genes whose expression is activated to a higher or lower level at different stages of the same disease. It is also understood that a differentially expressed gene may be either activated or inhibited at the nucleic acid level or protein level, or may be subject to alternative splicing to result in a different polypeptide product. Such differences may be evidenced by a change in mRNA levels, surface expression, secretion or other partitioning of a polypeptide, for example. Differential gene expression may include a comparison of expression between two or more genes or their gene products, or a comparison of the ratios of the expression between two or more genes or their gene products, or even a comparison of two differently processed products of the same gene, which differ between normal subjects and subjects suffering from a disease, specifically cancer, or between various stages of the same disease. Differential expression includes both quantitative, as well as qualitative, differences in the temporal or cellular expression pattern in a gene or its

expression products among, for example, normal and diseased cells, or among cells which have undergone different disease events or disease stages. For the purpose of this invention, "differential gene expression" is considered to be present when there is at least an about two-fold, preferably at least about four-fold, more preferably at least about six-fold, most preferably at least about ten-fold difference between the expression of a given gene in normal and diseased subjects, or in various stages of disease development in a diseased subject.

[0059] Gene expression profiling: includes all methods of quantification of mRNA and/or protein levels in a biological sample.

The term "prognosis" and "diagnosis" are used herein to refer to the prediction of the likelihood of cancer-attributable death or progression, including recurrence, metastatic spread, and drug resistance, of a neoplastic disease, such as breast cancer (breast tumor).

[0060] The term "prediction" is used herein to refer to the likelihood that a patient will respond either favorably or unfavorably to a drug or set of drugs, and also the extent of those responses, or that a patient will survive, following surgical removal or the primary tumor and/or chemotherapy for a certain period of time without cancer recurrence.

[0061] The predictive methods of the present invention are valuable tools in predicting if a patient is likely to respond favorably to a treatment regimen, such as, chemotherapy with a given drug or drug combination, and/or radiation therapy, or whether long-term survival of the patient, following surgery and/or termination of chemotherapy or other treatment modalities is likely.

[0062] The term "high risk" means the patient is expected to have a distant relapse in less than 10 years, 5 years, 4 years preferably in less than 3 years.

[0063] The term "low risk" means the patient is expected to have a distant relapse after 10 years, 5 years, 4 years preferably after more than 3 years.

[0064] The term "tumor (cancer) sample" corresponds to any sample obtained from a tissue or cell mammal subject (preferably a human patient that may present a predisposition to a cancer) and obtained from a biological fluid of a mammal subject (preferably a human patient) or a biopsy, including frozen or dried (paraffin embedded tumor sample, preferably human) tumor sample.

[0065] The term "tumor," as used herein, refers to all neoplastic cell growth and proliferation, whether malignant or benign, and all pre-cancerous and cancerous cells and tissues.

[0066] The terms "cancer" and "cancerous" refer to or describe the physiological condition in mammals that is typically characterized by unregulated cell growth. Examples of cancer include but are not limited to, breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, and brain cancer.

[0067] Raw "GGI" (Gene expression grade index) is the sum of the log expression (or log ratio) of all genes high-in-HG3 - sum of the log expression (or log ratio) of all genes high-in-HG1 and can be written as:

$$\sum_{j \in G_3} x_j - \sum_{j \in G_1} x_j$$

wherein:

x is the gene expression level of mRNA,

[0068] G_1 and G_3 are sets of genes up-regulated in HG1 and HG3, respectively, and j refers to a probe or probe set.

[0069] GGI may include cutoff and scale values chosen so that the mean GGI of the HG1 cases is about -1 and the mean GGI of the HG3 cases is about +1:

$$GGI = scale\left[\sum_{j \in G_3} x_j - \sum_{j \in G_1} x_j - cutoff \right]$$

The cutoff in GGI is 0 and corresponds to the mean of means.

GGI ranges in value from -4 to +4.

[0070] As above described proliferation capture by the Genomic Grade Index (GGI) is an important prognostic factor in breast cancer, far beyond estrogen receptor status and encompasses a significant portion of the predictive power of many previously published prognostic signatures.

[0071] Surprisingly, GGI correlates with response to chemotherapy in both ER-negative and ER-positive breast cancer patients, with only 11.9 % of patients having low GGI having complete response to chemotherapy, while 40.4 % of patients having high GGI have a complete response after the same treatment.

[0072] The inventors were also able to convert and validate by qRT-PCR assay the prognostic value of GGI using frozen (FS) and paraffin-embedded tumor samples (FFPE) from early breast cancer patients. The inventors have developed a qRT-PCR assay based on 8 selected GGI genes involved in different phases of the cell cycle and 4 reference genes. These selected genes are CNB1, CCNA2, CDC2, CDC20, MCM2, MYBL2, KPNA2 and STK6 (4 reference genes are TFRC, GUS,

RPLPO and TBP). The preferred 4 selected genes are either CDC2, CDC20, CCNB1 and MCM2 (assay 1) or more preferably CDC2, CDC20, MYBL2 and KPNA2 (assay 2).

[0073] The inventors have tested the accuracy of this qRT-PCR assay in concordance with the original micro-array derived GGI above described by using breast cancer population from which frozen, paraffin-embedded tumor samples tissues and micro-array data were available (N=30). A statistically significant correlation was observed between GGI generated by micro-array and qRT-PCR assays (1 and 2) using frozen material (for assay 2 : HR=0.945, (95%CI: 0.856-0.98, p=3.67E-09) and FFPE material (for assay 2: HR= 0.889, (95%CI:0.721-0.958), p=8.26E-07) as well as between GGI using qRT-PCR derived from frozen and FFPE tumor samples assays (1 and 2) (for assay 2: HR= 0.851, 95%CI: 0.636-0.943), p=7.73E-06).

[0074] The prognostic value of the qRT-PCR assay 1 and 2 has been tested upon a population of 78 hormono-dependant breast tumor of frozen sample tissue. Statistically significant correlation was observed between a high relapsing risk and an elevated expression of these 4 genes of the bio-assay 1 and 2 (HR for bioassay 2= 3.338(95%CI:1.189-9.374),p=0.022). The prognostic value of the bio-assay 1 and 2 remains significative during multivariable analyses (HR for bioassay 2=3.267 (95%CI:1.157-9.227),p=0.025) together with age (<50 years) and tumor size (>2cm).

[0075] The inventors have also assessed the prognostic value of this assay 2 on a population of 208 breast cancers operated consecutively at the Bordet Institute between 1995 and 1996.

[0076] These samples are paraffin-embedded tumor sample tissues. Statistically significant correlation has been observed between the high relapsing risk and high

expression of the 4 genes of this bio-assay in global population (HR=1.072 (95%CI:0.999-3.507), p=0.050) and in particular in sub-population of breast cancers hormone-dependant (HR=2.26(95%CI:1.075-4.751),p=0.032).

[0077] The prognostic value remains significant even during multivariable analyses together with nodal invasion for the global population (HR=1.880(95%CI:0.941-3.757),p=0.074) and the ER positive subgroup (HR=2.249(95%CI:0.982-5.150),p=0.055).

[0078] This prognostic value of the bio-assay 2 has been also validated upon another independent population of 106 paraffin-embedded breast tumor sample with similar results.

[0079] A bio-assay based upon a limited number of genes, such as the four genes selected from the set of genes as described in the present invention, preferably a qRT-PCR assays (assay 1 or assay 2) allows an accurate and reproducible manner the prognostic power of micro-array derived GGI using both frozen and paraffin-embedded tumor samples. As described in the figures 8 to 11 prognostic value of qRT-PCR assay 2 is comparable to a prognostic value of micro-array.

[0080] As showing in the figures, RT-PCR grade index in a high-risk tamoxifen-only treated patients (JNI) is unexpectedly a strong predictor assay for node negative ER-positive (but also ER-negative) breast type cancer patients, that can be use to avoid unnecessary (and possibly toxic) treatment by hormonotherapy with compounds that are not effective in the treatment of this detected specific group of human patients.

[0081] Conversely, RT-PCR grade index is unexpectedly a strong predictor of positive reponse to chemotherapy for node negative ER-positive (but also ER-negative) breast type cancer.

[0082] Therefore, the gene set diagnostic kit and method according to the invention is also a predictive assay and method for these patients.

[0083] Different embodiments of the present invention have been described according to the present invention. Many modifications and variations may be made to the techniques and structures described and illustrated herein without departing from the spirit and scope of the invention. Accordingly, it should be understood that the apparatuses described herein are illustrative only and are not limiting upon the scope of the invention.

CLAIMS

1. A gene set comprising at least 4, but less than 8 genes that are selected from the group consisting of the genes CDC2, CDC20, MYBL2, KPNA2, CCNB1, MCM2, CCNA2 and STK6.

2. The gene set according to the claim 1, wherein the genes are selected from the group consisting of CDC2, CDC20, MYBL2 and KPNA2 genes.

3. The gene set according to the claim 1, wherein the genes are selected from group consisting of CCNB1, CDC2, CDC20 and MCM2 genes.

4. The gene set according to the claims 1 to 3, wherein the genes sequences are bounded to a solid support surface, as an array.

5. A diagnostic kit or device comprising the gene set according to any of the claims 1 to 4, possibly including means for real time PCR analysis of a tumor sample.

6. The diagnostic kit or device according to the claim 5, wherein the means for real time PCR analysis are means for qRT-PCR.

7. The diagnostic kit or device according to the claim 5 or 6, which comprises one or more of the primer sequence(s) selected from the group consisting of SEQ ID NO 1 to SEQ ID NO 16.

8. The diagnostic kit or device according to the claim 7, which further comprises means for real time PCR analysis of reference genes.

9. The diagnostic kit or device according to the claim 8, wherein the reference genes are selected from the group consisting of TFRC, GUS, RPLPO and TBP genes.

10. The diagnostic kit or device according to the claim 8 or 9, which further comprises one or more primer sequence(s) selected from the group consisting of SEQ ID NO 17 to SEQ ID NO 24.

11. The kit or device according to any of the claims 8 to 10 which is a computerized system comprising :

a) a bio assay module configured for detecting gene expression for a tumor sample based on the gene set according to any of the claims 1 to 4 and,

b) a processor module configured to calculate gene-expression grade index (GGI) or relapse score(RS) based on the gene expression and to generate a risk assessment for the tumor sample.

12. The kit or device according to any of the preceding claims 5 to 11, wherein the tumor sample is from a tissue affected by a cancer selected from the group consisting of breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, or brain cancer, preferably a breast tumor sample.

13. The kit or device of claim 12, wherein the tumor sample is a frozen sample (FS) or a paraffin-embedded tumor sample/FPPE).

14. A method for the prognosis or diagnosis of cancer in a tumor sample which comprises the step of putting into contact nucleotide sequences obtained from this tumor sample with the gene set according to any of the claims 1 to 4 and measuring gene expression of the nucleotide sequences in the tumor sample and correlating the expression of the nucleotide sequences with cancer prognosis or diagnosis.

15. The method according to the claim 14, which is combined to an estrogen receptor and/or progesterone receptor gene expression detection.

16. The method of claim 14 or 15, wherein the tumor sample is a frozen or a paraffin-embedded tumor sample.