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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 June 2009 (18.06.2009)

PCT

(10) International Publication Number
WO 2009/074968 A2

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

(21) International Application Number:
PCT/IB2008/055252

(22) International Filing Date:
12 December 2008 (12.12.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/996,961 12 December 2007 (12.12.2007) US

(71) Applicant *(for all designated States except US)*: **ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE (EPFL)** [CH/CH]; CM-Ecublens, CH-1015 Lausanne (CH).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **FARMER, Pierre** [CA/FR]; 14, rue Bernard Moutardier, F-74500 Evian-les Bains (FR). **DELORENZI, Mauro** [CH/CH]; Route de Berne 228, CH-1066 Epalinges (CH). **BONNEFOI, Hervé** [FR/FR]; 72 quai des Chartrons, F-33000 Bordeaux (FR). **IGGO, Richard** [GB/FR]; 166 Cours de l'Argonne, apt. 2105, F-33076 Bordeaux (FR).

(74) Agent: **KATZAROV SA**; 19, rue des Epinettes, CH-1227 Geneva (CH).

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

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

Declaration under Rule 4.17:
— *of inventorship (Rule 4.17(iv))*

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

WO 2009/074968 A2

(54) Title: METHOD FOR PREDICTING THE EFFICACY OF CANCER THERAPY

(57) Abstract: The present invention relates to a method and a kit for predicting the efficacy of cancer therapy in a subject who has undergone or is undergoing chemotherapy treatment for cancer.

METHOD FOR PREDICTING THE EFFICACY OF CANCER THERAPY

FIELD OF THE INVENTION

The present invention relates to a method and a kit for predicting the efficacy of cancer therapy in a subject who has undergone or is undergoing chemotherapy treatment for cancer.

BACKGROUND OF THE INVENTION

Many cancer patients are diagnosed at a stage in which the cancer is too far advanced to be cured, and most cancer treatments are effective in only a minority of patients undergoing therapy. Therefore, there has been much interest in cancer signatures (herein biomarkers) in predicting future patterns of disease, especially as cancer treatment has made such positive strides in the last few years. Cancer signatures provide a powerful and dynamic approach to understanding the spectrum of malignancies with applications in observational and analytic epidemiology, randomized clinical trials, screening, diagnosis and prognosis. Defined as alterations in the constituents of tissues or body fluids, these signatures offer a means for homogeneous classification of a disease and risk factor, and they can extend one's basic information about the underlying pathogenesis of disease. The goals in cancer research include finding signatures (biomarkers) that can be used for the early detection of cancers, predict the efficacy of a cancer therapy and to identify underlying processes involved in the disease.

For example breast cancer is the most common malignancy among women, and has one of the highest fatality rates of all cancers affecting females. In fact, breast cancer remains the leading cause of cancer deaths in women aged 20-59. Adjuvant systemic chemotherapy for breast cancer decreases the risk of relapse and improves overall survival[1] by 10 to 50% depending on patient's age and tumour's estrogen receptors (ER) status of the tumour[1]. When administrating chemotherapy before surgery, a complete disappearance of the invasive component of the primary tumour (complete pathological response; pCR) is observed in approximately 10% of ER positive tumours and 20 to 30% of ER negative tumours[2-5], thus breast cancers have heterogeneous sensitivity to chemotherapy. Complete pathological

response correlates with a longer survival and is therefore considered as a surrogate measure of chemosensitivity[2,4,6-8].

In December 2006, Potti et al. [8] developed predictive signatures based on the *in vitro* response of cell lines to chemotherapy and validated these signatures in clinical datasets. However, the use of cell lines has the disadvantage of ignoring the influence of the tumour stroma microenvironment on drug response. Indeed, recent pre-clinical findings showed the influence of extracellular matrix proteins, such as fibronectin and laminin, on cell line sensitivity to cytotoxic drug and radiation[12-15]. However, these signatures did not give new insights in the understanding of chemosensitivity or resistance mechanisms *in vivo*.

In many cases of anticancer therapies, biomarkers are critical to predict efficacy of the therapy for individual subjects. Biomarkers can be used to predict efficacy before treatment or can be monitored to predict the therapeutic response shortly after initiation of treatment. These biomarkers are useful to select appropriate subjects for the therapy and to save remaining subjects, in whom the therapy is unlikely to exhibit any clinical benefit, from unnecessary side effects and costs. Although a variety of gene alterations have been identified, no single biomarker can reliably predict response to therapy and outcome. There still exists a need for additional sets of biomarkers for individuals having cancers. Currently, there is no diagnostic test, used in clinic that predicts sensitivity of subjects with cancer to various chemotherapy regimens.Therefore, it would be highly desirable to be able to identify whether a subject with cancer, and in particular with breast cancer or colorectal cancer, who has undergone or is undergoing chemotherapy treatment for cancer will be responsive to chemotherapy, in particular anthracycline-based neo-adjuvant chemotherapy.

SUMMARY OF THE INVENTION

This object has been achieved by the Applicants in the present invention which provides for a method for predicting the efficacy of cancer therapy in a subject who has undergone or is undergoing chemotherapy treatment for cancer, characterized in that said method comprises

(a) obtaining a stromal tissue sample from a subject,

(b) determining in said stromal tissue sample the expression values of at least two stromal genes and of at least two reference genes,

(c) defining stromal content (SC) from the expression values of step (b),
(d) comparing the stromal content (SC) with a reference threshold,
(e) predicting resistance to chemotherapy of said subject based on the step (d), wherein high stromal content is indicative of resistance to chemotherapy, while low stromal content is indicative of sensitivity to chemotherapy,
(f) adapting the treatment of said subject

The present invention further provides for a kit for predicting the efficacy of cancer therapy in a subject who has undergone or is undergoing chemotherapy treatment for cancer, characterized in that said kit comprises
(a) a reagent for detecting mRNA levels of at least two stromal genes selected from the group consisting of the genes of Table 4 and of at least two reference genes selected from the group consisting of the genes of Table 5 in a stromal tissue sample from a subject, and
(b) an instruction sheet.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1. Flow chart of experimental procedures. Flow chart showing the steps in the procedure for selecting and testing gene expression modules identified with the multiple regression model. The brackets indicate the dataset used for a particular step.

Figure 2. Gene Set Enrichment Analysis (GSEA) confirms that expression modules are associated to their respective biological process. The GSEA measures the distributional bias of a subset of genes defined a priori within a larger ordered list of genes. Nine independent tests were performed, one for each of the 9 biological process included in the study. All reference gene subsets were taken from MSigDM ((http://www.broad.mit.edu/gsea/). For each test, all genes of the NKI-EMC matrix were ordered (from left to right) and weighted according to their expression similarity to the representative gene indicated below the panel. The distributional bias of genes of the geneset indicated above each panel was measured. A positive relationship between the gene ordering and the tested geneset is visualized by a leftward distribution bias. Results are expressed as percentage of maximal theoretical score. P-values were obtained empirically by randomizing 100 000 times the composition of the genes included in the geneset.

Figure 3. Heatmap and response data for stromal genes. The expression data for the individual genes are shown in the heatmap after mean centring. Panel (A) Results observed in the EORTC study and in (B) the MDA study. The lower panels (C) and (D) show the stroma metagene score for the EORTC and MDA studies for the. The score is defined as the mean expression value of the 50 genes included in the metagene. Patient ordering in panels (A) and (B) are maintained in (C) and (D). pCR, red; non-pCR, blue.

Figure 4. Influence of the relative decision threshold on classification statistics. The upper and lower panels represent results obtained in the EORTC and MDA datasets respectively. The vertical gray line represents the decision threshold that maximizes the Youden index in the EORTC dataset and the associated circle symbols for each line indicate the points used for the classification statistic reported in table 2 c. The horizontal axis represents the various decision thresholds (cutoff) used to make the PCR-nPCR prediction. Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), Positive predictive value (PPV), Negative predictive value (NPV) and Youden index (YOU).

Figure 5. Prediction of pCR using metagenes derived from the DCN expression module. The horizontal axis represents different metagenes composed of genes chosen in decreasing order of association to the representative gene in the NKI-EMC dataset; window position 1 is constructed with the first 15 genes, window position 2, with the next non-overlapping 15 genes). Red points are metagenes where all genes in the expression module are significantly associated with the representative gene ($p \leq 0.05$). Black points are metagenes where at least one gene is not significantly associated with the representative gene. The vertical axis is the AUC of the ROC curve for predicting pCR. When the AUC is over 0.5, the metagene predicts resistance or response better than chance. The error bars show the 95% confidence interval of the bootstrap.

Figure 6. Kaplan Meier survival analysis of patients in the Amsterdam, Rotterdam and Duke datasets. For all panels, patients were divided into 2 groups based on their metagene score and survival of patients with tumours falling into the upper (red) and lower (blue) halves was compared. The NKI (n=189) [21] and EMC (n=286) [36] datasets comprise breast cancer patients that received no chemotherapy while all patients of the Duke dataset (n = 120)

received chemotherapy. Panels (A, B, C): The metagene score for the proliferation module was used to split the patients. Panels (D, E, F): The metagene score for the stroma was used to split patients. Significance was determined with the log rank statistics.

Figure 7. Activated stroma is associated to response to FEC chemotherapy. Panels(A) and (B): Scatter plots of metagene scores for individual tumours. The stroma metagene is plotted against SFT (A) or DTF (B) metagene scores derived from the published gene lists. Each point represents a single tumour: pCR, red; non-pCR, blue. Panel C: Gene Set Experiment Analysis (GSEA) measuring the distribution of Wnt target genes within the DCN expression module. All genes of the NKI-EMC expression matrix were ranked (from left to right) and weighted according to the meta-analytical t statistics for the DCN representative gene. Results are expressed as percentage of maximal theoretical score. Significance was determined by 100 000 permutations of the ranks of the DCN expression module. Panel (D): Stripchart of epithelial-stroma specific scores for selected gene lists. Scores were calculated as follows: Laser Dissection Microscopy (LDM) was performed on 3 colon carcinoma patients. For each patient, 2 fractions were isolated: (1) cancerous epithelial cells and reactive stroma. The score is the log ratio between the average expression in the epithelial fraction and average expression in the reactive stroma. A negative score suggest that the gene is more specifically expressed in reactive stroma compared to epithelial cells. Each symbol represents the mean value of an individual gene of the list; Red vertical bar represent the average score of all genes of the list. Significance was determined by randomly selecting the same number of genes, then comparing their mean score to the observed mean score. A total of 1000 permutations were performed. NBS and NBE are gene lists of normal mammary fibroblast and normal mammary epithelial tissues; DTF, desmoid-type fibromatosis and SFT solitary fibrous tumour.

Figure 8. Stripchart of epithelial-Cancer Associated Fibroblasts specific scores in function of selected gene lists. Scores were calculated as follows: Laser Dissection Microdisection (LDM) was performed on 3 colon carcinoma patients. Scores are defined as Average expression in the epithelial fraction minus average expression in the CAF fraction. A negative score suggest that the gene is more specifically expressed in CAF relative to epithelial cells. Each symbol represents the mean value of an individual gene of the list; Red vertical bar represent the average score of all genes of the list. Significance, determined

empirically was determined the probability that the observed average score be generated by the same number of randomly selected genes. A total of 1000 permutations were performed. NBS and NBE are respectively gene lists of normal mammary fibroblast and epithelial tissues; DTF, desmoid-type fibromatosis and SFT solitary fibrous tumour.

Figure 9. Prediction of pCR using metagenes derived from the DCN signature. Tumors from the EORTC dataset are ranked according to the probability of pCR after conversion of the metagene scores into probabilities by logistic regression. The metagene scores are the averages of the expression level of genes included in the DCN signatures. Open symbol (white) npCR, Closed symbol (black) pCR.

Figure 10. Prediction of pCR using metagenes derived from the DCN signature. Tumors from the MDA dataset are ranked according to the probability of pCR after conversion of the metagene scores into probabilities by logistic regression. The metagene scores are the averages of the expression level of genes included in the DCN signatures. Open symbol (white) npCR, Closed symbol (black) pCR.

Figure 11. Stromal signature associated with response in an independent cohort of 23 rectal carcinoma patients treated with preoperative fluorouracil (1,000 mg/m2/d), as single agent

DETAILED DESCRIPTION OF THE INVENTION

Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. The publications and applications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention. In addition, the materials, methods, and examples are illustrative only and are not intended to be limiting.

In the case of conflict, the present specification, including definitions, will control.

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as is commonly understood by one of skill in art to which the subject matter herein belongs. As used herein, the following definitions are supplied in order to facilitate the understanding of the present invention.

The term “comprise” is generally used in the sense of include, that is to say permitting the presence of one or more features or components.

As used in the specification and claims, the singular form "a", "an" and "the" include plural references unless the context clearly dictates otherwise. For example, the term "a cell" includes a plurality of cells, including mixtures thereof. The term "a gene" includes a plurality of genes.

As herein used, “a gene signature” is used to designate a specific combination of genes, which serves as a biomarker for a specific phenotype, state or outcome, herein chemosensitivity.

The abbreviation “FEC” refers to a combination of 5-Fluorouracil, Epirubicin, and Cyclophosphamide, a widely prescribed pre-operative chemotherapy regimen.

The abbreviation “LDM” refers to laser dissection microscopy.

The abbreviation “TOP2A” refers to TOPOISOMERASE, DNA, II, ALPHA. Enzymes that control and alter the topologic states of DNA in both prokaryotes and eukaryotes. Topoisomerase II from eukaryotic cells catalyzes the relaxation of supercoiled DNA molecules, catenation, decatenation, knotting, and unknotting of circular DNA. It appears likely that the reaction catalyzed by topoisomerase II involves the crossing-over of 2 DNA segments.

The abbreviation “T-FAC” refers to a chemotherapy regimen (i.e. a combination of chemotherapy drugs) given to breast cancer patients. T-FAC is the acronym for Taxotere® (Docetaxel), Fluorouracil® (5-FU), Adriamycin® (doxorubicin), Cytoxan® (cyclophosphamide).

The abbreviation "NKI-EMC" refers to two external publicly available data sets, the van de Vijver [3](n=295, Agilent platform, obtained from author's web-site) and Wang [4] (n=286, Affymetrix platfrom, GEO:GSE2034) datasets, used herein to define the expression modules. A total of 10317 genes could be cross-matched between these two platforms and were used to define the expression modules.

The term "Expression Module" is used herein to designate a group of genes significantly associated, in terms of similarity of expression, with that of the prototype gene.

The abbreviation "pCR" refers to Pathological Complete Response, i.e. disappearance of the tumour after treatment, with at most scattered tumour cells detected by the pathologist in the resection specimen.

The abbreviation "npCR" refers to Non Pathological Complete Response.

The abbreviation "MIAME" refers to Minimum Information About a Microarray Experiment, an international standard for annotation of microarray data.

The term "Multiple Regression" designs linear regression with two or more independent variables.

As used herein, the term "prototype gene" (or "representative gene") is used to designate a gene identified as a typically representative of a large, highly correlated cluster of genes. These cluster genes were observed to be made of functionally related genes that consistently cluster together regardless of the dataset being analyzed.

As used herein, the term "Metagene" designates the average expression value of a subset of genes, all belonging to the same expression module. Therefore a metagene is a "virtual" gene that summarizes the information contained in many real genes into a single value per sample.

The abbreviation "MDA dataset" is herein used to designate as follows: the nine metagenes were tested for their ability to predict pathological complete response by using the area under the receiver operating characteristic curves (AUC) as a measure of association (Fig. 1 step 5).

An AUC value significantly different from random association (AUC=0.5) was observed for the interferon signalling and stromal metagenes (Tab.1a). These two signatures were then tested in an independent cohort of ER negative tumours (referred to here as the "MDA" dataset).

The term "HUGO name" designates the unique name given by the Human Genome Organisation to a gene. Use of HUGO names is encouraged in order to prevent confusion when, as is often the case, multiple different names have been used in the literature.

The term "sensitivity" is herein used to measure the ability of a classification function to predict pCR when it is truly present. Sensitivity is the proportion of all pCR for whom there is a positive prediction, determined as the number of true positive divided by the sum of true positive + false negatives.

The term "specificity" is herein used to measure the ability of a classification function to predict the absence of pCR when a patient is truly npCR. Specificity is the proportion of npCR patients from whom there is a correct prediction, expressed as the number of true negatives divided by the sum of true negative + false positives.

As herein used, the term "ROC curve" is Receiver Operating Characteristic (curve), a plot of [sensitivity] vs [1-specificity] for a classification function. The AUC (area under the ROC curve) is a useful global measure for how well the two classes are separated, independent from a particular threshold.

The abbreviation "PPV" is Positive Predictive Value. It is the proportion of the correct decisions among the cases declared positive by a particular classification function and a particular threshold, that is the ratio of the true positives to the number of positive calls.

The abbreviation "NPV" refers to Negative Predictive Value. It is the proportion of predicted nPCR patients (negative tests) that are truly nPCR

The term "Bootstrapping" refers to sampling with replacement from a set of data to produce simulated data sets and approximately determine the variability of a parameter estimate. The

term "95% confidence interval of the AUC" is used here for the range of values bounded by the 2.5^{th} and 97.5^{th} centiles of the bootstrap distribution of the AUC.

The abbreviation “DTF” refers to Desmoid-Type Fibromatosis. A form of fibroblastic tumour occurring in patients with germline mutations in the adenomatous polyposis coli gene (APC).

The abbreviation “SFT” is herein used to designate Solitary Fibrous Tumour. A generally benign tumour of fibroblasts.

The abbreviation “CAF” refers to Cancer-Associated Fibroblasts

As herein used, the term “Multiple regression for the selection of expression modules” refers to:

$$GENE_i = \beta_{0i} + \beta_{1i}ESR1 + \beta_{2i}CLCA2 + \beta_{3i}FABP4 + \beta_{4i}GZMA + \beta_{5i}CD83 + \beta_{6i}MX1 + \beta_{7i}DCN + \beta_{8i}ADM + \beta_{9i}TPX2 + \varepsilon_i$$ *(sample's index are removed for clarity)*, wherein

$GENE_i$: the expression of gene i is an independent variable of the linear model. Its variation over the set of profiled tumours is decomposed into linear terms given by the prototype genes as explanatory variables.

ESR1, CLCA2, FABP4, GZMA, CD83, MX1, DCN, ADM, TPX2: the expression vectors of the nine prototype genes.

β_{0i}: the intercept term for gene i

β_{ji}: the regression coefficient for prototype j and gene i. It is a measure of the correlation between the expression vector of the genes j and i, adjusted by the presence of the other explanatory variables in the model.

ε_i: the unexplained variation (residual) term for gene i.

HUGO names of the prototypes (the independent variables):

ESR1: estrogen receptor α, classic marker that distinguishes ER+ and ER- breast cancer subtype

CLCA2: chloride-activated calcium channel 2 cells). CLCA2 is a marker that, conjointly with ESR1, distinguishes the molecular apocrine subtype postulated by Farmer et al [10] from the luminal and basal subtypes

FABP4: fatty acid binding protein 4, a marker for adipocytes
GZMA: granzyme A, a marker for T lymphocytes
CD83: cluster of Differentiation 83, a marker for B lymphocytes
MX1: myxovirus resistance gene 1, a marker for interferon signalling
DCN: decorin, a marker for stroma
ADM: adrenomedullin, a marker for hypoxia
TPX2: aurora kinase targeting subunit, a marker for proliferation.

The term "chemotherapy" generally refers to a treatment of a cancer using specific chemotherapeutic/chemical agents. A chemotherapeutic agent refers to a pharmaceutical agent generally used for treating cancer. The chemotherapeutic agents for treating cancer include, for example, cisplatin, carboplatin, etoposide, vincristine, cyclophosphamide, doxorubicin, ifosfamide, paclitaxel, gemcitabine, fluorouracil and docetaxel. More specifically, the chemotherapeutic agents of the present invention include 5-fluorouracil, epirubicin, and cyclophosphamide or combinations thereof (for example “FEC”).

The term “adapting the treatment” generally refers to the choice of a treatment among different options, based on the specificities of the disease, concomitant pathologies or patient conditions, or the switch from one treatment to another in the course of the therapy because of the non-response, progression or resistance of the disease to the initial treatment, with the intent to offer to the patients the best treatment for their diseases under the given circumstances.

“Stromal tissue” as referred herein is the supportive tissue of an epithelial organ, tumor, gonad, etc., consisting of connective tissues and blood vessels.

As used herein the terms "subject" or "patient" are well-recognized in the art, and, are used interchangeably herein to refer to a mammal, including dog, cat, rat, mouse, monkey, cow, horse, goat, sheep, pig, camel, and, most preferably, a human. In some embodiments, the subject is a subject in need of treatment. However, in other embodiments, the subject can be a normal subject. The term does not denote a particular age or sex. Thus, adult and newborn subjects, as well as fetuses, whether male or female, are intended to be covered. Patient or

subject are used interchangeably and refer to a subject with a disease or disorder. The term patient or subject includes human and veterinary subjects.

As used herein the term "biomarker" is virtually any detectable compound, such as, but not limited to, a protein, a peptide, a carbohydrate, a lipid, or a nucleic acid (e.g., DNA, such as cDNA or amplified DNA, or RNA, such as mRNA), that is present in or derived from a biological sample. "Derived from" as used in this context refers to a compound that, when detected, is indicative of a particular molecule being present in the biological sample. For example, detection of a particular mRNA can be indicative of the presence of the expression of a particular gene in the biological sample. A biomarker can, for example, be isolated from the biological sample, directly measured in the biological sample, or detected in or determined to be in the biological sample. "Biological sample" can be serum, blood, peripheral blood cells, plasma, saliva, amniotic fluid, synovial fluid, lacrimal fluid, milk, lymph and tissue. The tissue is usually a biopsy or surgical specimen taken at tumor removal.

As used herein the term "predicting the efficacy" means to assess the reaction of a cancer to treatment with chemotherapy, i.e to assess the ability of a cancer to respond favourably or to resists to the chemotherapy.

The standard approach to identify a biomarker is to randomly split the dataset into two parts: a learning and a validation set. The learning set is used to identify genes differentially expressed in the two phenotypes using a variable selection method, for example a t-test. Using this methodological framework, biomarkers for pathological complete response (pCR) to FEC neoadjuvant therapy were constructed and tested in three-fold cross-validation. However, statistically significance was never observed. The Applicants noticed that the selected genes varied considerably depending on which tumours were present in the learning set. This gene selection instability may account for the failure to identify a discriminatory signature in this case [16]. To circumvent this difficulty, the Applicants developed a new strategy that aimed to explicitly test the association of the most prominent cancer specific gene clusters with the phenotype of interest.

Figure 1 illustrates the strategy used to construct metagene associated with biological processes prominent in cancer gene expression data. The Applicants [17] identified nine major clusters of co-expressed genes. These clusters are related to epithelial tumour types (luminal,

basal and molecular apocrine), cell physiology (proliferation, hypoxia and interferon signalling) and the tumour microenvironment (T and B cells, adipocytes and stroma) and have been described by others [18-21] (Figure 1 step1). The Applicants aimed to identify groups of genes comprised in these nine clusters using an automated and non-biased procedure. This was achieved by identifying, a priori, a single "representative-gene" (prototype genes) that is typical of each cluster (Figure 1 step 2). These representative-genes (prototype genes) are included as the explanatory variables of a multi-linear regression model in order to identify other genes that are most strongly associated with them (see Example 1). The linear model was not fitted to the Applicants' data but to a large external dataset (NKI-EMC; see Example 1) comprising 583 tumours hybridized to two different types of microarray (Figure 1 step 3). Importantly, no information regarding response to therapy for the NKI-EMC patients was used in this process. The use of a large external dataset brings the advantages of a higher statistical power to correctly identify the most strongly correlated genes and a lower chance of observing a spurious correlation specific to a single study. For each representative gene, the 50 genes most strongly associated with its respective representative-gene were identified and map to the EORTC dataset. For each sample of the EORTC dataset, the expression values of these 50 genes were averaged to generate a single summary value per biological process. The Applicants refer to this value as a "metagene" (Figure 1 step 4). As a result of this procedure the information contained in the initial nine clusters of genes was condensed into 9 metagenes.

The ability of the linear model to single out functionally related genes was verified. For each of the studied process, The Applicants have identified a corresponding entry in the Molecular Signatures Database (MsigDB), a database of manually curated gene sets [22]. For each test, all genes of the NKI-EMC matrix were ordered and weighted according to their expression similarity to the representative-gene. In each case, gene set enrichment analysis (GSEA) revealed a significant association (see Figure 2). The Applicants conclude that each metagene is, as expected, a measure of the activity of the intended biological process at the gene expression level.

The nine metagenes were tested for their ability to predict pathological complete response by using the area under the receiver operating characteristic curves (AUC) as a measure of association (Figure 1 step 5). This statistic was chosen as it is not dependent on any decision threshold. An AUC value significantly different from random association (AUC=0.5) was observed for the interferon signalling and stromal metagenes (Table 1A). These two

signatures were then tested in an independent cohort of ER negative tumours (referred to here as the "MDA" dataset) included in a recent study of response to T-FAC chemotherapy [10] (Figure 1 step 6). The stromal metagene was significantly associated with response in this independent dataset (AUC = 0.70; $p < 0.01$), whereas the interferon metagene was not (Table 1B). Figure 3 shows that the stromal genes have a coherent expression pattern in both the EORTC and the MDA studies. The mean pairwise correlation of stroma genes in the EORTC dataset was 0.55 ($p < 0.0001$). High stromal gene expression is associated with resistance to chemotherapy (ie, absence of a pCR) (Figure 3c and d). Similar range in expression (about 6 fold in absolute change) was observed in the two datasets meaning that the stromal metagene reveals strong differences between these tumours. The datasets were sufficiently similar that a logistic regression model trained on the EORTC data could be successfully applied, with identical model coefficients and decision threshold, to the MDA data. A significant odds ratio of 3.41 ($p < 0.05$) was observed using a decision threshold set at 0.38 (Table 1C). The decision threshold was determined by maximizing the Youden index (specificity + sensitivity -1) on the EORTC data only. The impact of varying the decision threshold on PPV, NPV, sensitivity and specificity statistics is presented in Figure 4. The positive likelihood ratio, a good measure of the added information, is equal to 1.7 and 1.6, respectively, for the EORTC and MDA datasets. The maximum observed PPVs and NPVs are respectively 77% and 86% in the EORTC data and 76% and 89 % in the MDA data, indicating that the current classifier is better predicting resistance then sensitivity. A likely reason is that multiple different mechanisms can confer resistance to anthracycline-based chemotherapy. See also Figures 9 and 10.

Table 1. Prediction of pCR using metagene signatures. The 95% confidence intervals were calculated from the AUC distribution over 1000 cycles of bootstrapping. (A) Prediction of pCR using metagene signatures in the EORTC dataset. P-values were determined empirically and adjusted for False Discovery Rate. (B) Validation of the interferon and stroma metagenes in the MDA study. (C) Prediction of pCR using stromal signature using a logistic regression is trained on the EORTC data and applied on the MDA dataset. The accuracy (ACC), sensitivity (SENS), specificity (SPEC), positive (PPV) and negative (NPV) predictive value and Odds Ratio (OR) were determined at a probability threshold of 0.38 defined by maximizing the Youden index on the EORTC dataset. P-value of a fisher exact test is reported.

Table 1

A. EORTC

Biological Process	Representative Gene	HUGO Names	AUC [C.I. 95%]	P-Value (FDR)
Luminal-Basal	ESR1	Estrogen receptor alpha	0.53 [0.37-0.68]	0.40
Apocrine	CLCA2	Chloride-Associated Calcium Channel 2	0.46 [0.32-0.59]	0.70
Stroma	DCN	Decorin	0.68 [0.54-0.80]	0.03
T Cell	GZMA	Granzyme A	0.62 [0.48-0.75]	0.14
B Cell	CD83	CD83 Antigen	0.58 [0.44-0.71]	0.23
Adipocyte	FABP4	Fatty Acid Binding Protein 4	0.54 [0.38-0.68]	0.40
Proliferation	TPX2	TPX2 microtubule-associated homologue	0.55 [0.39-0.69]	0.40
Interferon	MX1	Myxovirus resistance gene 1	0.72 [0.59-0.84]	< 0.01
Hypoxia	ADM	Adrenomedulin	0.59 [0.44-0.73]	0.23

B.MDA

Biological Process	Representative Gene	HUGO Names	AUC [C.I. 95%]	P-Value
Stroma	DCN	Decorin	0.70 [0.52-0.85]	0.01
Interferon	MX1	Myxovirus resistance gene 1	0.50 [0.33-0.67]	0.60

C.

Dataset	ACC	PPV	SENS	PNV	SPEC	OR	P-Value
EORTC	0.65	0.57	0.86	0.81	0.49	5.51	0.01
MDA	0.65	0.64	0.78	0.67	0.50	3.41	0.05

The relationship of the stromal metagene with the clinical variables histological grade, node status and tumour size, and ERBB2 status were tested for association with pCR by logistic regression (Table 2). The stromal metagene was the only significant variable in the univariate model in both datasets. The multivariate model showed that the stromal metagene is an independent predictive factor in both the EORTC and the MDA datasets. Thus, the stroma metagene does not detect biological information measured by other clinical variables.

It was shown that ER negative tumours were comprised of least two different molecular subtype (basal and ERBB2 or molecular apocrine) [17,19]. The ERBB2-molecular apocrine class was too small to test, but a significant association of the stromal metagene with response was observed in the basal molecular type in both the EORTC dataset (AUC = 0.69 [0.51-0.84]; $p < 0.02$; N=39) and the MDA dataset (AUC= 0.73 [0.54-0.92]; $p < 0.01$; N=27). This result shows that stromal metagene signature is not confounded with different tumour molecular classes found within ER-negative breast cancers.

Table 2. Multivariate analysis of predictive factors for Pathological Complete Response (pCR). Univariate and multivariate logistic regression models for the Stroma signature and other clinical variables.

Table 2

EORTC	Univariate		Multivariate	
Variable	**Coef**	**P Value**	**Coef**	**P-Value**
Clinical Node (cN0 vs cN1 & cN2)	-0.099	0.862	0.014	0.982
Clinical Size (cT2 & cT2 vs cT3)	-0.076	0.895	-0.222	0.736
Grade (Grade 1 & 2 vs 3)	1.030	0.080	0.923	0.141
ERBB2 (Low vs High)	-0.588	0.362	0.114	0.882
DCN Metagene (Low vs High)	-1.658	0.022	1.673	0.036

MDA	Univariate		Multivariate	
Variable	**Coef**	**P Value**	**Coef**	**P-Value**
Clinical Node (cN0 vs cN1 & cN2)	0.789	0.231	1.844	0.044
Clinical Size (cT2 & cT2 vs cT3)	-0.357	0.540	-1.698	0.040
Grade (Grade 1 & 2 vs 3)	1.191	0.181	1.009	0.333
ERBB2 (Neg vs Pos)	0.577	0.336	0.489	0.477
DCN Metagene (Low vs High)	1.217	0.043	1.605	0.039

The Applicants' results show that the predictive value of the stromal metagene is independent of other clinical variables and this signature remains predictive when tested with tumours restricted to the basal molecular-type.

Using the output of the multiple regression model, all 10317 genes of the NKI-EMC datasets were ranked according to the strength of their association with the stromal representative gene decorin (DCN). Rather than using the top 50 genes, as done previously, the Applicants constructed a series of metagenes from non-overlapping groups of 15 genes (see Figure 5). Nine of the first 12 metagenes, involving a total of 180 genes, gave an AUC significantly greater than 0.5, with a maximum of 0.70 in the EORTC dataset. The Applicants found out that the predictive value f the stromal metagene is not limited to the 50 genes but can be broaden to other genes of the same biological process (see Table 4).

The stromal metagene is detecting intrinsically more aggressive tumours that would have more chance to resist chemotherapy (i.e. npCR). To verify this possibility, the impact of the stromal metagene on relapse-free survival was investigated in three cohorts of patients, one

treated with chemotherapy, the two others not. As a positive control, a proliferation metagene known to be associated with high grade and poor survival [20] was used. Patients in the NKI and EMC studies who did not receive either chemotherapy or hormonal therapy were taken as the reference untreated population. Patients were split into two equally sized groups based on the value of their metagenes. The proliferation metagene divided patients into 2 groups with significantly different survival in both datasets (Figure 6a and 6b) while the stromal metagene does not (Figure 6d and 6e). In the cohort of breast cancer patients that received adjuvant chemotherapy [23] the group of patients having higher stromal metagene signature showed significantly shorter relapse free survival (Figure 6f). This observation shows that the stromal metagene is a predictive rather than a prognostic signature and reinforces the hypothesis that the main association is between high stromal content and resistance to chemotherapy.

The Applicants have tested the ability of other published stroma-related gene lists to predict reponse to chemotherapy. West et al [24] compared two neoplastic conditions with fibroblastic features: desmoid-type fibromatosis (DTF) and solitary fibrous tumours (SFT). The former arise in patients with mutations in the adenomatous polyposis coli gene that leads to activation of Wnt signalling in these tumours. The set of genes up-regulated in DTF is enriched in genes involved in the fibrotic response, including genes for matrix remodelling, whereas the set up-regulated in SFT includes more basement membrane genes and lacks remodelling genes [24].

The DTF but not the SFT signature predicts response in the EORTC and MDA datasets (see Table 3). The Applicants' stromal metagene is closely related to the DTF metagene but not to the SFT metagene, with correlation coefficients respectively of 0.91 and 0.41 in the EORTC dataset (Figure 7 a and b). GSEA confirmed a significant relationship ($p < 0.01$) between the stromal metagene and a published list of Wnt target genes (Figure 7c). This shows that it is not just a difference in the amount of fibroblasts but a difference in the type of stroma that is associated with response.

Table 3. Prediction of pCR using stroma related gene list from the literature. The 95% confidence intervals were calculated from the AUC distribution over 1000 cycles of bootstrapping.

Table 3

Name	Nb Genes	Reference	Eortc - AUC [C.I. 95%]	MDA- AUC [C.I. 95%]
DTF	415	West et al. 2005	0.66 [0.51-0.79]	0.67 [0.51-0.81]
SFT	194	West et al. 2005	0.52 [0.38-0.66]	0.53 [0.37-0.69]
LDM-Stroma	143	Finak et al. 2006	0.53 [0.37-0.68]	0.62 [0.45-0.76]

Further evidence linking the stromal metagene with the stromal tissue in tumours was obtained from gene expression profiles of samples in an ongoing study of colon carcinoma. Well-defined tumour epithelial tissue and stromal tissue were isolated from tumours of three patients by laser dissection microscopy (LDM). Cancer-associated fibroblasts (CAFs) were also isolated. mRNA from the LDM material and CAFs was hybridised to microarrays. To confirm that the microarrayed samples showed the expected expression patterns, genes known to be expressed selectively in normal mammary epithelial cells (Figure 7d, NBE) and normal mammary stroma (Figure 7d NBS) were identified from the literature [25]. Figure 7d shows the "epithelial : stromal" expression ratio of these genes in the microarrayed samples. A negative ratio indicates that genes are more strongly expressed in stromal compared to epithelial tissues. Normal mammary epithelial and stromal genes show positive and negative mean log ratio values of 0.95 and -2.41, respectively. Almost all of the 50 genes included in the stromal metagene were more highly represented in microdissected stromal than epithelial tissue (mean $\log_2$ difference –4.68; $p < 0.001$). The DTF gene list shows a similar pattern. To further confirm that genes included in the stromal metagene are expressed by fibroblasts, rather than some other cellular component of the tumour stroma, the analysis was repeated with purified CAFs in place of microdissected stroma (see Figure 8). The stromal-metagene is shown to be significantly associated with the CAF fraction with a mean $\log_2$ difference of -4.12, $p < 0.001$. The Applicants found out that the stromal metagene does indeed measure the amount of the activated stroma in the tumour.

Thus the Applicants have identified stromal gene signatures that predict poor pathological response to anthracycline-based neo-adjuvant chemotherapy in two independent datasets. These signatures were shown to be a reflection of the activation state of the tumour stroma. Preferably the anthracycline-based neo-adjuvant chemotherapy is the neoadjuvant therapy with both FEC and T-FAC. The Applicants identified stromal genes signature that influences

the response of cancers to anthracycline-based neo-adjuvant chemotherapy. Preferably, the breast cancers are estrogen (ER) and progesterone receptor negative cancers.

The Applicants have identified several specific combinations of stromal genes (Table 4), which are part of the stromal metagene and which are biomarkers for chemosensitivity of cancer subjects to the anthracycline-based neo-adjuvant chemotherapy. Quantified mRNA levels for genes included in the expression signature within tumour biopsies are associated with the cancer sensitivity to chemotherapy. More precisely, high levels of mRNA for genes of the expression signature is indicative of non pathological complete response (resistance) to anthracycline-based neo-adjuvant chemotherapy while low mRNA levels for these genes is associated with pathological complete response (sensitivity) to this treatment.

Table 4 Genes associated with the tumour sensitivity to chemotherapy. The list contains official gene symbol defined by the Human Genome Organisation (HUGO; http://www.hugo-international.org/) and gene bank accession numbers http://www.genenames.org/cgi-bin/hgnc_search.pl) :

DCN SEQ ID NO:1	DECORIN	AF138300 NM_133507	RODH	HYDROXYSTEROID (17-BETA) DEHYDROGENASE 6 HOMOLOG (MOUSE)	AF016509 NM_003725
CSPG2	VERSICAN	X15998 NM_004385	ADAMTS2	ADAM METALLOPEPTIDASE WITH THROMBOSPONDIN TYPE 1 MOTIF, 2	AJ003125 NM_014244
CDH11 SEQ ID NO:2	CADHERIN 11, TYPE 2, OB-CADHERIN (OSTEOBLAST)	D21255 NM_033664	SART2	DERMATAN SULFATE EPIMERASE	AF098066 NM_013352
COL3A1 SEQ ID NO:3	COLLAGEN, TYPE III, ALPHA 1	X15332 NM_000090	SCGF	C-TYPE LECTIN DOMAIN FAMILY 11, MEMBER A	AF087658 NM_002975
FAP SEQ ID NO:4	FIBROBLAST ACTIVATION PROTEIN, ALPHA	U09278	PPIC	PEPTIDYLPROLYL ISOMERASE C (CYCLOPHILIN C)	S71018 NM_000943
SERPINF1 SEQ ID NO:5	SERPIN PEPTIDASE INHIBITOR, CLADE F (ALPHA-2 ANTIPLASMIN, PIGMENT EPITHELIUM DERIVED FACTOR), MEMBER 1	M76979 NM_002615	PDGFC	PLATELET DERIVED GROWTH FACTOR C	AF091434
FBN1 SEQ ID NO:6	FIBRILLIN 1	X63556	POSTN	PERIOSTIN, OSTEOBLAST SPECIFIC FACTOR	D13665 NM_006475
PDGFRL SEQ ID NO:7	PLATELET-DERIVED GROWTH FACTOR RECEPTOR-LIKE	D37965 NM_006207	CBFA2T1	RUNT-RELATED TRANSCRIPTION FACTOR 1; TRANSLOCATED TO, 1 (CYCLIN D-RELATED)	D43638 NM_004349 NM_175635
CTSK SEQ ID NO:8	CATHEPSIN K	BC016058 NM_000396	PLAUR	PLASMINOGEN ACTIVATOR, UROKINASE RECEPTOR	
PRSS11	HTRA SERINE PEPTIDASE 1	AF097709 NM_002775	LAMC1	LAMININ, GAMMA 1 (FORMERLY LAMB2)	J03202 NM_002293
ASPN SEQ ID NO:9	ASPORIN	AF316824 NM_017680	MATN3	MATRILIN 3	AJ001047 NM_002381
SPARC SEQ ID NO:10	SECRETED PROTEIN, ACIDIC, CYSTEINE-RICH (OSTEONECTIN)	NM_003118	ACVR1	ACTIVIN A RECEPTOR, TYPE I	NM_001105
COL5A2	COLLAGEN, TYPE V, ALPHA 2	Y14690	COMP		

SEQ ID NO:11		NM_000393			
LOXL1 SEQ ID NO:12	LYSYL OXIDASE-LIKE 1	L21186 NM_005576	ANGPTL2	ANGIOPOIETIN-LIKE 2	AF125175 NM_012098
MMP2 SEQ ID NO:13	MATRIX METALLOPEPTIDASE 2 (GELATINASE A, 72KDA GELATINASE, 72KDA TYPE IV COLLAGENASE)		ID3	INHIBITOR OF DNA BINDING 3, DOMINANT NEGATIVE HELIX-LOOP-HELIX PROTEIN	X69111 NM_002167
SPON1 SEQ ID NO:14	SPONDIN 1, EXTRACELLULAR MATRIX PROTEIN	AB018305	RIN2	RAS AND RAB INTERACTOR 2	AB050339
SFRP4 SEQ ID NO:15	SECRETED FRIZZLED-RELATED PROTEIN 4	AF026692 NM_003014	BAPX1	NK3 HOMEOBOX 2	AF009801
ITGBL1 SEQ ID NO:16	INTEGRIN, BETA-LIKE 1 (WITH EGF-LIKE REPEAT DOMAINS)	AF072752 NM_004791	TPM2	TROPOMYOSIN 2 (BETA)	NM_003289
CALD1 SEQ ID NO:17	CALDESMON 1	M64110 NM_033138	PDLIM2	PDZ AND LIM DOMAIN 2 (MYSTIQUE)	AY007729
COPZ2 SEQ ID NO:18	COATOMER PROTEIN COMPLEX, SUBUNIT ZETA 2	AB037938 NM_016429	KDELC1	KDEL (LYS-ASP-GLU-LEU) CONTAINING 1	BC001297
MFAP2 SEQ ID NO:19	MICROFIBRILLAR-ASSOCIATED PROTEIN 2	BC015039 NM_002403	COL5A1	COLLAGEN, TYPE V, ALPHA 1	D90279 NM_000093
ANGPTL2 SEQ ID NO:20	ANGIOPOIETIN-LIKE 2	AF125175 NM_012098	TGFB1I1	TRANSFORMING GROWTH FACTOR BETA 1 INDUCED TRANSCRIPT 1	AB007836
PLAU SEQ ID NO:21	PLASMINOGEN ACTIVATOR, UROKINASE	BC013575 NM_002658	PRRX2	PAIRED RELATED HOMEOBOX 2	AF061970 NM_016307
COL1A2 SEQ ID NO:22	COLLAGEN, TYPE I, ALPHA 2	Z74616 NM_000089	TPST1	TYROSYLPROTEIN SULFOTRANSFERASE 1	AF038009 NM_003596
LRRC17 SEQ ID NO:23	LEUCINE RICH REPEAT CONTAINING 17	U32907 NM_005824	AKAP12	A KINASE (PRKA) ANCHOR PROTEIN 12	U81607

C1QTNF3 SEQ ID NO:24	C1Q AND TUMOR NECROSIS FACTOR RELATED PROTEIN 3	AF329837 NM_030945	COL18A1	COLLAGEN, TYPE XVIII, ALPHA 1	
SNAI2 SEQ ID NO:25	SNAIL HOMOLOG 2 (DROSOPHILA)	U97060 NM_003068	PARVA	PARVIN, ALPHA	AF237771 NM_018222
PCOLCE SEQ ID NO:26	PROCOLLAGEN C-ENDOPEPTIDASE ENHANCER	L33799 NM_002593	COX7A1	CYTOCHROME C OXIDASE SUBUNIT VIIA POLYPEPTIDE 1 (MUSCLE)	BC002757 NM_001864
POSTN SEQ ID NO:27	PERIOSTIN, OSTEOBLAST SPECIFIC FACTOR	D13665 NM_006475	TIMP3	TIMP METALLOPEPTIDASE INHIBITOR 3	NM_000362
ECM2 SEQ ID NO:28	EXTRACELLULAR MATRIX PROTEIN 2, FEMALE ORGAN AND ADIPOCYTE SPECIFIC	AB011792 NM_001393	LRRC17	LEUCINE RICH REPEAT CONTAINING 17	U32907 NM_005824
FBLN1 SEQ ID NO:29	FIBULIN 1	NM_006486	PLSCR4	PHOSPHOLIPID SCRAMBLASE 4	AF199023 NM_020353
ADAM12 SEQ ID NO:30	ADAM METALLOPEPTIDASE DOMAIN 12	AF023476	NBL1	NEUROBLASTOMA, SUPPRESSION OF TUMORIGENICITY 1	NM_005380
MMP11 SEQ ID NO:31	MATRIX METALLOPEPTIDASE 11 (STROMELYSIN 3)	NM_005940	TGFB3	TRANSFORMING GROWTH FACTOR, BETA 3	NM_003239
AEBP1 SEQ ID NO:32	AE BINDING PROTEIN 1	D86479 NM_001129	FOXF2	FORKHEAD BOX F2	U13220
PDGFRB SEQ ID NO:33	PLATELET-DERIVED GROWTH FACTOR RECEPTOR, BETA POLYPEPTIDE	M21616	OLFML3	OLFACTOMEDIN-LIKE 3	AF201945 NM_020190
GAS1 SEQ ID NO:34	GROWTH ARREST-SPECIFIC 1	NM_002048	ZNF423	ZINC FINGER PROTEIN 423	AB018303 NM_015069
COL6A3 SEQ ID NO:35	COLLAGEN, TYPE VI, ALPHA 3	X52022 NM_004369	COL11A1	COLLAGEN, TYPE XI, ALPHA 1	J04177 NM_080630
RARRES2	RETINOIC ACID RECEPTOR RESPONDER (TAZAROTENE INDUCED) 2	U77594	DFNA5	DEAFNESS, AUTOSOMAL DOMINANT 5	AF007790 NM_004403

SEQ ID NO:36					
COL6A1 SEQ ID NO:37	COLLAGEN, TYPE VI, ALPHA 1	M20776 NM_001848	TGFBI	TRANSFORMING GROWTH FACTOR, BETA-INDUCED, 68KDA	M77349
C1R	COMPLEMENT COMPONENT 1, R SUBCOMPONENT	M14058	FLJ10134	TRANSMEMBRANE PROTEIN 45A	AK000996 NM_018004
NDN SEQ ID NO:38	NECDIN HOMOLOG (MOUSE)	U35139 NM_002487	OLFML2A	OLFACTOMEDIN-LIKE 2A	AK092255 NM_182487
TGFB3 SEQ ID NO:39	TRANSFORMING GROWTH FACTOR, BETA 3	NM_003239	ISLR	IMMUNOGLOBULIN SUPERFAMILY CONTAINING LEUCINE-RICH REPEAT	AB003184 NM_005545
LRP1 SEQ ID NO:40	LOW DENSITY LIPOPROTEIN-RELATED PROTEIN 1 (ALPHA-2-MACROGLOBULIN RECEPTOR)	X13916 NM_002332	TNFAIP6	TUMOR NECROSIS FACTOR, ALPHA-INDUCED PROTEIN 6	NM_007115
COL10A1 SEQ ID NO:41	COLLAGEN, TYPE X, ALPHA 1		LAMA4	LAMININ, ALPHA 4	NM_001105206
DPYSL3 SEQ ID NO:42	DIHYDROPYRIMIDINASE-LIKE 3	D78014 NM_001387	COL5A3	COLLAGEN, TYPE V, ALPHA 3	AF177941 NM_015719
OLFML2B SEQ ID NO:43	OLFACTOMEDIN-LIKE 2B	BX648975 NM_015441	FBLN1	FIBULIN 1	NM_006486
MMP14 SEQ ID NO:44	MATRIX METALLOPEPTIDASE 14 (MEMBRANE-INSERTED)		SPON2	SPONDIN 2, EXTRACELLULAR MATRIX PROTEIN	AB027466
DACT1 SEQ ID NO:45	DAPPER, ANTAGONIST OF BETA-CATENIN, HOMOLOG 1 (XENOPUS LAEVIS)	AF251079 NM_016651	MMP13	MATRIX METALLOPEPTIDASE 13 (COLLAGENASE 3)	X75308 NM_002427
MGC3047			CNIH3	CORNICHON HOMOLOG 3 (DROSOPHILA)	AF070524 NM_152495
THBS2 SEQ ID NO:46	THROMBOSPONDIN 2	NM_003247	WNT2	WINGLESS-TYPE MMTV INTEGRATION SITE FAMILY MEMBER 2	X07876 NM_003391
CYR61	CYSTEINE-RICH, ANGIOGENIC INDUCER, 61	AF031385 NM_001554	ZFPM2	ZINC FINGER PROTEIN, MULTITYPE 2	AF119334

EGR1	EARLY GROWTH RESPONSE 1	M62829 NM_001964	DAF	CD55 MOLECULE, DECAY ACCELERATING FACTOR FOR COMPLEMENT (CROMER BLOOD GROUP)	BC001288 NM_000574
CTGF	CONNECTIVE TISSUE GROWTH FACTOR	X78947 NM_001901	LEPRE1	LEUCINE PROLINE-ENRICHED PROTEOGLYCAN (LEPRECAN) 1	AK027648 NM_022356
DUSP1	DUAL SPECIFICITY PHOSPHATASE 1	X68277 NM_004417	BHLHB3	BASIC HELIX-LOOP-HELIX DOMAIN CONTAINING, CLASS B, 3	AI3044088
ZFP36	ZINC FINGER PROTEIN 36, C3H TYPE, HOMOLOG (MOUSE)	M63625	FEZ1	FASCICULATION AND ELONGATION PROTEIN ZETA 1 (ZYGIN I)	U60060 NM_005103
FOS	V-FOS FBJ MURINE OSTEOSARCOMA VIRAL ONCOGENE HOMOLOG	K00650 NM_005252	TUBB6	TUBULIN, BETA 6	AK001295 NM_032525
EGR2	EARLY GROWTH RESPONSE 2 (KROX-20 HOMOLOG, DROSOPHILA)	BC035625 NM_000399	ST5	SUPPRESSION OF TUMORIGENICITY 5	U15131 NM_005418
ITGB5	INTEGRIN, BETA 5	J05633 NM_002213	ADAMTS5	ADAM METALLOPEPTIDASE WITH THROMBOSPONDIN TYPE 1 MOTIF, 5	AF142099
PLEKHC1	FERMITIN FAMILY HOMOLOG 2 (DROSOPHILA)	Z24725 NM_006832	HOM-TES-103	INTERMEDIATE FILAMENT FAMILY ORPHAN 1	AF124432 NM_080730
FOSB	FBJ MURINE OSTEOSARCOMA VIRAL ONCOGENE HOMOLOG B	NM_006732	SERPINH1	SERPIN PEPTIDASE INHIBITOR, CLADE H (HEAT SHOCK PROTEIN 47), MEMBER 1, (COLLAGEN BINDING PROTEIN 1)	X61598 NM_004353
GEM	GTP BINDING PROTEIN OVEREXPRESSED IN SKELETAL MUSCLE	NM_181702	AD-017	GLYCOSYLTRANSFERASE 8 DOMAIN CONTAINING 1	AY358579 NM_152932
DACT1	DAPPER, ANTAGONIST OF BETA-CATENIN, HOMOLOG 1 (XENOPUS LAEVIS)	AF251079 NM_016651	VEGFC	VASCULAR ENDOTHELIAL GROWTH FACTOR C	BC035212 NM_005429
ATF3	ACTIVATING TRANSCRIPTION FACTOR 3	L19871 NM_001674	CNN2	CALPONIN 2	D83735 NM_004368
SPARC	SECRETED PROTEIN, ACIDIC, CYSTEINE-RICH (OSTEONECTIN)	NM_003118	ADARB1	ADENOSINE DEAMINASE, RNA-SPECIFIC, B1 (RED1 HOMOLOG RAT)	U76420 NM_015833
LAMB1	LAMININ, BETA 1	M61916 NM_002291	RNF144	RING FINGER PROTEIN 144A	D79983 NM_014746
FSTL1	FOLLISTATIN-LIKE 1	U06863 NM_007085	LOX	LYSYL OXIDASE	

RECK	REVERSION-INDUCING-CYSTEINE-RICH PROTEIN WITH KAZAL MOTIFS	E13833	BGN	BIGLYCAN	AK092954 NM_001711
COL6A3	COLLAGEN, TYPE VI, ALPHA 3	X52022 NM_004369	LHFP	LIPOMA HMGIC FUSION PARTNER	AF098807 NM_005780
NR4A1	NUCLEAR RECEPTOR SUBFAMILY 4, GROUP A, MEMBER 1	L13740	THBS1	THROMBOSPONDIN 1	NM_003246
JAM3	JUNCTIONAL ADHESION MOLECULE 3	AF356518 NM_032801	ITGA5	INTEGRIN, ALPHA 5 (FIBRONECTIN RECEPTOR, ALPHA POLYPEPTIDE)	
PDGFRA	PLATELET-DERIVED GROWTH FACTOR RECEPTOR, ALPHA POLYPEPTIDE	D50001 NM_006206	SLC15A3	SOLUTE CARRIER FAMILY 15, MEMBER 3	AB020598 NM_016582
RGS2	REGULATOR OF G-PROTEIN SIGNALING 2, 24KDA	L13463 NM_002923	PPGB	CATHEPSIN A	M22960 NM_000308
RGS16	REGULATOR OF G-PROTEIN SIGNALING 16	U70426 NM_002928	LGALS1	LECTIN, GALACTOSIDE-BINDING, SOLUBLE, 1	NM_002305
ACTA2	ACTIN, ALPHA 2, SMOOTH MUSCLE, AORTA	X13839 NM_001613	SERPING1	SERPIN PEPTIDASE INHIBITOR, CLADE G (C1 INHIBITOR), MEMBER 1	X54486 NM_000062
SERPINF1	SERPIN PEPTIDASE INHIBITOR, CLADE F (ALPHA-2 ANTIPLASMIN, PIGMENT EPITHELIUM DERIVED FACTOR), MEMBER 1	M76979 NM_002615	TPM1	TROPOMYOSIN 1 (ALPHA)	AB209041 NM_001018004
NNMT	NICOTINAMIDE N-METHYLTRANSFERASE	U08021 NM_006169	CD9	CD9 MOLECULE	M38690
DPYSL3	DIHYDROPYRIMIDINASE-LIKE 3	D78014 NM_001387	EGFR	EPIDERMAL GROWTH FACTOR RECEPTOR (ERYTHROBLASTIC LEUKEMIA VIRAL (V-ERB-B) ONCOGENE HOMOLOG, AVIAN)	NM_005228
COL5A2	COLLAGEN, TYPE V, ALPHA 2	Y14690 NM_000393	KRT18	KERATIN 18	NM_199187
IL6	INTERLEUKIN 6 (INTERFERON, BETA 2)	M18403	CD24	CD24 MOLECULE	
TIMP2	TIMP METALLOPEPTIDASE INHIBITOR 2	NM_003255	KRT7	KERATIN 7	NM_005556
THBS2	THROMBOSPONDIN 2	NM_003247	CNN3	CALPONIN 3, ACIDIC	BC025372 NM_001839
SPRY1	SPROUTY HOMOLOG 1, ANTAGONIST OF FGF SIGNALING (DROSOPHILA)	AF041037	ANXA3	ANNEXIN A3	M63310 NM_005139

WO 2009/074968 PCT/IB2008/055252

SRPX	SUSHI-REPEAT-CONTAINING PROTEIN, X-LINKED	U78093 NM_006307	ZNF146	ZINC FINGER PROTEIN 146	X70394 NM_007145
MYLK	MYOSIN LIGHT CHAIN KINASE	X85337 NM_053025	NME7	NON-METASTATIC CELLS 7, PROTEIN EXPRESSED IN (NUCLEOSIDE-DIPHOSPHATE KINASE)	AF153191 NM_013330
DLC1	DELETED IN LIVER CANCER 1	AF035119 NM_182643, NM_006094	DCBLD2	DISCOIDIN, CUB AND LCCL DOMAIN CONTAINING 2	NM_080927
LOC83468			MGST1	MICROSOMAL GLUTATHIONE S-TRANSFERASE 1	U46494
PRKD1	PROTEIN KINASE D1	NM_002742	COTL1	COACTOSIN-LIKE 1 (DICTYOSTELIUM)	L54057 NM_021149
PRSS11	HTRA SERINE PEPTIDASE 1	AF097709 NM_002775	MAL2	MAL, T-CELL DIFFERENTIATION PROTEIN 2	AL117612 NM_052886
EMP1	EPITHELIAL MEMBRANE PROTEIN 1	U43916 NM_001423	CRIM1	CYSTEINE RICH TRANSMEMBRANE BMP REGULATOR 1 (CHORDIN-LIKE)	AF168681 NM_016441
SPON1	SPONDIN 1, EXTRACELLULAR MATRIX PROTEIN	AB018305	MIRN21	MICRORNA 21	NT_010783
COL1A2	COLLAGEN, TYPE I, ALPHA 2	Z74616 NM_000089	TMEPAI	PROSTATE TRANSMEMBRANE PROTEIN, ANDROGEN INDUCED 1	AF224278 NM_020182
AEBP1	AE BINDING PROTEIN 1	D86479 NM_001129	FN1	FIBRONECTIN 1	NM_212475
TAGLN	TRANSGELIN	M95787 NM_001001522	CLECSF2	C-TYPE LECTIN DOMAIN FAMILY 2, MEMBER B	X96719 NM_005127
FBLN2	FIBULIN 2	X82494 NM_001004019	DDR2	DISCOIDIN DOMAIN RECEPTOR TYROSINE KINASE 2	AK095975 NM_006182
DAB2	DISABLED HOMOLOG 2, MITOGEN-RESPONSIVE PHOSPHOPROTEIN (DROSOPHILA)	U53446 NM_001343	MFAP2	MICROFIBRILLAR-ASSOCIATED PROTEIN 2	BC015039 NM_002403
FLJ10357			GARP	CYCLIC NUCLEOTIDE GATED CHANNEL BETA 1	AF042498 NM_001297
ITGBL1	INTEGRIN, BETA-LIKE 1 (WITH EGF-LIKE REPEAT DOMAINS)	AF072752 NM_004791	NINJ2	NINJURIN 2	AF205633 NM_016533
EFEMP2	EGF-CONTAINING FIBULIN-LIKE EXTRACELLULAR MATRIX PROTEIN 2	AF109121 NM_016938	DOC1		

C1S	COMPLEMENT COMPONENT 1, S SUBCOMPONENT	NM_001734	CILP	CARTILAGE INTERMEDIATE LAYER PROTEIN, NUCLEOTIDE PYROPHOSPHO-HYDROLASE	AY358904 NM_003613
FLRT2	FIBRONECTIN LEUCINE RICH TRANSMEMBRANE PROTEIN 2	AF169676	EMILIN1	ELASTIN MICROFIBRIL INTERFACER 1	AF088916 NM_007046
COL15A1	COLLAGEN, TYPE XV, ALPHA 1	L25286 NM_001855	BMP1	BONE MORPHOGENETIC PROTEIN 1	NM_006132
PMP22	PERIPHERAL MYELIN PROTEIN 22	D11428 NM_000304	MAFB	V-MAF MUSCULOAPONEUROTIC FIBROSARCOMA ONCOGENE HOMOLOG B (AVIAN)	AF134157
PDGFRB	PLATELET-DERIVED GROWTH FACTOR RECEPTOR, BETA POLYPEPTIDE	M21616	GFPT2	GLUTAMINE-FRUCTOSE-6-PHOSPHATE TRANSAMINASE 2	AB016789 NM_005110
CD200	CD200 MOLECULE		FKBP14	FK506 BINDING PROTEIN 14, 22 KDA	AK000738 NM_017946
COL3A1	COLLAGEN, TYPE III, ALPHA 1	X15332 NM_000090	LMCD1	LIM AND CYSTEINE-RICH DOMAINS 1	AF169284 NM_014583
NID2	NIDOGEN 2 (OSTEONIDOGEN)	AB009799	SPOCK	SPARC/OSTEONECTIN, CWCV AND KAZAL-LIKE DOMAINS PROTEOGLYCAN (TESTICAN) 1	AF231124 NM_004598
DKK3	DICKKOPF HOMOLOG 3 (XENOPUS LAEVIS)	AF177396 NM_013253	ZFHX4	ZINC FINGER HOMEOBOX 4	NM_024721
GRP	GASTRIN-RELEASING PEPTIDE	NM_002091	WISP2	WNT1 INDUCIBLE SIGNALING PATHWAY PROTEIN 2	AF100780 NM_003881
ARK5	NUAK FAMILY, SNF1-LIKE KINASE, 1	AB011109 NM_014840	ADCY7	ADENYLATE CYCLASE 7	D25538
COL10A1	COLLAGEN, TYPE X, ALPHA 1		VIM	VIMENTIN	M14144 NM_003380
SNAI2	SNAIL HOMOLOG 2 (DROSOPHILA)	U97060 NM_003068	KIAA1199	KIAA1199	BC020256 NM_018689
NID	NIDOGEN 1	BC045606 NM_002508	RAI14	RETINOIC ACID INDUCED 14	AB037755 NM_015577
TMAP1	MATRIX-REMODELLING ASSOCIATED 7	BC053983 NM_001008529	CDC10	SEPTIN 7	S72008 NM_001788
PLS3	PLASTIN 3 (T ISOFORM)	L05491	SULF1	SULFATASE 1	AB029000

WO 2009/074968 PCT/IB2008/055252

					NM_015170
NOX4	NADPH OXIDASE 4	AF254621 NM_016931	HNT	RAS RESPONSIVE ELEMENT BINDING PROTEIN 1	U26914
C1QTNF3	C1Q AND TUMOR NECROSIS FACTOR RELATED PROTEIN 3	AF329837 NM_030945	SHOX2	SHORT STATURE HOMEOBOX 2	AJ002368
COPZ2	COATOMER PROTEIN COMPLEX, SUBUNIT ZETA 2	AB037938 NM_016429	MN1	MENINGIOMA (DISRUPTED IN BALANCED TRANSLOCATION) 1	X82209 NM_002430
ECM2	EXTRACELLULAR MATRIX PROTEIN 2, FEMALE ORGAN AND ADIPOCYTE SPECIFIC	AB011792 NM_001393	RUNX2	RUNT-RELATED TRANSCRIPTION FACTOR 2	AF001450 NM_004348
CTSK	CATHEPSIN K	BC016058 NM_000396	PROS1	PROTEIN S (ALPHA)	NM_000313
LOXL1	LYSYL OXIDASE-LIKE 1	L21186 NM_005576	PTK7	PTK7 PROTEIN TYROSINE KINASE 7	AF447176
GAS6	GROWTH ARREST-SPECIFIC 6	NM_000820	FMO1	FLAVIN CONTAINING MONOOXYGENASE 1	M64082 NM_002021
PDGFRL	PLATELET-DERIVED GROWTH FACTOR RECEPTOR-LIKE	D37965 NM_006207	CALD1	CALDESMON 1	M64110 NM_033138
CORIN	CORIN, SERINE PEPTIDASE	AF133845	NRN1	NEURITIN 1	AF136631
LOXL2	LYSYL OXIDASE-LIKE 2	U89942	HS3ST3A1	HEPARAN SULFATE (GLUCOSAMINE) 3-O-SULFOTRANSFERASE 3A1	AF105376 NM_006042
NAP1L3	NUCLEOSOME ASSEMBLY PROTEIN 1-LIKE 3	NM_004538	CRHR1	CORTICOTROPIN RELEASING HORMONE RECEPTOR 1	L23332
F13A1	COAGULATION FACTOR XIII, A1 POLYPEPTIDE	M14539	TNFSF4	TUMOR NECROSIS FACTOR (LIGAND) SUPERFAMILY, MEMBER 4	D90224
MITF	MICROPHTHALMIA-ASSOCIATED TRANSCRIPTION FACTOR	NM_198159	PDLIM7	PDZ AND LIM DOMAIN 7 (ENIGMA)	BC001093 NM_005451
FBN1	FIBRILLIN 1	X63556	EVC	ELLIS VAN CREVELD SYNDROME	AF216184
HEPH	HEPHAESTIN	AB014598 NM_138737	MSX1	MSH HOMEOBOX 1	M97676
IGFBP7	INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN 7	S75725	PCSK5	PROPROTEIN CONVERTASE SUBTILISIN/KEXIN TYPE 5	

RAB31	RAB31, MEMBER RAS ONCOGENE FAMILY	U59877	LEPREL2	LEPRECAN-LIKE 2	U47926 NM_014262
INHBA	INHIBIN, BETA A		LRP1	LOW DENSITY LIPOPROTEIN-RELATED PROTEIN 1 (ALPHA-2-MACROGLOBULIN RECEPTOR)	X13916 NM_002332
SRPX2	SUSHI-REPEAT-CONTAINING PROTEIN, X-LINKED 2	AF393649 NM_014467	GALNT1	UDP-N-ACETYL-ALPHA-D-GALACTOSAMINE: POLYPEPTIDE N-ACETYLGALACTOSAMINYL-TRANSFERASE 1 (GALNAC-T1)	NM_020474
AXL	AXL RECEPTOR TYROSINE KINASE	M76125	PLA2G5	PHOSPHOLIPASE A2, GROUP V	U03090 NM_000929
LUM	LUMICAN	BT006707 NM_002345	SPARCL1	SPARC-LIKE 1 (HEVIN)	X86693
GREM1	GREMLIN 1, CYSTEINE KNOT SUPERFAMILY, HOMOLOG (XENOPUS LAEVIS)	NM_013372	ANXA1	ANNEXIN A1	X05908 NM_000700
PCOLCE	PROCOLLAGEN C-ENDOPEPTIDASE ENHANCER	L33799 NM_002593	WBP5	WW DOMAIN BINDING PROTEIN 5	BC023544 NM_016303
PLXDC1	PLEXIN DOMAIN CONTAINING 1	AF279144 NM_020405	GNG11	GUANINE NUCLEOTIDE BINDING PROTEIN (G PROTEIN), GAMMA 11	NM_004126
LIMS2	LIM AND SENESCENT CELL ANTIGEN-LIKE DOMAINS 2	AF520987 NM_017980	NT5E	5'-NUCLEOTIDASE, ECTO (CD73)	X55740
KIAA0992	PALLADIN, CYTOSKELETAL ASSOCIATED PROTEIN	AB023209 NM_016081	LTBP2	LATENT TRANSFORMING GROWTH FACTOR BETA BINDING PROTEIN 2	NM_000428
CTSB	CATHEPSIN B	M14221 NM_147780	FST	FOLLISTATIN	M19481 NM_013409
MFAP5	MICROFIBRILLAR ASSOCIATED PROTEIN 5	AK124368 NM_003480	THY1	THY-1 CELL SURFACE ANTIGEN	M11749 NM_006288

Moreover, this expression signature predicts resistance to each of individual drugs given in the FEC chemotherapy regimen (i.e. Cyclophosphamide, Adriamycin/Doxorubicin and Fluorouracil). The gene signature predicts chemotherapy regimen composed of drugs having similar mechanism of action such as other pyrimidine analogues (ex: *Capecitabine, Cytarabine, Floxuridine, Gemcitabine*), purine analogues (*Cladribine, Clofarabine, Fludarabine, Mercaptopurine, Pentostatin, Thioguanine*), Cytotoxic/antitumor antibiotics of the anthracyclin family (*Daunorubicin, Doxorubicin, Epirubicin, Idarubicin, Mitoxantrone, Valrubicin*) or nitrogen mustard agent (*Chlorambucil, Chlormethine, Cyclophosphamide, Ifosfamide, Melphalan*).

The present invention further provides for that mRNA levels of genes included in expression signatures was not associated in predicting pathological or clinical response for subjects treated by Epirubicin / Taxotere (ET) chemotherapy regimen. Thus the expression signature will not be associated to pathological or clinical response to drugs having similar mechanism of actions such as spindle poison/mitotic inhibitor of the taxanes family (Docetaxel, Paclitaxel). Therefore, the stromal metagene signature could serve also as a treatment indicator. Subject having a high stromal metagene signature should be treated with taxane based chemotherapy.

The stromal metagene was shown to be associated with relapse-free survival in a third independent study performed in an adjuvant setting. That the signature is not merely detecting a difference in the innate aggressiveness of cancers is confirmed since the stromal metagene signature was not associated with differences in relapse-free survival in the patients in the NKI-EMC datasets who had not been given adjuvant systemic therapy. Thus the stromal metagene signature only predicts response to therapy, meaning it is predictive rather than prognostic.

The mRNA levels of genes included in the gene signatures are also associated with a well define physiological activation state of the tissue surrounding of the tumour termed “stroma activation”. The present invention can also be used for any clinical situations where the state of “stroma activation” is required.

The Applicants’ results also show that the stroma within the tumour needs to be in an activated state in order to confer resistance to anthracycline-based neo-adjuvant

chemotherapy. Indeed neither the normal stromal expression profile nor the SFT signature was able to predict FEC response. Interestingly, the two signatures that produced a significant prediction (stroma and DTF) comprised genes that are more strongly associated with CAF purified from colon carcinomas (Figure 7). Technical limitations prevented the isolation of the stroma tissues by laser dissection microscopy (LDM) from breast tumours samples used for the microarray study. However, genes shown, by LDM, to be specifically expressed in normal breast stroma and epithelial compartments [25] were equally associated to their respective tumour compartments in colon carcinomas. Furthermore, the stromal genes signatures of the present invention were strongly associated with the stromal compartment of the colon carcinomas suggesting that phenotype modelled by the signature is specific to the stroma tissue regardless of the studied organ.

A precedent, and possible mechanism for a direct role for stroma as a resistance factor, is provided by studies showing that β4-integrin activation renders mammary epithelial cells resistant to apoptosis induction by a wide range of different treatments [26]. Hyaluronic acid was found to promote doxorubicin resistance in the MCF7 human mammary carcinoma cell line [27]. Adhesion of multiple myeloma cells to fibronectin was also shown to provide a survival advantage in the presence of doxorubicin [14]. The same group proved that adhesion to fibronectin by means of β1-integrins protects U937 human monocytic leukaemia cells from doxorubicin-induced DNA damage by reducing the activity and modifying the subcellular distribution of topoisomerase II [28]. Growth factors linked to the stromal compartment could also contribute to drug resistance. Fibroblast growth factors were shown to induce paclitaxel, doxorubicin and 5-fluorouracil resistance to human prostate PC3 tumours cells and rat MAT-LyLu tumours cells in vitro [29] while suramin, an inhibitor of FGF and other growth factors, sensitized these tumour cells to doxorubicin both in vitro and in vivo experimental models. Interestingly, despite including TOP2A, the target of the epirubicin, the proliferation metagene was not associated with response. This signature was shown to be functional as it successfully identified patients having the worst prognosis in both the NKI and EMC datasets (Figure 6). The proliferation metagene was not associated with response in the MDA dataset 0.63 [0.39-0.83]. However, a significant AUC (0.68[0.58-78]) was observed when all ER positive tumours were included in the MDA dataset. Therefore, proliferation is a prognostic but not a predictive signature. From these results, the Applicants discovered that the low response rate to chemotherapy observed in ER positive tumours is not explained by their

lower proliferation relative to ER negative tumours but rather due to differences intrinsic to their epithelial biology.

Compared to traditional supervised analysis, building a predictor from expression modules has several advantages. First, testing groups of co-expressed genes related to a defined biological process facilitates the interpretation of the results. Second, random effects in single measurements are tamed by averaging over redundant measurements (because metagene contains many similarly expressed genes), reducing the problem of chance associations between one gene and the response. Over a wide range, the choice of genes among the stromal module used to produce the metagene did not affect the result (Figure 5). This has as a practical consequence that a customised diagnostic kit would not have to include all the genes in the expression module and the genes themselves can be flexibly chosen. Third, the approach reduces the number of variables that are tested, which decreases the probability of observing a significant association by chance. Fourth, performing the multiple regressions on data from two gene expression studies that used different microarray platforms reduces the probability of observing spurious associations. Fifth, the use of metagenes eases cross-platform mapping because many genes are available within each metagene, and the loss of a few genes after mapping does not impair the predictive power of the metagene.

The present invention provides for a method for predicting the efficacy of cancer therapy in a subject who has undergone or is undergoing chemotherapy treatment for cancer, characterized in that said method comprises

(a) obtaining a stromal tissue sample from a subject,

(b) determining in said stromal tissue sample the expression values of at least two stromal genes and of at least two reference genes,

(c) defining stromal content (SC) from the expression values of step (b),

(d) comparing the stromal content (SC) with a reference threshold,

(e) predicting resistance to chemotherapy of said subject based on the step (d), wherein high stromal content is indicative of resistance to chemotherapy, while low stromal content is indicative of sensitivity to chemotherapy,

(f) adapting the treatment of said subject

Preferably said cancer is selected from the group comprising breast cancer, colon cancer, lung cancer, colorectal cancer, head and neck cancer, or ovarian cancer.

Preferably said chemotherapy is anthracycline-based neo-adjuvant chemotherapy and the most preferably said chemotherapy is selected from the group comprising fluorouracil, epirubicin and cyclophosphamide based chemotherapy or combination thereof (two or three compounds together), such as for example “FEC” chemotherapy.

The stromal tissue sample is taken from a tumour biopsy. "Sample" or "tissue sample" refers to a collection of similar cells obtained from a tissue of a subject or patient. The source of the tissue sample may be solid tissue as from a fresh, frozen and/or preserved organ or tissue sample or biopsy or aspirate; blood or any blood constituents; bodily fluids such as cerebral spinal fluid, amniotic fluid, peritoneal fluid, or interstitial fluid; or cells from any time in gestation or development of the subject. The tissue sample may contain compounds which are not naturally intermixed with the tissue in nature such as preservatives, anticoagulants, buffers, fixatives, nutrients, antibiotics, or the like. In one aspect of the invention, tissue samples or patient samples are fixed, particularly conventional formalin-fixed paraffin-embedded samples. Such samples are typically used in an assay for receptor complexes in the form of thin sections, e.g. 3-10 μm thick, of fixed tissue mounted on a microscope slide, or equivalent surface. Such samples also typically undergo a conventional re-hydration procedure, and optionally, an antigen retrieval procedure as a part of, or preliminary to, assay measurements.

Biopsy refers to the removal of a sample of tissue for purposes of diagnosis. For example, a biopsy is from a cancer or tumour, including a sample of tissue from an abnormal area or an entire tumour. A non-limiting list of different types of cancers include lymphoma, B cell lymphoma, T cell lymphoma, mycosis fungoides, Hodgkin's Disease, myeloid leukemia, bladder cancer, brain cancer, nervous system cancer, head and neck cancer, kidney cancer, lung cancers such as small cell lung cancer and non-small cell lung cancer, brain cancers such as neuroblastoma and glioblastoma, ovarian cancer, pancreatic cancer, prostate cancer, liver cancer, melanoma, squamous cell carcinomas, cervical carcinoma, breast cancer, renal cancer, genitourinary cancer, esophageal carcinoma, hematopoietic cancers, testicular cancer, or colon and rectal cancers.

The stromal genes of the present invention are selected from the group consisting of the genes of Table 4. Preferably said stromal genes are DCN, CSPG2, CDH11, ASPN, SPARC, ITGBL1, PLAU, COL1A2, SNAI2, POSTN and THBS2.

The reference genes are selected from the group consisting of the genes of Table 5. Preferably said reference genes are GAPDH, ACTB, TCF2, ZNF333, ADH6, FOXH1, TPX2 1, CENPA, BIRC5, TOP2A.

In another embodiment of the present invention, said reference genes are genes selected from the group consisting of the genes of Table 4 or Table 5 determined within stromal tissue of non-tumor reference biopsy.

Preferably the expression values of at least 5 stromal genes and of at least 5 reference genes are determined in the stromal tissue sample. The most preferably the expression values of at least 10 stromal genes and of at least 10 reference genes are determined in the stromal tissue sample.

The said subject is preferably a mammal and the most preferably a human.

Table 5. Reference genes not associated with the tumour sensitivity to chemotherapy. The list contains official gene symbol defined by the Human Genome Organisation (HUGO; http://www.hugo-international.org/) and gene bank accession numbers http://www.genenames.org/cgi-bin/hgnc_search.pl) :

TPX2 SEQ ID NO:47	MICROTUBULE-ASSOCIATED, HOMOLOG	AF098158	RUTBC2	SMALL G PROTEIN SIGNALING MODULATOR 1	XM_059318 AB075821
CENPA SEQ ID NO:48	CENTROMERE PROTEIN A	U14518 NM_001809	FLJ13941	MORN REPEAT CONTAINING 1	NM_024848 AK024003
PKMYT1 SEQ ID NO:49	PROTEIN KINASE, MEMBRANE ASSOCIATED TYROSINE/THREONINE 1	NM_004203 AK097642	PGLYRP1	PEPTIDOGLYCAN RECOGNITION PROTEIN 1	NM_005091 AF076483
KIF4A SEQ ID NO:50	KINESIN FAMILY MEMBER 4A	NM_012310 AF179308	C21orf90	CHROMOSOME 21 OPEN READING FRAME 90	AF426270
KIF2C SEQ ID NO:51	KINESIN FAMILY MEMBER 2C	NM_006845 U63743	ALPP	ALKALINE PHOSPHATASE, PLACENTAL (REGAN ISOZYME)	NM_001632 M14169
CCNB2 SEQ ID NO:52	CYCLIN B2	NM_004701 AF002822	BCL6B	B-CELL CLL/LYMPHOMA 6, MEMBER B	NM_181844 AI672318
CDCA8 SEQ ID NO:53	CELL DIVISION CYCLE ASSOCIATED 8	NM_018101 BG354581	MGC33962		
CDCA3 SEQ ID NO:54	CELL DIVISION CYCLE ASSOCIATED 3	NM_031299 BG354576	FLJ38792	TRANSMEMBRANE PROTEIN 105	NM_178520 AK096111
CDC2 SEQ ID NO:55	CELL DIVISION CYCLE 2	NM_001786 BC014563	MMPL1	MATRIX METALLOPEPTIDASE 25	NM_022468 AF145442
ESPL1 SEQ ID NO:56	EXTRA SPINDLE POLE BODIES HOMOLOG 1	NM_012291 D79987	KRTAP7-1	KERATIN ASSOCIATED PROTEIN 7-1	AJ457063
CCNB1 SEQ ID NO:57	CYCLIN B1	NM_031966 U22364	DDX4	DEAD (ASP-GLU-ALA-ASP) BOX POLYPEPTIDE 4	NM_024415 AF262962
NEK2 SEQ ID NO:58	NIMA (NEVER IN MITOSIS GENE A)-RELATED KINASE 2	NM_002497 U11050	LOC149773		
KIF20A SEQ ID NO:59	KINESIN FAMILY MEMBER 20A	NM_005733 AF070672	NPHP1	NEPHRONOPHTHISIS 1	NM_000272 AF023674
FOXM1 SEQ ID NO:60	FORKHEAD BOX M1	NM_021953 Y12773	LOC161394		
BRRN1	NON-SMC CONDENSIN I COMPLEX, SUBUNIT H	NM_015341 BC024211	HEMGN	HEMOGEN	NM_197978 AF228713
	RAC GTPASE ACTIVATING PROTEIN 1	NM_013277	FLJ12975		NM_024809

RACGAP1				TECTONIC FAMILY MEMBER 2	AK056924
STK6	AURORA KINASE A	NM_003600 BC001280	SLC23A3	SOLUTE CARRIER FAMILY 23 (NUCLEOBASE TRANSPORTERS), MEMBER 3	NM_144712 BC030243
TRIP13 SEQ ID NO:61	THYROID HORMONE RECEPTOR INTERACTOR 13	NM_004237 L40384	RNF31	RING FINGER PROTEIN 31	NM_017999 AK000973
PRC1 SEQ ID NO:62	PROTEIN REGULATOR OF CYTOKINESIS 1	NM_003981 AF044588	IL11	INTERLEUKIN 11	NM_000641 X58377
EXO1 SEQ ID NO:63	EXONUCLEASE 1	NM_006027 AF042282	FLJ43965	BEN DOMAIN CONTAINING 4	NM_207406 AK092951
CCNA2 SEQ ID NO:64	CYCLIN A2	NM_001237	BCL2L10	BCL2-LIKE 10 (APOPTOSIS FACILITATOR)	AF285092
DLG7	DISCS, LARGE (DROSOPHILA) HOMOLOG-ASSOCIATED PROTEIN 5	NM_014750 D13633	LOC283140		
CDC20 SEQ ID NO:65	CELL DIVISION CYCLE 20 HOMOLOG	NM_001255 U05340	LOC340107		
TYMS SEQ ID NO:66	THYMIDYLATE SYNTHETASE	NM_001071 X02308	LOC339146		
MELK SEQ ID NO:67	MATERNAL EMBRYONIC LEUCINE ZIPPER KINASE	NM_014791 D79997	LOC115110		
AURKB SEQ ID NO:68	AURORA KINASE B	NM_004217 AF004022	FLJ37964	THIOESTERASE SUPERFAMILY MEMBER 5	NM_182578 AK095283
ZWINT SEQ ID NO:69	ZW10 INTERACTOR	AF067656	CACNB4	CALCIUM CHANNEL, VOLTAGE-DEPENDENT, BETA 4 SUBUNIT	AF038852
BIRC5 SEQ ID NO:70	BACULOVIRAL IAP REPEAT-CONTAINING 5	NM_001168 U75285	GAGED2	X ANTIGEN FAMILY, MEMBER 1A	NM_020411
MAD2L1 SEQ ID NO:71	MAD2 MITOTIC ARREST DEFICIENT-LIKE 1	U65410	C19orf26	CHROMOSOME 19 OPEN READING FRAME 26	NM_152769 BC028156
TIMELESS	TIMELESS HOMOLOG	NM_003920 AF098162	RNF157	RING FINGER PROTEIN 157	XM_290732 AK091467
TROAP SEQ ID NO:72	TROPHININ ASSOCIATED PROTEIN (TASTIN)	NM_005480 U04810	LOC283639		
	FAMILY WITH SEQUENCE SIMILARITY 64,	NM_019013	IL17C		NM_013278

FLJ10156	MEMBER A			INTERLEUKIN 17C	AF152099
ASPM SEQ ID NO:73	ASP (ABNORMAL SPINDLE) HOMOLOG, MICROCEPHALY ASSOCIATED	NM_018136 AY367065	GSTA3	GLUTATHIONE S-TRANSFERASE ALPHA 3	AF020919
CDKN3 SEQ ID NO:74	CYCLIN-DEPENDENT KINASE INHIBITOR 3	U02681	DC36		
PTTG3 SEQ ID NO:75	PITUITARY TUMOR-TRANSFORMING 3	NM_021000 AF095289	ALDRL6	MYO-INOSITOL OXYGENASE	NM_017584 AF197129
MYBL2 SEQ ID NO:76	V-MYB MYELOBLASTOSIS VIRAL ONCOGENE HOMOLOG (AVIAN)-LIKE 2	NM_002466	RAG1	RECOMBINATION ACTIVATING GENE 1	NM_000448 M29474
MTB			NYX	NYCTALOPIN	NM_022567 AF254868
RAD54L SEQ ID NO:77	ALPHA THALASSEMIA/MENTAL RETARDATION SYNDROME X-LINKED – LIKE	NM_003579 X97795	MT3	METALLOTHIONEIN 3	NM_005954 BC035624
RRM2 SEQ ID NO:78	RIBONUCLEOTIDE REDUCTASE M2 POLYPEPTIDE	Enzyme ID: 1.17.4.1	GABRR2	GAMMA-AMINOBUTYRIC ACID (GABA) RECEPTOR, RHO 2	
TOPK	PDZ BINDING KINASE	NM_018492 AB027249	TBXA2R	THROMBOXANE A2 RECEPTOR	
TTK SEQ ID NO:79	TTK PROTEIN KINASE		CRH	CORTICOTROPIN RELEASING HORMONE	NM_000756
OIP5 SEQ ID NO:80	OPA INTERACTING PROTEIN 5	AF025441 NM_007280	OR51E2	OLFACTORY RECEPTOR, FAMILY 51, SUBFAMILY E, MEMBER 2	AY033942 NM_030774
SPAG5 SEQ ID NO:81	SPERM ASSOCIATED ANTIGEN 5	AF063308 NM_006461	MGC39526	CHROMOSOME 8 OPEN READING FRAME 56	BC029562
MCM2 SEQ ID NO:82	MINICHROMOSOME MAINTENANCE COMPLEX COMPONENT 2	X67334	FUT5	FUCOSYLTRANSFERASE 5 (ALPHA (1,3) FUCOSYLTRANSFERASE)	NM_002034
RAMP	DENTICLELESS HOMOLOG (DROSOPHILA)	AF195765 NM_016448	C20orf136	STERILE ALPHA MOTIF DOMAIN CONTAINING 10	NM_080621
KIF11 SEQ ID NO:83	KINESIN FAMILY MEMBER 11	X85137 NM_004523	ZFHX2	ZINC FINGER HOMEOBOX 2	AB051549
E2F1 SEQ ID NO:84	E2F TRANSCRIPTION FACTOR 1		SLC12A4	SOLUTE CARRIER FAMILY 12 (POTASSIUM/CHLORIDE TRANSPORTERS), MEMBER 4	NM_005072
	NDC80 HOMOLOG, KINETOCHORE	AF017790	MGC26598		BC024183

KNTC2	COMPLEX COMPONENT (S. CEREVISIAE)	NM_006101		CHROMOSOME 12 OPEN READING FRAME 12	NM_152638
TACC3 SEQ ID NO:85	TRANSFORMING, ACIDIC COILED-COIL CONTAINING PROTEIN 3	AF093543	KCNS2	POTASSIUM VOLTAGE-GATED CHANNEL, DELAYED-RECTIFIER, SUBFAMILY S, MEMBER 2	AB032970 NM_020697
KNSL7	KINESIN FAMILY MEMBER 15	AB035898	FLJ40244	TBC1 DOMAIN FAMILY, MEMBER 28	NM_001039397
ACTB	ACTIN, BETA	M28424 NM_001101	FBX30	F-BOX PROTEIN 30	AF248640
GAPDH	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE	AF261085 NM_002046	AGXT2	ALANINE-GLYOXYLATE AMINOTRANSFERASE 2	AJ292204 NM_031900
KRT3	KERATIN 3	NM_057088	TBX4	T-BOX 4	AF188703 NM_018488
KRT9	KERATIN 9	NM_000226	PADI1	PEPTIDYL ARGININE DEIMINASE, TYPE I	AB033768 NM_013358
KRT1	KERATIN 1	X69725	NPY6R	NEUROPEPTIDE Y RECEPTOR Y6 (PSEUDOGENE)	D86519
KRT18	KERATIN 18	NM_199187	MOG	MYELIN OLIGODENDROCYTE GLYCOPROTEIN	
KRT12	KERATIN 12	NM_000223	ST18	SUPPRESSION OF TUMORIGENICITY 18 (BREAST CARCINOMA) (ZINC FINGER PROTEIN)	AB011107
KRT24	KERATIN 24	NM_019016	PSORS1C2	PSORIASIS SUSCEPTIBILITY 1 CANDIDATE 2	AB031480
KRT20	KERATIN 20	BC031559	PSG9	PREGNANCY SPECIFIC BETA-1-GLYCOPROTEIN 9	M34481 NM_002784
KRT13	KERATIN 13	NM_153490	LOC93432		
KRT23	KERATIN 23 (HISTONE DEACETYLASE INDUCIBLE)	AF102848	NOXA1	NADPH OXIDASE ACTIVATOR 1	AF039697
KRT16	KERATIN 16	S79867 NM_005557	HSA251708		
KRT19	KERATIN 19	NM_002276	PCTK3	PCTAIRE PROTEIN KINASE 3	X66362 NM_002596
KRT8	KERATIN 8	BC000654 NM_002273	FOXP2	FORKHEAD BOX P2	U80741 NM_014491
		NM_002272	FGF17		AB009249

KRT4	KERATIN 4			FIBROBLAST GROWTH FACTOR 17	NM_003867
KRT15	KERATIN 15	NM_002275	FATE	FETAL AND ADULT TESTIS EXPRESSED 1	AF249872 NM_033085
KRT14	KERATIN 14	BC002690 NM_000526	SPATA3	SPERMATOGENESIS ASSOCIATED 3	AY032925 NM_139073
KRT6A	KERATIN 6A	BC014152, L42593, L42610 NM_005554	LOC144305		
KRT6B	KERATIN 6B	BC034535 NM_005555	PRDM8	PR DOMAIN CONTAINING 8	AF275815
KRTHB6	KERATIN 86	X99142 NM_002284	LOC255798		
KRT17	KERATIN 17	X62571	FLJ35834	IQ MOTIF AND UBIQUITIN DOMAIN CONTAINING	AK093153 NM_178827
KRT7	KERATIN 7	NM_005556	DBH	DOPAMINE BETA-HYDROXYLASE	X13256 NM_000787
KRT5	KERATIN 5		P2RX3	PURINERGIC RECEPTOR P2X, LIGAND-GATED ION CHANNEL, 3	Y07683 NM_002559
KRT10	KERATIN 10	J04029 NM_000421	MGC23244	TRANSMEMBRANE AND IMMUNOGLOBULIN DOMAIN CONTAINING 2	BC015655 NM_144615
C10orf95	CHROMOSOME 10 OPEN READING FRAME 95	AK024342 NM_024886	SRY	SEX DETERMINING REGION Y	
LOC157627			RNU71B	SMALL NUCLEOLAR RNA, H/ACA BOX 71B	Y11166 NR_002910
LOC221122			CETN1	CENTRIN, EF-HAND PROTEIN, 1	U03270 NM_004066
LMOD2	LEIOMODIN 2 (CARDIAC)	AC006333	CDCA8	CELL DIVISION CYCLE ASSOCIATED 8	BG354581 NM_018101
ADH6	ALCOHOL DEHYDROGENASE 6 (CLASS V)	AK092768 NM_000672	C20orf71	CHROMOSOME 20 OPEN READING FRAME 71	NM_178466
ZNF333	ZINC FINGER PROTEIN 333	NM_032433	MYCBPAP		BC028393

				MYCBP ASSOCIATED PROTEIN	NM_032133
TCF2	HNF1 HOMEOBOX B	BC017714 NM_000458	LOC151484		
MGC13034	TRANSMEMBRANE PROTEIN 174	BC019346 NM_153217	PTGER1	PROSTAGLANDIN E RECEPTOR 1 (SUBTYPE EP1), 42KDA	
LOC339742			LMX1B	LIM HOMEOBOX TRANSCRIPTION FACTOR 1, BETA	U77457
MLL2	MYELOID/LYMPHOID OR MIXED-LINEAGE LEUKEMIA 2	AF010403	FLJ21736	CARBOXYLESTERASE 3	AK025389 NM_024922
MGC5242	TRAFFICKING PROTEIN PARTICLE COMPLEX 5	BC042161 XM_058961	CNTFR	CILIARY NEUROTROPHIC FACTOR RECEPTOR	M73238
DKFZP434G072	CHROMOSOME 4 OPEN READING FRAME 17	AL136838 NM_032149	WFDC9	WAP FOUR-DISULFIDE CORE DOMAIN 9	AL031671
FOXH1	FORKHEAD BOX H1	AF076292	RNU6C	SMALL NUCLEOLAR RNA, C/D BOX 8	AJ243222 NR_002916
FLJ36119	TUBULIN TYROSINE LIGASE-LIKE FAMILY, MEMBER 10	AK093438 NM_153254	GPR145	MELANIN-CONCENTRATING HORMONE RECEPTOR 2	AF347063 NM_032503
LOC149464			TAL2	T-CELL ACUTE LYMPHOCYTIC LEUKEMIA 2	NM_005421
PCYT1B	PHOSPHATE CYTIDYLYLTRANSFERASE 1, CHOLINE, BETA	AF052510 NM_004845	TAC4	TACHYKININ 4 (HEMOKININ)	AF521560 NM_170685
TAL1	T-CELL ACUTE LYMPHOCYTIC LEUKEMIA 1	M29038 NM_003189	CA9	CARBONIC ANHYDRASE IX	X66839 NM_001216
AQP2	AQUAPORIN 2 (COLLECTING DUCT)	NM_000486	HPS6	HERMANSKY-PUDLAK SYNDROME 6	BC009258 NM_024747
TRPC4	TRANSIENT RECEPTOR POTENTIAL CATION CHANNEL, SUBFAMILY C, MEMBER 4	U40983 NM_003306	OPN5	OPSIN 5	AY288419 NM_181744
ATCAY	ATAXIA, CEREBELLAR, CAYMAN TYPE		ASB15	ANKYRIN REPEAT AND SOCS BOX-CONTAINING 15	AF403033
		AF231917			

HAO2	HYDROXYACID OXIDASE 2 (LONG CHAIN)	NM_001005783	FMN	FORMIN 1	AH002864 NM_001103184
ABO	ABO BLOOD GROUP (TRANSFERASE A, ALPHA 1-3-N-ACETYLGALACTOSAMINYLTRANSFERASE; TRANSFERASE B, ALPHA 1-3-GALACTOSYLTRANSFERASE)	AF134415 NM_020469	FLJ35785	GOLGI AUTOANTIGEN, GOLGIN SUBFAMILY A, 9 PSEUDOGENE	AK093104 XR_017720
APOBEC3F	APOLIPOPROTEIN B MRNA EDITING ENZYME, CATALYTIC POLYPEPTIDE-LIKE 3F	BC038808 NM_145298	LOC145694		
ROCK1	RHO-ASSOCIATED, COILED-COIL CONTAINING PROTEIN KINASE 1	NM_005406	IL13	INTERLEUKIN 13	U31120 NM_002188
PTCRA	PRE T-CELL ANTIGEN RECEPTOR ALPHA	AF084941 NM_138296	FLJ38716	CHROMOSOME 1 OPEN READING FRAME 161	AK096035 NM_152367
SLC22A7	SOLUTE CARRIER FAMILY 22 (ORGANIC ANION TRANSPORTER), MEMBER 7	AF097518	MGC13168	BESTROPHIN 3	AF440758 NM_152439
ONECUT1	ONE CUT HOMEOBOX 1	U77975	CHRNA2	CHOLINERGIC RECEPTOR, NICOTINIC, ALPHA 2 (NEURONAL)	U62431
FLJ42461	SMOOTHELIN-LIKE 2	AK124452 NM_198501	BARX2	BARX HOMEOBOX 2	AF031924 NM_003658
SLC6A18	SOLUTE CARRIER FAMILY 6, MEMBER 18	AK055798 NM_182632	IPF1	PANCREATIC AND DUODENAL HOMEOBOX 1	AF035260 NM_000209
LOC286031			OR1A1	OLFACTORY RECEPTOR, FAMILY 1, SUBFAMILY A, MEMBER 1	AF087918 NM_014565
OR10H2	OLFACTORY RECEPTOR, FAMILY 10, SUBFAMILY H, MEMBER 2	AC004597	FLJ42291		
PLA2G3	PHOSPHOLIPASE A2, GROUP III	AF220490 NM_015715	WNT7A	WINGLESS-TYPE MMTV INTEGRATION SITE FAMILY, MEMBER 7A	D83175 NM_004625
DGCR5	DIGEORGE SYNDROME CRITICAL REGION GENE 5 (NON-PROTEIN CODING)	X91348 NR_002733	PAQR10	MONOCYTE TO MACROPHAGE DIFFERENTIATION-ASSOCIATED 2	BC037881 NM_198403
C9orf93	CHROMOSOME 9 OPEN READING FRAME 93	AY422473 NM_173550	PKHD1	POLYCYSTIC KIDNEY AND HEPATIC DISEASE 1 (AUTOSOMAL RECESSIVE)	AF480064 NM_138694
HTR5A	5-HYDROXYTRYPTAMINE (SEROTONIN) RECEPTOR 5A	NM_024012	LLGL1	LETHAL GIANT LARVAE HOMOLOG 1	

WO 2009/074968 PCT/IB2008/055252

				(DROSOPHILA)	
MGC34919			ADAM21	ADAM METALLOPEPTIDASE DOMAIN 21	AF029900
DKFZp434I1020	CHROMOSOME 15 OPEN READING FRAME 51	AK127834 NM_194295	MUC3A	MUCIN 3A, CELL SURFACE ASSOCIATED	AF113616
C9orf100	CHROMOSOME 9 OPEN READING FRAME 100	AK001187 NM_032818	FLJ25168	CHROMOSOME 17 OPEN READING FRAME 69	AK057897 NM_152466
LOC143425			DLGAP2	DISCS, LARGE (DROSOPHILA) HOMOLOG-ASSOCIATED PROTEIN 2	AB000275 NM_004745
USH1C	USHER SYNDROME 1C	AB006955 NM_005709	FLJ22175	CHROMOSOME 17 OPEN READING FRAME 70	BC008883 NM_025161
PRM3	PROTAMINE 3	Z46940	AAT1	CHROMOSOME 3 OPEN READING FRAME 15	AB063296 NM_033364
ENPP7	ECTONUCLEOTIDE PYROPHOSPHATASE/PHOSPHODIESTERASE 7	AY230663 NM_178543	SLC17A4	SOLUTE CARRIER FAMILY 17 (SODIUM PHOSPHATE), MEMBER 4	AB020527
FLJ14442	ATP/GTP BINDING PROTEIN-LIKE 4	AK027348 NM_032785	SLC10A2	SOLUTE CARRIER FAMILY 10, MEMBER 2	U10417
NEUROD6	NEUROGENIC DIFFERENTIATION 6	AF248954 NM_022728	TACR1	TACHYKININ RECEPTOR 1	M76675 NM_001058
GOR	REX1, RNA EXONUCLEASE 1 HOMOLOG (S. CEREVISIAE)-LIKE 1	AF495523 NM_172239			

The stromal content (SC) can be determined by any technique or calculation method known to the person skilled in the art. For example the stromal content (SC) can be defined as follows:

$SC = \log_2 [(CTE_1 + \text{Stromal Metagene Score}) / (CTE_2 + \text{Reference Metagene Score})] + CTE_3$

The stromal metagene score (equivalent to stromal metagene signature) is a weighted average of the expression values of the stromal genes of Table 4 measured within the tumor biopsy.

Stromal Metagene Score = $\frac{1}{n}\sum_{i=1}^{n}(Stromal\_Gene_i \bullet ks_i)$, wherein

- n can be any value within the range of 2 (inclusive) and 300 (inclusive)
- "Stromal $Gene_i$" represents the expression value of each selected stromal gene.
- ks_i is specific for each stromal gene and defines the importance of the corresponding stromal gene in the calculation of the weighted average of the Stromal Metagene Score. The variable ks_i may take any positive real value within the range of zero (inclusive) and 1000 times the maximal expression value of the stromal gene included in the calculation of the Stromal Metagene score.

Preferably the expression value of at least 5 stromal genes is used to calculate the Stromal Metagene Score. The most preferably the expression value of at least 10 stromal genes is used to calculate the Stromal Metagene Score.

The purpose of the variable ks_i is to adjust (or correct) for the difference in expression magnitude between stromal genes and therefore will make these expression values more similar to all other stromal genes included in the calculation of the Stromal Metagene Score.

The variable CTE1 may take any real value within the range of plus / minus 1000 times the average of the stromal metagene score. The purpose to the CTE1 variable is to adjust for differences in efficiency in extracting the mRNA of stromal genes from the tumor sample relative to the reference genes.

The reference metagene score (equivalent to reference metagene signature) is a weighted average of the expression values of the reference genes of Table 5 measured within the tumor

biopsy or the reference genes of Table 5 and/or the stromal genes of Table 4 measured within stromal tissue of non-tumour reference biopsy (normal non pathological biopsy).

The Reference Metagene Score = = $\frac{1}{n}\sum_{t=1}^{n}(\mathrm{Re}\,ference\_Gene_t \bullet kr_t)$, wherein

- n can be any value within the range of 2 (inclusive) and 237 (inclusive)
- "Reference $Gene_t$" represents the expression value of each selected reference gene.
- kr_t is specific to each reference gene and defines the importance of the corresponding reference gene in the calculation of the weighted average of the Reference Metagene Score. The variable kr_t may take any positive real value within the range of zero (inclusive) and 1000 times the maximal expression value of the reference gene included in the calculation the Reference Metagene Score.

Preferably the expression value of at least 5 reference genes is used to calculate the Reference Metagene Score. The most preferably the expression value of at least 10 reference genes is used to calculate the Reference Metagene Score.

The purpose of the variable kr_t is to adjust (or correct) for the difference in expression magnitude between reference genes and therefore will make these expression values more similar to all other reference genes included in the calculation of the Reference Metagene Score.

The variable CTE2 may take any real value within the range of plus / minus 1000 times the average of the reference metagene score. The purpose to the CTE2 variable is to adjust for differences in efficiency in extracting the mRNA of reference genes from the tumor sample relative to the stromal genes.

The purpose of the variable CT3 is to adjust for systematic bias due to experimental measurements.

A tumor sample is considered as having high stromal content if the score SC is grater than the threshold TH1 (i.e. SC > TH1), which is indicative of resistance to chemotherapy.

A tumor sample is considered as having low stromal content if the score SC is lower than the threshold TH2 (i.e. SC < TH2), which is indicative of sensitivity to chemotherapy.

The variables TH1 and TH2 can take any real value between –50 and + 50, depending on the selected stromal genes, reference genes and the method used to determine the expression values of said genes. The purpose TH1 constant is to adjust for the desire sensitivity and specificity in declaring a tumor sample as having high tumour content. As the threshold TH1 increases, there will be an increase in the true positive rate when classifying a tumour sample as having high stroma. The purpose of the TH2 constant is to adjust for the desire sensitivity and specificity in declaring a tumour sample as having low tumour content. As the value of TH2 decreases, the higher will be the true positive rate of classifying a sample as having low stroma content. The use of both constant brings the advantage of controlling specificity and selectivity of both samples having low stromal content and high stromal content and thus leaving a security margin for samples having "dubious" (i.e. ambiguous) stromal content. (see Example 2).

Generally the determining of expression values of stromal genes and reference genes of the present invention is obtained by detecting mRNA levels of said stromal genes and said reference genes. Usually the detecting of mRNA levels is obtained through, but mot limited to, Microarray hybridization, real-time polymerase chain reaction, Northern blot, In Situ Hybridization, sequencing-based methods, reverse transcription-polymerase chain reaction, RNA expression microarray or RNAse protection assay.

The present invention also provides for a kit for predicting the efficacy of cancer therapy in a subject who has undergone or is undergoing chemotherapy treatment for cancer, characterized in that said kit comprises (a) a reagent for detecting mRNA levels of at least two stromal genes selected from the group consisting of the genes of Table 4 and of at least two reference genes selected from the group consisting of the genes of Table 5 in a stromal tissue sample from a subject, and (b) an instruction sheet. Said reagent comprises buffers and premeasured portions of probes that hybridize to mRNA of at least two stromal genes of Table 4 and to mRNA of at least two reference genes of Table 5. The kit of the present invention further comprises a reagent for preparing and processing a stromal tissue sample from the subject.

For all cancer patients that require chemotherapy, the kit of the present invention informs clinician of probable therapy outcome. Such kit allows the clinicians to prescribe the

chemotherapy regimen whose probable favourable outcome is highest, based on the result of the prediction test leading to an overall increase in treatment efficacy.

The kit of the present invention can further include reagents for collecting a stromal tissue sample from a subject, such as by biopsy, and reagents for preparing and processing the stromal tissue. The kit can also include one or more reagents for performing a gene expression analysis, such as reagents for performing, but not limited to, to determine mRNA expression levels in a tumor sample. Suitable techniques for the determination of mRNA expression levels can be, but not limited to, Microarray hybridization, real-time polymerase chain reaction, Northern blot, In Situ Hybridization, sequencing-based methods, reverse transcription-polymerase chain reaction, RNA expression microarray or RNAse protection assay. For example Northern hybridization is known in the art. See, for example, Sambrook and Russell, Molecular Cloning: A Laboratory Manual, 3rd Edition, 2001, Cold Spring Harbor Laboratory Press; and Harlow and Lane, Using Antibodies, supra. For example, probes for performing Northern blot analyses can be included in such kits. Appropriate buffers for the assays can also be included. Detection reagents required for the assay can also be included. The kits featured herein can also include an instruction sheet describing how to perform the assay for measuring gene expression.

Alternatively, the kit can include reagents for detecting protein levels, said proteins being encoded by the stromal genes of the present invention. Such analysis can be performed, but not limited to, Western Blotting, ELISA and Immunohistochemistry.

The instruction sheet can also include instructions for how to define stromal content (SC) and the thresholds, including how to determine expression levels of the reference genes of the present invention in stromal tissue from a tumor biopsy or from a non-pathological reference biopsy. The instruction sheet can also include instructions to subsequently determine the appropriate chemotherapy for the subject. Methods for determining the appropriate chemotherapy are described above and can be described in detail in the instruction sheet.

The kit can contain separate containers, dividers or compartments for the reagents and informational material. A container can be labeled for use for the determination of gene expression levels and the subsequent determination of an appropriate chemotherapy for the human. The informational material of the kits is not limited in its form. In many cases, the

informational material, e.g., instructions, is provided in printed matter, e.g., a printed text, drawing, and/or photograph, e.g., a label or printed sheet. However, the informational material can also be provided in other formats, such as Braille, computer readable material, video recording, or audio recording. Of course, the informational material can also be provided in any combination of formats.

Those skilled in the art will appreciate that the invention described herein is susceptible to variations and modifications other than those specifically described. It is to be understood that the invention includes all such variations and modifications without departing from the spirit or essential characteristics thereof. The invention also includes all of the steps, features, compositions and compounds referred to or indicated in this specification, individually or collectively, and any and all combinations or any two or more of said steps or features. The present disclosure is therefore to be considered as in all aspects illustrated and not restrictive, the scope of the invention being indicated by the appended Claims, and all changes which come within the meaning and range of equivalency are intended to be embraced therein.

Various references are cited throughout this specification, each of which is incorporated herein by reference in its entirety.

The foregoing description will be more fully understood with reference to the following Examples. Such Examples, are, however, exemplary of methods of practising the present invention and are not intended to limit the scope of the invention.

EXAMPLES

EXAMPLE 1

Patient selection and sample processing

This study was performed in the context of a prospective phase III intergroup trial of neoadjuvant chemotherapy (EORTC 10994/BIG 00-01). Eligible patients had no evidence of metastatic disease, and had histologically confirmed large operable invasive tumor or locally advanced breast cancer. This sub-study was restricted to all cases evaluated at the EORTC data center on April 1st, 2005 meeting the following criteria: (1) estrogen receptor negative tumors defined as <10% of tumor cells stained positive for ER by immunohistochemistry of

the pretreatment formalin-fixed biopsy; (2) patients who had completed the planned chemotherapy regimen with no major protocol violation; (3) non T4 tumors; (4) good quality and >200 ng yield of RNA available from a pretreatment frozen biopsy. Ethical approval for the clinical trial and associated translational projects was obtained in all participating institutions. Patients gave signed informed consent for both the clinical and translational studies. Patients randomized to FEC received either six cycles of FEC 100, a European non-taxane regimen consisting of 500 mg/m^2 5-fluorouracil, 100 mg/m^2 epirubicin, 500 mg/m^2 cyclophosphamide, or six cycles of dose escalated FEC (Swedish patients) [30]. At the completion of chemotherapy all patients underwent either tumorectomy or mastectomy. Pathological complete response (pCR) was used as the outcome measure, defined as disappearance of the invasive component of the primary tumor after treatment, with at most scattered tumor cells detected by the pathologist in the resection specimen. Analysis of pCR was performed locally in each centre. All patients had one incisional or two trucut biopsies frozen before starting chemotherapy. Frozen sections of these biopsies were examined centrally by one pathologist and excluded if the tumor cell content was below 20%. RNA was extracted from frozen sections as previously described and hybridized to Affymetrix X3P chips.

Laser dissection microscopy and cell culture

Samples were collected from fresh colon tumor tissue from 3 independent patients, embedded in OCT, and frozen by immersion in dry ice/ethanol. 12 µm frozen sections were cut and mounted on membrane slides and stained with hematoxilin and eosin solution. Samples were then processed using a laser dissecting microscope, coupled to a CCD camera. Cancer-associated fibroblast (CAF) cultures were prepared as described in [31]. Total RNA of microdissected samples and cell cultures was extracted and analyzed using HU133 Plus 2.0 chips (Affymetrix, USA) after assessing the quality by Bioanalyzer. A tissue specific score defined as the mean expressions of the epithelial fraction (tumour epithelial cells, EPI) minus the Reatctive Stroma (or CAF) fractions was calculated for each gene of the tested list. Negative value imply that the genes are preferentially expressed in RS (or CAF) compared to EPI. Significance was measured empirically by estimating the probability score (i.e. departure from zero) by randomly selecting the same number of genes and measuring its average score (1000 randomization were performed).

Microarray data analysis

MIAME-compliant data were deposited in the Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo) under accession number GSE4779. Raw data were processed with the statistical programming language R (cran.r-project.org), and Bioconductor packages (www.bioconductor.org). Gene expression was normalized with the rma package and transformed to a $\log_2$ scale. Four exclusion criteria were applied to all probesets: (1) A consensus sequence inferior to 56 nucleotides, (2) No annotation to a defined Entrez-gene id, (3) A standard deviation inferior to 0.5 across all experiments and (4) in case of multiple probesets representing a single Entrez-gene id, only the most variable was considered. Cross-platform mapping was performed by taking the Entrez-gene id as primary key. Two external publicly available data sets, the van de Vijver [32] (n=295, Agilent platform, obtained from author's web-site) and Wang [33] (n=286, Affymetrix platfrom, GEO:GSE2034) datasets were used to define the expression modules. A total of 10317 genes could be cross-matched between these two platforms and were used to define the expression modules. For convenience, these two datasets are referred as NKI-EMC. Two external validation set were equally used. From the study by Hess et al [34], the 51 IHC-ER negative patients of the 133 stage 1-3 breast cancer patients were included and referred to as "MDA". The "CEL" files were obtained from the authors' web page (http://www.bioinformatics.mdanderson.org/pubdata.html) and normalized using rma. The Duke University dataset (DUKE) from the study by Bild et al. [35] was obtained from the authors' web-site (http://data.cgt.duke.edu). A total of 120 patients were treated with the following chemotherapies: A3CMF (35), AC4 (11), FAC (48), CMF (26) and total of 44 events. For both the MDA and DUKE datasets, only probesets having a standard variation greater than 0.5 were analysed. When multiple probesets were mapped to the same Entrez-gene id, only the most variable one was kept.

Procedure for analysis

Classifier based on variable selection by association with outcome

A classifier was built by selecting on a training set, the 50 genes most strongly associated with pCR by a two-sample student t-test. The average of these 50 genes was calculated for each tumour. The resulting vector was used to fit a logistic regression model for pCR. Alternatively, the ranksum statistics was used for gene selection. The performance of the classifier was tested on the respective test set in three-fold cross-validation and by pooling the

predictions from the three test sets, so that each sample was classified once. The full cross-validation procedure was repeated 1000 times.

Classifier based on gene modules defined on external data

A method was developed to measure a statistical association between large clusters of functionally related genes that were observed repeatedly [36-39] with response using public datasets. For each of these biological processes, a prototype gene was chosen as the representative genes of the cluster. The prototype gene is a gene regularly found grouped with other functionally related genes when clustering different breast cancer profile datasets. A more quantitative system than hierarchical clustering was used to identify groups of genes associated with each prototype. The method is based on a multiple linear regression model. Linear models provide the framework to allow easy adjustment for potential confounding effects and integration of data coming from different technological platforms. This selection of the genes belonging to a module was unsupervised and was performed using the NKI-EMC datasets. The multilinear model was fitted separately for each study and the associated t-statistics were combined using the fixed-effect meta-analytical method [40]. P-values were estimated by random permutation according to the method of Westfall [41] in order to correct for multiple testing. For all expression modules, the number of genes significantly associated ($p < 0.05$) with the prototype was over 50 genes. The expression of the module genes in the Applicants' data was visualized with heatmaps which are color-coded representations of the mean centred expression matrix. In the definition of the metagenes, the Applicants fixed the number of genes used per module to 50. A sensitivity-analysis to test for the impact of the number of genes to be included was performed. For this, the Applicants ranked the genes by the strength of association with the prototype gene that defines the expression module, which is in decreasing order of the meta-analytical t. Then, the Applicants took from the top consecutive non-overlapping groups of 15 genes and computed their average.

Prediction of Accuracy

The ability of the metagenes to classify the samples by their pCR status, was assessed with the area under the receiver operating characteristic curve (AUC). In all cases, gene selection was performed using external NKI-EMC data only; therefore no pCR-npCR labels were used. In consequence, no cross-validation is needed for unbiased assessment of classification performance. The classification function would be identical as it is not trained on the dataset

where it is applied, that is a cross-validation procedure would give always the same results as the full data set. The 95% confidence intervals for the AUC were estimated by bootstrapping (1000 iterations). P-values were adjusted for multiple testing using the false discovery rate (FDR) method [42].

Gene Set Enrichment Assay (GSEA)

Nine genesets from the MSigDB database [43] (http://www.broad.mit.edu/gsea/) were tested by weighted GSEA as previously described [43]. The genes were ranked and weighted according to the t statistic for each prototype. The test was performed using R statistical software. P-values were obtained empirically by randomizing the ranking 100 000 times the gene composition of the geneset.

Molecular Sub-type Classification

Molecular classification of tumours was performed as previously described [39]. Briefly, the expression profile of each tumour was compared to reference profiles based on the mean expression level of the previously defined Luminal, Basal, Molecular-Apocrine (LAB) gene lists.

Stromal signature and colorectal cancer

To confirm the association of the stromal signature with the fluoruracil activity, we have tested if the stromal signature is associated with response in an independent cohort of 23 rectal carcinoma patients [44] treated with preoperative fluorouracil (1,000 mg/m2/d), as single agent. Response was measured as by T level downsizing and histopathologic tumour regression grading (see Figure 11). Results show a significant association between response to fluorourail (5-FU) and the stromal's metagene scores AUC 0.77; p = 0.032; figure xx). In summary, results show that the stromal signature predicts response to fluorouracil in rectal cancer patients. A significant association with 5-FU resistance in both of these studies confirming results derived form the NCI60 datasets. More importantly, these results show that the stromal signature could be clinical useful in other cancer, such as colon and rectal cancer in addition to breast carcinoma.

EXAMPLE 2:

The following genes are selected to be part of the stromal signature: DCN, PLAU, CYR61, and SPARC and have the respective expression values of 8.0, 4.6, 8.2, 7.6. Their respective coefficients ks are 1, 2, 1, and 1.

The Stroma Metagene Score = average (ks_1*DCN + ks_2 * PLAU + ks_3 * CYR61 + ks_4 *SPARC) = average (1 * 8.0 + 2 * 4.6 + 1 * 8.2 + 1 * 7.6) = 8.2

The ks_2 has a value of 2 rather then 1 to compensate for the property of PLAU to have a lower expression level relative to the 3 other genes.

The following genes are selected to be part of the reference signature: GAPDH, KRT5, BIRC5 and TPX2 and have the respective expression values of 4.2, 5.1, 4.7, and 3.9. Their respective coefficients kr are 1, 2, 1, and 1.

The Reference Metagene Score = average (kr_1* GAPDH + kr_2 * KRT5 + kr_3 * BIRC5 + kr_4 * TPX2) = average (1 * 4.2 + 1 * 5.1 + 1 * 4.7 + 1 * 3.9) = 4.475

The efficiency of extracting mRNA for stromal and reference genes is equal, therefore the values of both CT1 and CTE2 constants will be zero. As there is no batch specific bias, the value of CTE3 constant will be also zero.

The TH1 threshold is defined as +1, the TH2 is defined a –1.

The SC value is (8.2+0) / (4.475+0) = 1.82. As 1.82 is > +1, the tumor will be classified as having a high tumor content.

Reference list

1. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* **365**, 1687-717 (2005).
2. Colleoni, M. et al. Chemotherapy is more effective in patients with breast cancer not expressing steroid hormone receptors: a study of preoperative treatment. *Clin Cancer Res* **10**, 6622-8 (2004).
3. Hannemann, J. et al. Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer. *J Clin Oncol* **23**, 3331-42 (2005).
4. Guarneri, V. et al. Prognostic value of pathologic complete response after primary chemotherapy in relation to hormone receptor status and other factors. *J Clin Oncol* **24**, 1037-44 (2006).
5. Fisher, E.R. et al. Pathobiology of preoperative chemotherapy: findings from the National Surgical Adjuvant Breast and Bowel (NSABP) protocol B-18. *Cancer* **95**, 681-95 (2002).
6. Bonadonna, G. et al. Primary chemotherapy in operable breast cancer: eight-year experience at the Milan Cancer Institute. *J Clin Oncol* **16**, 93-100 (1998).
7. Fisher, B. et al. Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. *J Clin Oncol* **16**, 2672-85 (1998).
8. Potti, A. et al. Genomic signatures to guide the use of chemotherapeutics. *Nat Med* **12**, 1294-300 (2006).
9. Gianni, L. et al. Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. *J Clin Oncol* **23**, 7265-77 (2005).
10. Hess, K.R. et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol* **24**, 4236-44 (2006).
11. Thuerigen, O. et al. Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer. *J Clin Oncol* **24**, 1839-45 (2006).
12. Aoudjit, F. & Vuori, K. Integrin signaling inhibits paclitaxel-induced apoptosis in breast cancer cells. *Oncogene* **20**, 4995-5004 (2001).
13. Cordes, N., Blaese, M.A., Plasswilm, L., Rodemann, H.P. & Van Beuningen, D. Fibronectin and laminin increase resistance to ionizing radiation and the cytotoxic drug Ukrain in human tumour and normal cells in vitro. *Int J Radiat Biol* **79**, 709-20 (2003).
14. Damiano, J.S., Cress, A.E., Hazlehurst, L.A., Shtil, A.A. & Dalton, W.S. Cell adhesion mediated drug resistance (CAM-DR): role of integrins and resistance to apoptosis in human myeloma cell lines. *Blood* **93**, 1658-67 (1999).
15. Sethi, T. et al. Extracellular matrix proteins protect small cell lung cancer cells against apoptosis: a mechanism for small cell lung cancer growth and drug resistance in vivo. *Nat Med* **5**, 662-8 (1999).
16. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* **103**, 5923-8 (2006).
17. Farmer, P. et al. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* **24**, 4660-71 (2005).

18. Perou, C.M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747-52 (2000).
19. Sorlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869-74 (2001).
20. Sotiriou, C. et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* **98**, 262-72 (2006).
21. van de Vijver, M.J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009 (2002).
22. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
23. Bild, A.H., Potti, A. & Nevins, J.R. Linking oncogenic pathways with therapeutic opportunities. *Nat Rev Cancer* **6**, 735-41 (2006).
24. West, R.B. et al. Determination of stromal signatures in breast carcinoma. *PLoS Biol* **3**, e187 (2005).
25. Finak, G. et al. Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. *Breast Cancer Res* **8**, R58 (2006).
26. Weaver, V.M. et al. beta4 integrin-dependent formation of polarized three-dimensional architecture confers resistance to apoptosis in normal and malignant mammary epithelium. *Cancer Cell* **2**, 205-16 (2002).
27. Misra, S., Ghatak, S. & Toole, B.P. Regulation of MDR1 expression and drug resistance by a positive feedback loop involving hyaluronan, phosphoinositide 3-kinase, and ErbB2. *J Biol Chem* **280**, 20310-5 (2005).
28. Hazlehurst, L.A. et al. Reduction in drug-induced DNA double-strand breaks associated with beta1 integrin-mediated adhesion correlates with drug resistance in U937 cells. *Blood* **98**, 1897-903 (2001).
29. Song, S., Wientjes, M.G., Gan, Y. & Au, J.L. Fibroblast growth factors: an epigenetic mechanism of broad spectrum resistance to anticancer drugs. *Proc Natl Acad Sci U S A* **97**, 8658-63 (2000).
30. Bergh, J. et al. Dosage of adjuvant G-CSF (filgrastim)-supported FEC polychemotherapy based on equivalent haematological toxicity in high-risk breast cancer patients. Scandinavian Breast Group, Study SBG 9401. *Ann Oncol* **9**, 403-11 (1998).
31. Orimo, A. et al. Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. *Cell* **121**, 335-48 (2005).
32. van de Vijver, M.J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009 (2002).
33. Wang, Y. et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-9 (2005).
34. Hess, K.R. et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol* **24**, 4236-44 (2006).
35. Bild, A.H., Potti, A. & Nevins, J.R. Linking oncogenic pathways with therapeutic opportunities. *Nat Rev Cancer* **6**, 735-41 (2006).
36. Perou, C.M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747-52 (2000).

37. Sorlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869-74 (2001).
38. Sotiriou, C. et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* **98**, 262-72 (2006).
39. Farmer, P. et al. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* **24**, 4660-71 (2005).
40. Hedges, L. & Olkin, I. *Statistical methods for meta-analysis.*, (London, 1985).
41. Westfall, P. & Young, S. *Resampling-based multiple testing: Examples and methods for p-value adjustment.*, (New York, 1993).
42. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerfull approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 289-300 (1995).
43. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
44. B. Michael Ghadimi, Marian Grade, Michael J. Difilippantonio, Sudhir Varma, Richard Simon, Cristina Montagna, Laszlo Fu¨zesi, Claus Langer, Heinz Becker, Torsten Liersch, and Thomas Ried Effectiveness of Gene Expression Profiling for Response Prediction of Rectal Adenocarcinomas to Preoperative Chemoradiotherapy, J. Clin. Oncol. 2005

CLAIMS

1. A method for predicting the efficacy of cancer therapy in a subject who has undergone or is undergoing chemotherapy treatment for cancer, characterized in that said method comprises

(a) obtaining a stromal tissue sample from a subject,

(b) determining in said stromal tissue sample the expression values of at least two stromal genes and of at least two reference genes,

(c) defining stromal content (SC) from the expression values of step (b),

(d) comparing the stromal content (SC) with a reference threshold,

(e) predicting resistance to chemotherapy of said subject based on the step (d), wherein high stromal content is indicative of resistance to chemotherapy, while low stromal content is indicative of sensitivity to chemotherapy,

(f) adapting the treatment of said subject

2. The method of claim 1, wherein cancer is selected from the group comprising breast cancer, colon cancer, lung cancer, colorectal cancer, head and neck cancer, or ovarian cancer.

3. The method of claims 1 to 2, wherein the chemotherapy is anthracycline-based neo-adjuvant chemotherapy.

4. The method of claims 1 to 3, wherein the chemotherapy is selected from the group comprising fluorouracil, epirubicin and cyclophosphamide based chemotherapy or combination thereof.

5. The method of claims 1 to 4, wherein said stromal tissue sample is taken from a tumour biopsy.

6. The method of claims 1 to 5, wherein said stromal genes are selected from the group consisting of the genes of Table 4.

7. The method of claim 6, wherein said stromal genes are DCN, CSPG2, CDH11, ASPN, SPARC, ITGBL1, PLAU, COL1A2, SNAI2, POSTN and THBS2.

8. The method of claims 1 to 7, wherein said reference genes are selected from the group consisting of the genes of Table 5.

9. The method of claim 8, wherein said reference genes are GAPDH, ACTB, TCF2, ZNF333, ADH6, FOXH1, TPX2 1, CENPA, BIRC5, TOP2A.

10. The method of claims 1 to 9, wherein said reference genes are genes selected from the group consisting of the genes of Table 4 or Table 5 determined within stromal tissue of non-tumor reference biopsy.

11. The method of claims 1 to 10, wherein determining in said stromal tissue sample the expression values of at least 5 stromal genes and of at least 5 reference genes.

12. The method of claims 1 to 11, wherein the determining of expression values of said stromal genes and said reference genes is obtained by detecting mRNA levels of said stromal genes and said reference genes.

13. A kit for predicting the efficacy of cancer therapy in a subject who has undergone or is undergoing chemotherapy treatment for cancer, characterized in that said kit comprises
(a) a reagent for detecting mRNA levels of at least two stromal genes selected from the group consisting of the genes of Table 4 and of at least two reference genes selected from the group consisting of the genes of Table 5 in a stromal tissue sample from a subject, and
(b) an instruction sheet.

14. The kit of claim 13, wherein said reagent comprises buffers and premeasured portions of probes that hybridize to mRNA of at least two stromal genes of claim 6 and to mRNA of at least two reference genes of claim 8.

15. The kit of claims 13 to 14, further comprising a reagent for preparing and processing a stromal tissue sample from the subject.

Figure 1

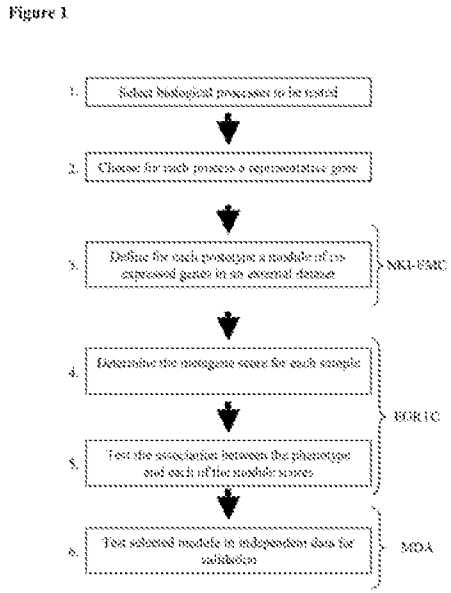
Figure 1
1. Select biological processes to be tested
2. Choose for each process a representative gene
3. Define for each prototype a module of co-expressed genes in an external dataset
NKI-ERIC
4. Determine the metagene score for each sample
5. Test the association between the phenotype and each of the module scores
EGRAC
6. Test selected models in independent data for validation
MDA

Figure 2

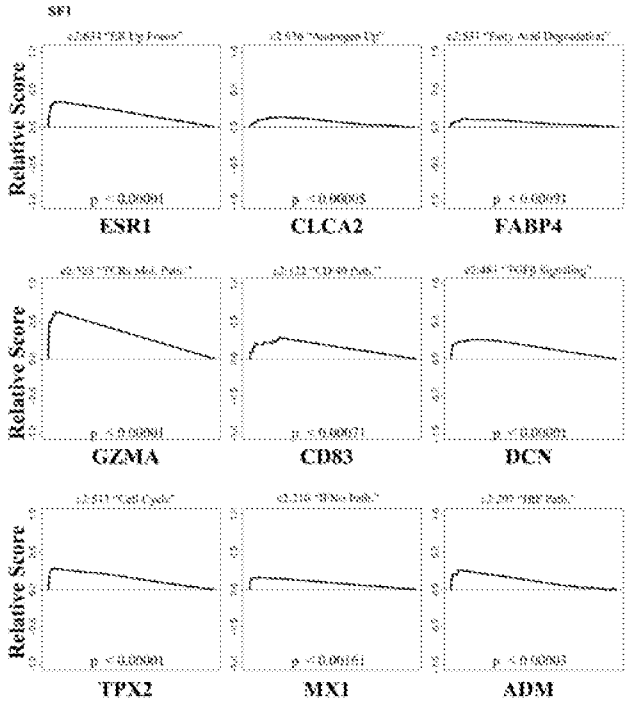
Relative Score
ESR1
CLCA2
FABP4
Relative Score
GZMA
CD83
DCN
Relative Score
TPX2
MX1
ADM

Figure 3

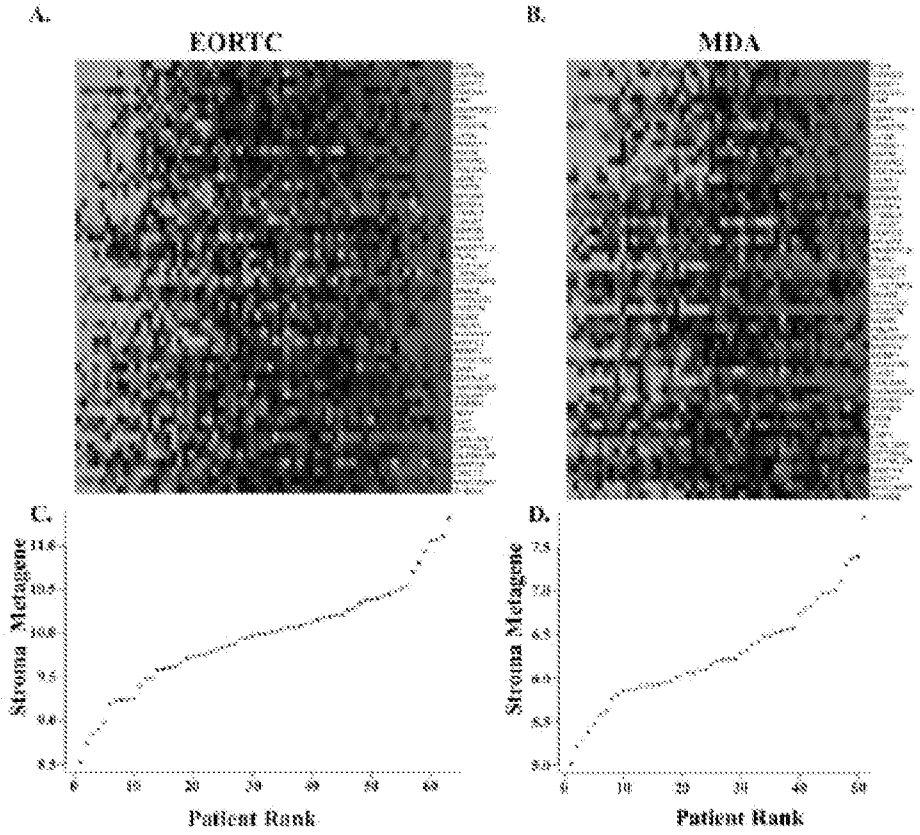
A.
EORTC
B.
MDA
C.
Stroma Metagene
Patient Rank
D.
Stroma Metagene
Patient Rank

Figure 4

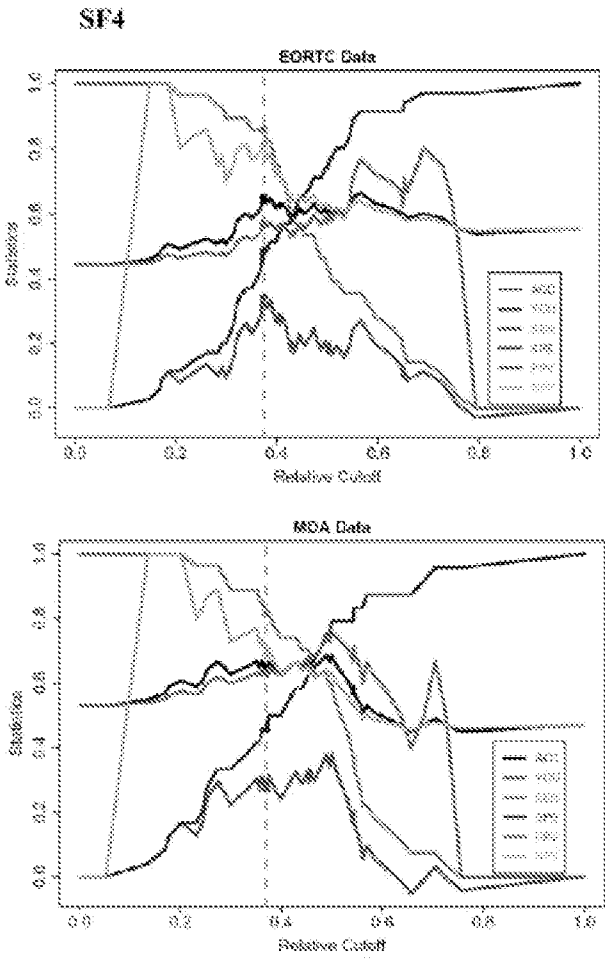
SF4
EORTC Data
Sensitivity
Relative Cutoff
MDA Data
Sensitivity
Relative Cutoff

Figure 5

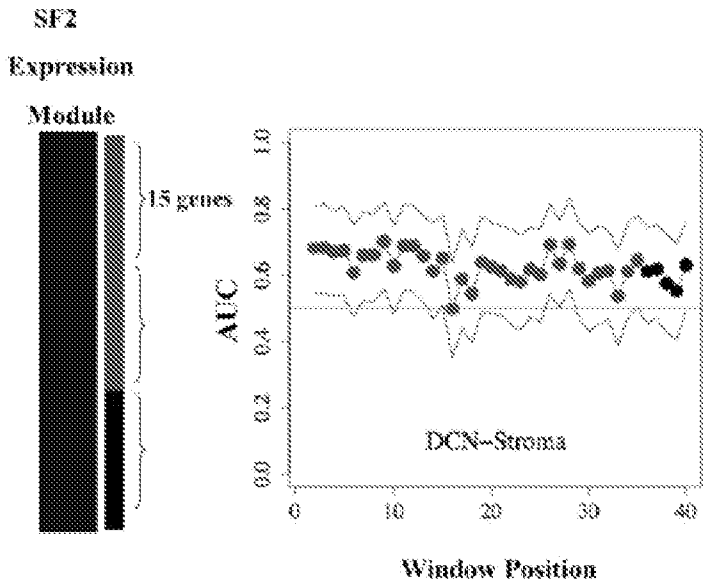

Figure 6

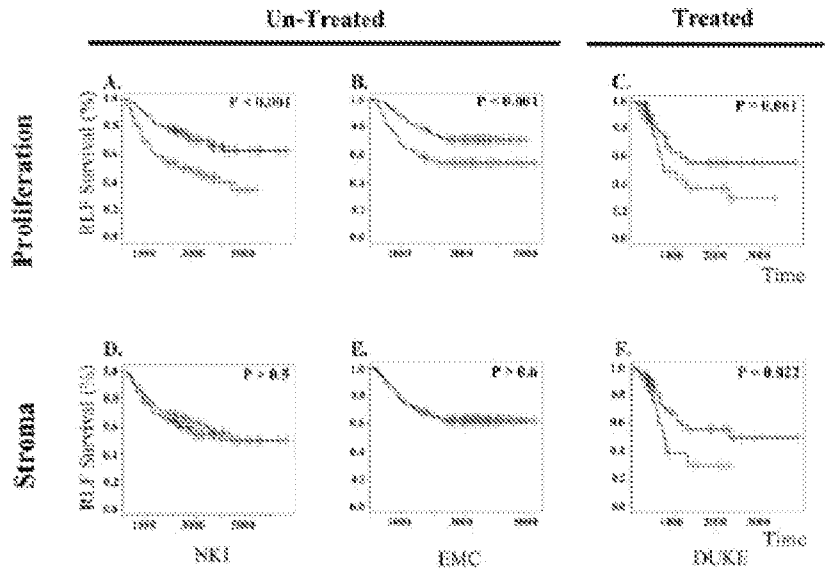

Figure 7

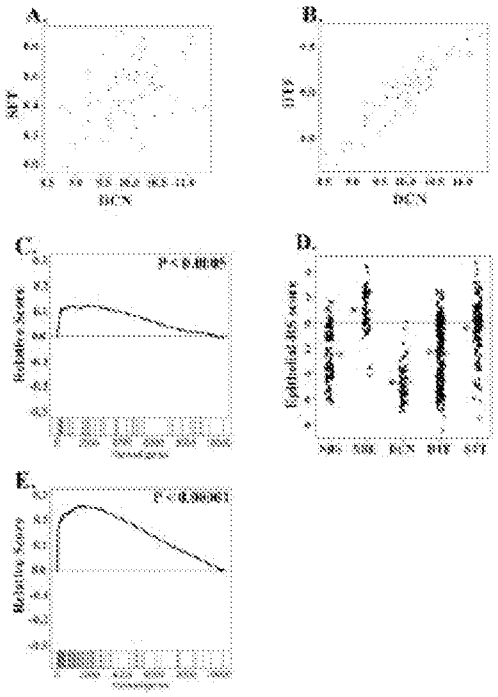

Figure 8

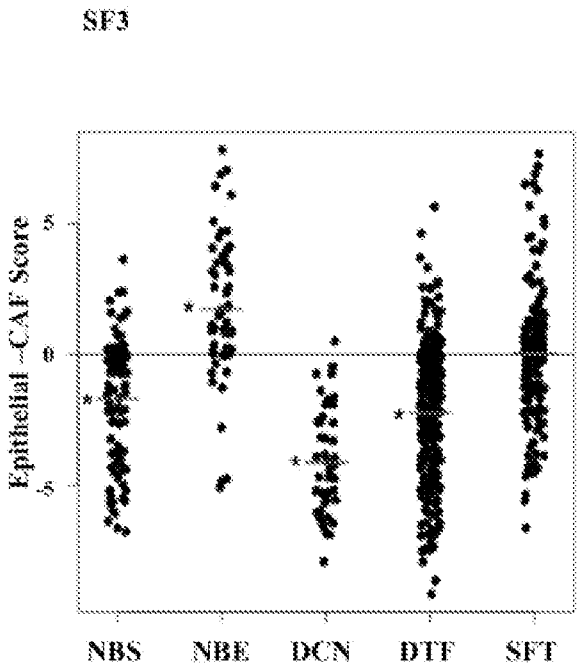

Figure 9

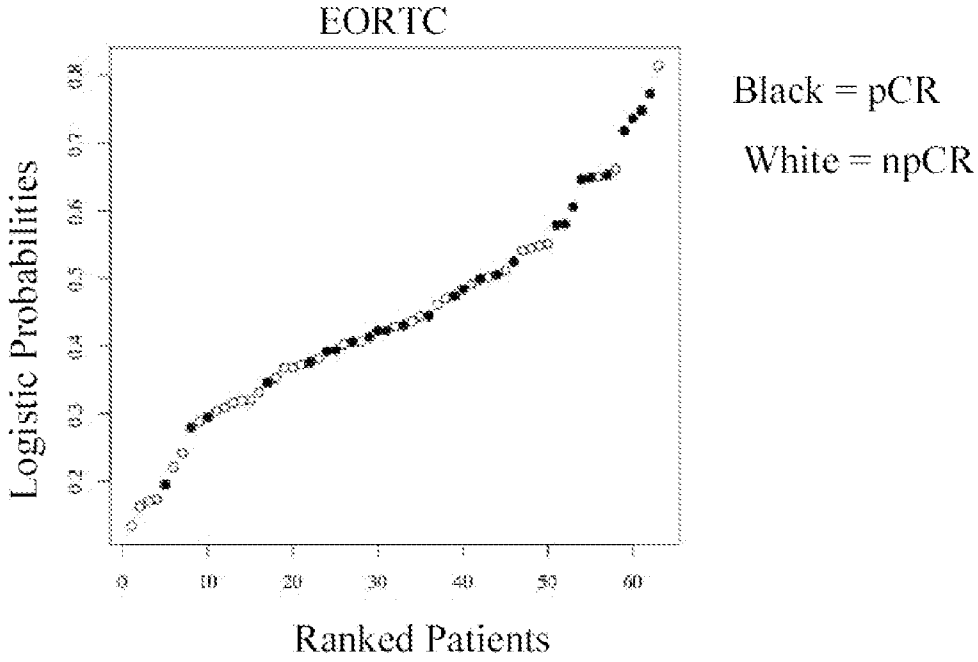

Figure 10

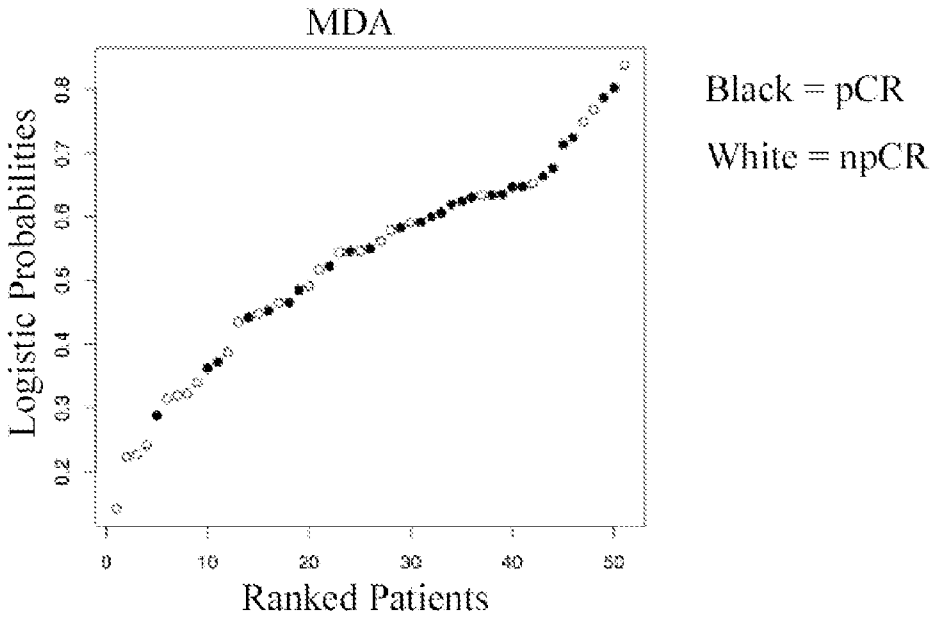

Figure 11