

MOLECULAR MARKER, REFERENCE GENE, AND APPLICATION AND TEST KIT THEREOF, AND METHOD FOR CONSTRUCTING TESTING MODEL

CROSS REFERENCE TO RELATED APPLICATIONS

5 **[0001]** This application claims the priority of Chinese Patent Application No. 201610509983.0, filed to SIPO on June 30, 2016, and titled with “MOLECULAR MARKER, REFERENCE GENE, AND APPLICATION AND TEST KIT THEREOF, AND METHOD FOR CONSTRUCTING TESTING MODEL”, and the entire disclosures of which are hereby incorporated by reference.

10

FIELD

[0002] The present disclosure relates to the technical field of biotechnology field, specifically to molecular markers, internal reference genes and application thereof, test kit and method for constructing test model.

15

BACKGROUND

[0003] Breast cancer is one of the main reasons that threaten the health of women around the world. In 2013, the National cancer statistics published by the American Cancer Society showed that the incidence of breast cancer ranked the first among women’s cancers, and the mortality ranked the second. According to the latest data in the National Cancer Center, in 2013, 232,340 new cases of breast cancer were reported and 39,620 cases of death were reported among women in the United States. In the United States, an average of one out of eight women suffers from breast cancer. Although China is a country with low incidence of breast cancer, its incidence and mortality have increased significantly in recent years. About 15% of the 1.3 million newly diagnosed breast cancer patients every year in the world are from China. According to the statistics of China Breast Cancer Network, the annual increase of breast cancer in China is 3%-4%, which is more than 1%-2% of the world level. The incidence is the first among the cancers easy to be suffered from in women. Urgent development of prevention, diagnosis, prognosis, and individualized treatment technologies of breast cancer is needed.

20

25

[0004] Breast cancer is a kind of highly heterogeneous tumor with many prognostic factors. Breast cancer patients having the same clinical stage, histological grade and hormone receptor expression and receiving the same treatment plan may have different prognosis. How to accurately determine the prognosis of breast cancer patients and formulate corresponding individualized treatment programs to avoid the harm and burden caused by over-treatment and improper treatment is an urgent problem to be solved in clinical practice.

[0005] With the rapid development of molecular biology technology, it is possible to find and detect prognosis-related genes in breast cancer using molecular biology methods such as polymerase chain reaction (PCR), probe hybridization and gene chip. In 2002, Van't Veer et al. screened 117 cases of breast cancer by DNA chip technology and found 70 genes related to breast cancer prognosis. In 2004, American researchers used RT-PCR method to verify 675 breast cancer samples, and obtained 21 prognostic-related genes. Genomic Health Co. developed Oncotype DX[®], a prognostic-related product for breast cancer. Oncotype DX[®] is the only prognostic test product for breast cancer recommended by NCCN guidelines, ASCO clinical guidelines and St Gallen's clinical consensus. In addition, Yasuto Naoi et al. used DNA chip technology to study cancer tissue samples from patients with ER-positive and lymph node-negative breast cancer in the Japanese population, and found 95 prognostic-related genes. Torsten O. Nielsen's team found that 50 gene combinations could provide more information on prognostic prediction of breast cancer as compared with clinical factors and immunohistochemical staining, and using FFPE samples instead of fresh or frozen samples for testing could expand the range of the detectable samples. The prognostic-related test products for breast cancer—Oncotype DX[®], Mammaprint, Prosigna[™], and MapQuant Dx[™]—have successively been certified by the FDA and CE in 2002-2013. However, these products are currently based on research and development for the European and American people. These products are expensive after entering China, and whether the genes and their test models are applicable to the Chinese people has yet to be verified. Therefore, it is of great significance to develop an economical and efficient prognostic test technology for breast cancer patients in China.

SUMMARY

[0006] In view of this, the present disclosure provides molecular markers and application thereof, test kit, and method for constructing test model. The kit has a superior performance in breast cancer prognosis evaluation test as compared with the clinical pathological evaluation results. It can reduce the over-treatment and improper treatment caused by pathological diagnosis errors to a certain extent, meeting the requirements of individualized and precise treatment of breast cancer patients, and further improving the technical methods of the prognosis prediction of breast cancer for patients in China.

[0007] In order to achieve the above object, the present disclosure provides the following technical solutions.

[0008] The present disclosure provides a gene composition comprising molecular marker(s) *MAPT* and/or *MS4A1*.

[0009] The present disclosure provides a gene composition comprising molecular markers *BCL2*, *PGR*, *SCUBE2*, *ESR1*, *MKI67*, *CCNB1*, *MYBL2*, *GRB7*, *ERBB2*, *MMP11*, *CD68*, *BAG1*, *MAPT* and *MS4A1*.

[0010] In some embodiments of the present disclosure, the gene composition further comprises internal reference genes *ACTB*, *GAPDH*, *GUSB*, *NUP214* and *VCAN*.

[0011] The present disclosure further provides use of the gene composition in the manufacture of a test device for risk prediction of recurrence and/or mortality in 3-10 years after breast cancer surgery.

[0012] In some embodiments of the present disclosure, the evaluating and testing of the risk of recurrence and/or mortality in 3-10 years after breast cancer surgery comprises specifically steps of: obtaining total RNA of a test sample, obtaining cDNA by reverse transcription, obtaining Ct values of the molecular marker and the internal reference genes by a fluorescent quantitative PCR, calculating average Ct value of the internal reference genes to obtain an average Ct value (Ct') of internal reference gene combination, normalizing Ct value of the molecular marker respectively by subtracting Ct' value of the internal reference gene combination to obtain ΔCt , and then analyzing ΔCt value, and age, pT value and LN value of the subject by a model for risk

prediction of recurrence or mortality in 3-10 years after breast cancer surgery constructed by random forest algorithm to obtain a result. Among them, pT value refers to pathological stage, and the LN value refers to number of lymph node metastasis. The value obtained by the analysis is compared with a threshold value to obtain the result, wherein the threshold value is 5. The value ≥ 5 obtained by the analysis represents a good prognosis, and the value < 5 obtained by the analysis represents a poor prognosis.

[0013] In some embodiments of the present disclosure, the method for constructing the model for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery comprises steps of: constructing a matrix by using ΔCt value of the molecular marker of the test sample, and age, pT value and LN value of the subject, randomly selecting half as a training set and another half as a verification set, establishing a prediction model containing 10,000 decision trees by random forest algorithm, randomly sampling for ≥ 1000 times and establishing ≥ 1000 prediction models, selecting ≥ 39 optimal models having the highest concordance rate with follow-up information from the ≥ 1000 prediction models as submodels of final model, and setting the median of the ≥ 39 submodels as a final prediction value of prognostic risk.

[0014] The random forest is composed of many decision trees. The decision tree is constructed by the method of double randomness of attributes and samples, so it is also called random decision tree. In the random forest, there is no correlation between decision trees. When the test data enter the random forest, it is classified by each decision tree. Finally, the category with the most classification in all the decision trees is the final result, i.e., the “voting” result of the decision tree. In other words, the random forest is a classifier including multiple decision trees, and the category of its output is determined by the mode of the category of the output of the individual tree. In the present disclosure, we have optimized the traditional random forest algorithm. We randomly sample for 1000 times and establish 1000 models. 39 optimal models having both higher accuracy and specificity from the 1000 model are selected as submodels of the final model, and the median of the 39 submodels is set as the final prediction result.

[0015] Age of the untreated early or intermediate stage ER or PR positive breast cancer newly diagnosed patient, PT stage, number of LN metastases, and Ct values of 14 molecular markers *BCL2*, *PGR*, *SCUBE2*, *ESR1*, *MKI67*, *CCNB1*, *MYBL2*, *GRB7*, *ERBB2*, *MMP11*, *CD68*, *BAG1*, *MAPT*, *MS4A1* and 5 housekeeping genes *ACTB*, *GAPDH*, *GUSB*, *NUP214*, *VCAN* are input into

39 prediction models for analysis to obtain predictive analysis results, thus obtaining the risk values of the recurrence or mortality in 3-10 years, which will be determined as a good prognosis or a poor prognosis based on the risk threshold (the threshold being 5).

[0016] In some embodiments of the present disclosure, the test sample in the present disclosure is a FFPE sample from an untreated early or intermediate stage ER or PR positive breast cancer newly diagnosed patient.

[0017] The present disclosure further provides a primer set for amplifying the gene composition, sequences of which are represented by SEQ ID NO: 1 to SEQ ID NO: 28.

[0018] The present disclosure further provides a probe set for amplifying the gene composition, sequences of which are represented by SEQ ID NO: 29 to SEQ ID NO: 42.

[0019] The present disclosure further provides a primer set for amplifying the internal reference genes of the gene composition, sequences of which are represented by SEQ ID NO: 43 to SEQ ID NO: 47.

[0020] The present disclosure further provides a probe set for amplifying the internal reference genes of the gene composition, sequences of which are represented by SEQ ID NO: 48 to SEQ ID NO: 52.

[0021] The present disclosure further provides a test kit for risk prediction of recurrence and/or mortality in 3-10 years after breast cancer surgery, comprising the primer set and/or the probe set, and reagents commonly used in a kit.

[0022] The present disclosure further provides a method for constructing the model for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery, comprising steps of: constructing a matrix by using ΔC_t value of the molecular marker of the test sample, and age, pT value and LN value of the subject, randomly selecting half as a training set and another half as a verification set, establishing a prediction model containing 10,000 decision trees by random forest algorithm, randomly sampling for ≥ 1000 times and establishing ≥ 1000 prediction models, selecting ≥ 39 optimal models having the highest concordance rate with follow-up information from the ≥ 1000 prediction models as submodels of final model, and setting the median of the ≥ 39 submodels as a final prediction value of prognostic risk.

[0023] The present disclosure further provides a method for evaluating and testing the risk of recurrence or mortality in 3-10 years after breast cancer surgery, comprising steps of: obtaining total RNA of a test sample, obtaining cDNA by reverse transcription, obtaining Ct values of the molecular marker and the internal reference genes by a fluorescent quantitative PCR, calculating average Ct value of the internal reference genes to obtain an average Ct value (Ct') of internal reference gene combination, normalizing Ct value of the molecular marker respectively by subtracting Ct' value of the internal reference gene combination to obtain Δ Ct, and then analyzing Δ Ct value, and age, pT value and LN value of the subject by a model for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery constructed by random forest algorithm to obtain a result, i.e., obtain the risk values of the recurrence or mortality in 3-10 years and predict a good prognosis or a poor prognosis based on the risk threshold (the threshold being 5).

[0024] In the present disclosure, the test sample is a FFPE sample from an untreated early or intermediate stage ER or PR positive breast cancer newly diagnosed patient.

[0025] The technical solutions used in the present disclosure include (1) through literature and database research, selecting 192 breast cancer-related candidate genes (not limited to breast cancer prognosis-related genes, containing internal reference genes) and customizing TLDA gene expression chips (Applied Biosystems Co.); (2) systematically collecting complete demographic data, clinical data and follow-up data (recurrence and metastasis time, survival time), and selecting FFPE samples of untreated early or intermediate stage ER or PR positive breast cancer newly diagnosed patients, using customized TLDA chips to test 192 genes and screening the internal reference genes and molecular markers related to breast cancer prognosis; (3) verifying the candidate molecular markers and reference genes in independent samples and constructing a model for risk prediction of recurrence or mortality in 3-10 years after surgery by random forest algorithm, and evaluating the concordance rate between the prediction models and the follow-up results; (4) conducting further verification using independent clinical samples: FFPE samples of 19 cases of ER or PR positive early or intermediate stage breast cancer newly diagnosed patient whose clinical follow-up data are known and evaluating the concordance rate between the prediction models and the follow-up results.

[0026] The present disclosure provides a gene assay system for prognostic evaluation of

recurrence or mortality in 3-10 years after surgery in untreated early or intermediate stage ER or PR positive breast cancer newly diagnosed patients. By extracting total RNA from Formalin-Fixed and Paraffin-Embedded (FFPE) breast cancer tissue samples, after reverse transcription by universal primers, Ct values of the expression of the breast cancer prognosis-related 14 molecular markers and 5 internal reference genes are detected by using PCR. The Ct value and the age, pT value and LN number of the subject are introduced into the model for risk prediction of recurrence or mortality in 3-10 years after surgery in ER or PR positive early or intermediate stage breast cancer patients constructed by random forest algorithm, to evaluate a good prognosis or a poor prognosis. Compared with the follow-up information, the system achieves an accuracy of 70%. Except for patient age, pT stage, and LN number, there is no need to rely on other clinical pathological information.

[0027] The kit provided in the present disclosure provides a prediction accuracy rate of 81.1% for newly diagnosed breast cancer patients with low risk of recurrence or mortality in 3-10 years, while a pathological prediction test accuracy is 71.9%. In addition, the kit provides a sensitivity of prediction accuracy rate of 54.4% for newly diagnosed breast cancer patients with high risk of recurrence or mortality in 3-10 years, which is close to the accuracy of the corresponding pathological prediction test of 56.8%. Compared with the clinical follow-up information, the kit achieves a concordance rate of 70%. Except for patient age, pT stage, and LN number, there is no need to rely on other clinical pathological information.

[0028] The assay system and kit have a superior performance in prognosis evaluation test for breast cancer as compared with the clinical pathological evaluation results. It can reduce the over-treatment and improper treatment caused by pathological diagnosis errors to a certain extent, meeting the requirements of individualized and precise treatment of breast cancer patients, and further improving the technical methods for the prognosis prediction of breast cancer patients in China.

DESCRIPTION OF THE DRAWINGS

[0029] In order to more clearly illustrate the embodiments of the present disclosure or the technical solutions in the prior art, the drawings used in the embodiments or the description of the

prior art will be briefly described below.

[0030] Figure 1 shows the results of analysis of the correlation between the internal reference genes and the test genes;

[0031] Figure 2 shows the establishment of a model for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery.

DETAILED DESCRIPTION

[0032] The present disclosure discloses molecular markers, internal reference genes and application thereof, test kit and method for constructing test model. Those skilled in the art can learn from the contents of the text and appropriately improve the process parameters. It is to be understood that all such alternatives and modifications are obvious to those skilled in the art and are considered to be included in the present disclosure. The method and application in the present disclosure have been described in the preferred embodiments, and it is obvious for the relevant people to change, make appropriate modification and combination the method and application described herein without departing from the content, spirit and scope of the present disclosure, and carry out and apply the technique of the present disclosure.

[0033] The technical solutions used in the present disclosure includes (1) through literature and database research, selecting 192 breast cancer-related candidate genes (not limited to breast cancer prognosis-related genes, containing internal reference genes) and customizing TLDA gene expression chips (Applied Biosystems Co.); (2) systematically collecting complete demographic data, clinical data and follow-up data (recurrence and metastasis time, survival time), and selecting FFPE samples of untreated early or intermediate stage ER or PR positive breast cancer newly diagnosed patients, using customized TLDA chips to test 192 genes and screening the internal reference genes and molecular markers related to breast cancer prognosis; (3) verifying the candidate molecular markers and reference genes in independent samples and constructing a model for risk prediction of recurrence or mortality in 3-10 years after surgery by random forest algorithm, and evaluating the concordance rate between the prediction models and the follow-up results; (4) conducting further verification using independent clinical samples: FFPE samples of 19 cases of ER or PR positive early or intermediate stage breast cancer newly diagnosed patient

whose clinical follow-up data are known and evaluating the concordance rate between the prediction models and the follow-up results.

1. Selection of research samples

[0034] (1) untreated early or intermediate stage breast cancer newly diagnosed patients;

5 **[0035]** (2) untreated early or intermediate stage ER or (and) PR positive breast cancer patients;

[0036] (3) having or (and) having no lymph node metastasis, and the number of lymph node metastasis;

[0037] (4) having accurate and detailed follow-up information.

[0038] A total of 339 samples meeting the criteria are used in this study.

10 2. Total RNA extraction of FFPE samples

[0039] Total RNA is extracted from FFPE samples using High Pure FFPET RNA Isolation Kit (Roche), and the obtained RNA has a concentration ranging from 25- 400 ng/ μ L, OD260/280 in the range of 1.8-2.0, and OD260/230 in the range of 1.5-2.0.

3. TLDA (Applied Biosystems) chip assay.

15 **[0040]** The following experiments are performed on 26 pairs of samples with good prognosis and poor prognosis from the known clinical follow-up information:

[0041] (1) conducting reverse transcription reaction on the total RNA to obtain a cDNA sample;

[0042] (2) conducting TLDA chip assay on the cDNA sample;

20 **[0043]** (3) analyzing and processing the data to obtain candidate molecular markers and internal reference genes.

4. Real-time quantitative RT-PCR (qRT-PCR)

[0044] Candidate molecular markers are verified on 289 samples from the known clinical follow-up information:

25 **[0045]** (1) conducting reverse transcription reaction on the total RNA to obtain a cDNA sample;

[0046] (2) conducting RT-PCR test on the cDNA sample;

[0047] (3) analyzing and processing the data.

5. Method for preparing a diagnostic kit

[0048] Through a method of customizing TLDA chip, the internal reference genes and genes with different expression between 26 samples of breast cancer with good prognosis and 26 samples of breast cancer with poor prognosis are determined, and the internal reference genes and differentially expressed genes are screened. Candidate molecular markers are verified with a large quantity of samples by reverse transcription fluorescent quantitative PCR. Finally, 14 genes and 5 internal reference genes (*BCL2*, *PGR*, *SCUBE2*, *ESR1*, *MKi67*, *CCNB1*, *MYBL2*, *GRB7*, *ERBB2*, *MMP11*, *CD68*, *BAG1*, *MAPT*, *MS4A1*; *ACTB*, *GAPDH*, *GUSB*, *NUP214*, *VCAN*) related to breast cancer prognosis were screened out, which were used to constitute a diagnostic kits. The diagnostic kit contains primers and probes of these genes, and other conventional reagents for qRT-PCR. The kit further comprises a prediction model, wherein the expression levels of *BCL2*, *PGR*, *SCUBE2*, *ESR1*, *MKi67*, *CCNB1*, *MYBL2*, *GRB7*, *ERBB2*, *MMP11*, *CD68*, *BAG1*, *MAPT* and *MS4A1* are tested by using the mean value of *ACTB*, *GAPDH*, *GUSB*, *NUP214* and *VCAN* as reference genes. At the same time, the clinical information such as the age of the breast cancer patients, PT and LN are used to comprehensively evaluate the prognosis risk of recurrence or mortality in 3-10 years after breast cancer surgery, and good prognosis or poor prognosis is determined.

20 6. Establishment of risk evaluation model

[0049] (1) Establishment of a model for predicting the risk of recurrence or mortality in 3-10 years after breast cancer surgery

[0050] Random forest algorithm in the machine learning method is used to evaluate and test the risk value of recurrence or mortality of the samples in 3-10 years after surgery, and to establish a gene assay model for breast cancer prognosis evaluation. The random forest is composed of many decision trees. The decision tree is constructed by the method of double randomness of attributes and samples, so it is also called random decision tree. In the random forest, there is no correlation between decision trees. When the test data enter the random forest, it is classified by each decision tree. Finally, the category with the most classification in all the decision trees is the final

result, i.e., the “voting” result of the decision tree. In other words, the random forest is a classifier including multiple decision trees, and the category of its output is determined by the mode of the category of the output of the individual tree. In the present disclosure, we have optimized the traditional random forest algorithm. We randomly sample for 1000 times and establish 1000 models. 39 optimal models having both higher accuracy and specificity from the 1000 model are selected as submodels of the final model, and the median of the 39 submodels are set as the final prediction result.

[0051] In the first phase of the study, total 192 candidate genes related to breast cancer were tested by TLDA technique. The gene expression differences between 26 breast cancer samples with good prognosis and 26 breast cancer samples with poor prognosis are detected, and the genes with different expression are screened. The gene expression difference is expressed as $2^{-\Delta Ct}$, wherein $\Delta Ct = Ct \text{ sample} - Ct \text{ reference}$, and the reference genes are used as a reference to normalize and calculate the relative expression level. The screening process of the internal reference gene comprises steps of: screening candidate internal reference genes using genorm, bestkeeper, normfinder, delta Ct based on the stability algorithm and considering the biological function of the less fluctuation genes and their relationship with tumor; calculating the correlation between the mean Ct of all candidate internal reference gene combinations and the mean Ct of 192 genes, and the combination with the highest correlation is the internal reference gene combination, which contains *ACTB*, *GAPDH*, *GUSB*, *NUP214* and *VCAN*. The screening criteria of the candidate gene include (1) overall analysis - the fold difference between the two groups of good prognosis and poor prognosis being 2 times or less than 0.5, and the proportion of cases with $Ct < 35$ being 50%; (2) stratified analysis – regarding no lymph nodes metastasis, the fold difference between the two groups of good prognosis and poor prognosis being over 2 times, and the statistical difference being <0.05 ; (3) in the overall analysis, the fold difference between the two groups is not significant, but it was reported in the literature related to the breast cancer prognosis, and the proportion of cases with $Ct < 35$ being 90%. The genes that meet the above screening criteria include *BCL2*, *PGR*, *SCUBE2*, *ESR1*, *MKI67*, *CCNB1*, *MYBL2*, *GRB7*, *ERBB2*, *MMP11*, *CD68*, *BAG1*, *MAPT* and *MS4A1*.

[0052] Firstly, the present disclosure provides prognostic evaluation genes for breast cancer patients in China. At present, the foreign similar products are developed based on European and

American people, and different ethnic groups have different gene expressions. In the present disclosure, 19 genes are screened out, wherein *MAPT* and *MS4A1* are genes related to the prediction of recurrence or mortality in 3-10 years after surgery based on data of breast cancer patients in China. Although the genes have been reported to be related to breast cancer, no direct report about its role in the prognosis of breast cancer. Secondly, the present disclosure establishes a new internal reference gene combination, which is different from other application and products, and is less affected by the RNA quality in FFPE samples, making the test result of the molecular markers more reliable. Thirdly, the prediction model based on the random forest algorithm is used for comprehensive analysis, which carries out the prediction of risk of recurrence or mortality in 3-10 years after surgery in the early or intermediate stage ER⁺ or PR⁺ breast cancer newly diagnosed patient.

[0053] In view of the above, the present disclosure provides a prognostic evaluation gene assay system for recurrence or mortality in 3-10 years after surgery in untreated stage I and stage II ER or PR positive breast cancer patients. Compared with the follow-up information, the system achieves an accuracy of 70%. Except for age, pT stage, and LN number of the patient, there is no need to rely on other clinical pathological information.

[0054] The raw materials and reagents used in the present disclosure of molecular markers, internal reference genes and application thereof, test kit and method for constructing test model are all commercially available.

[0055] The present disclosure is further illustrated below on the basis of examples.

Example 1 Collect sample and organize sample information

[0056] FFPE samples of untreated newly diagnosed breast cancer patients were chosen, and complete clinical follow-up data were systematically collected. Through organizing the sample information, 341 samples that meet the following criteria were selected as the experimental samples for TLDA (Taqman Low Density Array, TLDA) chips test and the subsequent series of qRT-PCR:

[0057] (1) untreated early or intermediate stage breast cancer newly diagnosed patients;

- [0058]** (2) untreated early or intermediate stage ER or (and) PR positive breast cancer patients;
- [0059]** (3) having or (and) having no lymph node metastasis, and the number of lymph node metastasis;
- [0060]** (4) having accurate and detailed follow-up information.

5

Example 2 Screen molecular markers and internal reference genes by TLDA chips

[0061] TLDA chips test was performed on 26 breast cancer samples with good prognosis and 26 breast cancer samples with poor prognosis that meet the above criteria, and the results were obtained. The specific steps were as follows.

10 **[0062]** (1) Extract RNA from the FFPE samples: 4 slices with a thickness of 20 μm or 8 slices with a thickness of 10 μm per sample were taken; RNA was extracted according to the protocol of High Pure FFPE RNA Isolation Kit (Roche); and the extracted RNA passed the quantity and quality control using NanoDrop-2000 was subjected to downstream reverse transcription experiment.

15 **[0063]** (2) Obtain cDNA sample by conducting reverse transcription reaction on the total RNA: 1 μg of total RNA was subjected to reverse transcription according to the protocol of SuperScript[®] VILO[™] Master Mix kit (Invitrogen).

[0064] (3) TLDA on the cDNA: The above cDNA products were thoroughly mixed with TaqMan[®] Universal PCR Master Mix, and then testing experiments were conducted on ABI 7900
20 fluorescence quantitative PCR instrument according to TLDA standard procedure.

[0065] (4) Analyze and process the data:

[0066] In the first phase of the study, total 192 candidate genes related to breast cancer were tested by TLDA technique. The gene expression differences between 26 breast cancer samples with good prognosis and 26 breast cancer samples with poor prognosis are detected, and the
25 genes with different expression are screened. The gene expression difference is expressed as $2^{-\Delta\text{Ct}}$, wherein $\Delta\text{Ct} = \text{CT sample} - \text{CT reference}$, and the reference genes are used as a reference to normalize and calculate the relative expression level. The screening process of the internal reference gene comprises steps of: screening candidate internal reference genes using genorm,

bestkeeper, normfinder, delta Ct based on the stability algorithm and considering the biological function of the less fluctuation genes and their relationship with tumor; calculating the correlation between the mean Ct of all candidate internal reference gene combinations and the mean Ct of 192 genes, and the combination with the highest correlation is the internal reference gene combination, which contains *ACTB*, *GAPDH*, *GUSB*, *NUP214* and *VCAN*. The screening criteria of the candidate gene include (1) overall analysis - the fold difference between the two groups of good prognosis and poor prognosis being 2 times or less than 0.5, and the proportion of cases with Ct < 35 being 50%; (2) stratified analysis – regarding no lymph nodes metastasis, the fold difference between the two groups of good prognosis and poor prognosis being over 2 times, and the statistical difference being <0.05; (3) in the overall analysis, the fold difference between the two groups is not significant, but it was reported in the literature related to the breast cancer prognosis, and the proportion of cases with Ct < 35 being 90%. The genes that meet the above screening criteria include *BCL2*, *PGR*, *SCUBE2*, *ESR1*, *MKi67*, *CCNB1*, *MYBL2*, *GRB7*, *ERBB2*, *MMP11*, *CD68*, *BAG1*, *MAPT* and *MS4A1*. The functions of the above 14 genes are shown in the following Table 1.

Table 1 Gene function analysis

No.	Gene Name	Functional Relevance
1.	<i>BCL2</i>	Estrogen related
2.	<i>PGR</i>	Estrogen related
3.	<i>SCUBE2</i>	Estrogen related
4.	<i>ESR1</i>	Estrogen related
5.	<i>MKi67</i>	Proliferation related
6.	<i>MKi67</i>	proliferation related
7.	<i>MYBL2</i>	proliferation related
8.	<i>GRB7</i>	Her-2 related
9.	<i>ERBB2</i>	Her-2 related
10.	<i>MMP11</i>	Invasion related
11.	<i>CD68</i>	Differentiation family 68
12.	<i>BAG1</i>	BCL2 bound anti-apoptotic gene 1
13.	<i>MAPT</i>	Microtubule- related protein tau

14.	<i>MS4A1</i>	Transmembrane 4 domain subfamily A member 1
-----	--------------	--

Example 3 Verify molecular markers by qRT-PCR with a large sample size

[0067] 14 molecular markers and 5 internal reference genes screened out by TLDA: *BCL2*, *PGR*, *SCUBE2*, *ESR1*, *MKI67*, *CCNB1*, *MYBL2*, *GRB7*, *ERBB2*, *MMP11*, *CD68*, *BAG1*, *MAPT*, *MS4A1*; *ACTB*, *GAPDH*, *GUSB*, *NUP214* and *VCAN*. A total of 289 FFPE samples that meet the above sample requirements and have complete clinical follow-up information were subjected to single-tube qRT-PCR verification.

[0068] (1) Extract RNA from 289 FFPE samples: 4 slices with a thickness of 20 µm or 8 slices with a thickness of 10 µm per sample were taken; RNA was extracted according to the protocol of High Pure FFPE RNA Isolation Kit (Roche); and the extracted RNA passed the quantity and quality control using NanoDrop-2000 was subjected to downstream reverse transcription experiment.

[0069] (2) Obtain cDNA sample of 289 FFPE samples by conducting reverse transcription reaction on the total RNA: 1 µg of total RNA was subjected to reverse transcription according to the protocol of SuperScript® VILO™ Master Mix kit (Invitrogen).

[0070] (3) qPCR on the cDNA from 289 FFPE samples: the above cDNA samples were thoroughly mixed with TaqMan® Universal PCR Master Mix, and then testing experiments were conducted on ABI 7900 fluorescence quantitative PCR instrument according to TLDA standard procedure. The sequences of qPCR primers and probes were shown in Tables 2 to 5.

Table 2 qRT-PCR primer sequences

Gene Name	NO.	Upstream Primer (5'-3')	NO.	Downstream Primer (5'-3')
<i>BCL2</i>	SEQ ID NO:1	CATGTGTGTGGAGAGC GTCAA	SEQ ID NO:2	GCCGGTTCAGGTACTCAGTC A
<i>PGR</i>	SEQ ID	CAGGTCTACCCGCCCT	SEQ ID	GGCTCTGGCTGGCTTCTG

	NO:3	ATCTC	NO:4	
<i>SCUBE 2</i>	SEQ ID NO:5	AAGGATCACGGCTGTA GTCACA	SEQ ID NO:6	AACCAGGCCTGCACTCACA
<i>ESR1</i>	SEQ ID NO:7	TGATCAGGTCCACCTT CTAGAATG	SEQ ID NO:8	GCGCCAGACGAGACCAAT
<i>MKi67</i>	SEQ ID NO:9	GCACACTCCACCTGTC CTGAA	SEQ ID NO:10	CCTGACTCTTGTTTTTCCTGAT GGT
<i>MKi67</i>	SEQ ID NO:11	CCAAGCCCAATGGAAA CATCT	SEQ ID NO:12	ATTACATCAGAGAAAGCCTG ACACA
<i>MYBL2</i>	SEQ ID NO:13	AAGACAACAGCTTGCT CAACCA	SEQ ID NO:14	GGGCCACTGCTGCCTTCT
<i>GRB7</i>	SEQ ID NO:15	TCCAGGACAAGGGCAC ACA	SEQ ID NO:16	ATGGCGTCTGAGCGAGAGA
<i>ERBB2</i>	SEQ ID NO:17	CCGAGGGCCGGTATAC ATT	SEQ ID NO:18	ATCCCACGTCCGTAGAAAGG T
<i>MMP11</i>	SEQ ID NO:19	GATGCCCAGGGCCACA T	SEQ ID NO:20	TTCACCGTCGTACACCCAGT AC
<i>CD68</i>	SEQ ID NO:21	AAGTTTCTCCTGCCCCA GTGA	SEQ ID NO:22	AAGAAGGATCAGGCCGATG AT
<i>BAG1</i>	SEQ ID NO:23	GCTGGGCTCACCGTGA CT	SEQ ID NO:24	GAGGTAACATGAAGGTCGTG CTT
<i>MAPT</i>	SEQ ID NO:25	ACCAAGACCAAGAGGG TGACA	SEQ ID NO:26	CAGAGCCCGGTTTCCTCAGA
<i>MS4A1</i>	SEQ ID NO:27	AACTCCCATCTACCC AATACTGTT	SEQ ID NO:28	ATCACTGACAAAATGCCCAA GA

Table 3 qRT-PCR probe sequences

Gene Name	Sequence Number	Probe Sequence
<i>BCL2</i>	SEQ ID NO:29	TGGACAACATCGCCCTG
<i>PGR</i>	SEQ ID NO:30	ACTACCTGAGGCCGGAT

<i>SCUBE2</i>	SEQ ID NO:31	TGCAAGGAGGCCC
<i>ESR1</i>	SEQ ID NO:32	CTGGCTAGAGATCCTGAT
<i>MKi67</i>	SEQ ID NO:33	TCATCAAGGAACAGCCTC
<i>MKi67</i>	SEQ ID NO:34	ATGTGCCCCTGCAGAA
<i>MYBL2</i>	SEQ ID NO:35	CTTCTTGCAGGCCAAG
<i>GRB7</i>	SEQ ID NO:36	CTGGTTCCGTTAAGCC
<i>ERBB2</i>	SEQ ID NO:37	CGCCAGCTGTGTGAC
<i>MMP11</i>	SEQ ID NO:38	TGGTTCTTCCAAGGTGCT
<i>CD68</i>	SEQ ID NO:39	CCATCTTGCTGCCTCT
<i>BAG1</i>	SEQ ID NO:40	TCACCCACAGCAATGA
<i>MAPT</i>	SEQ ID NO:41	CCTGAAAGAATCTCC
<i>MS4A1</i>	SEQ ID NO:42	CAGCATACAATCTCTG

Table 4 qRT-PCR primer sequences for housekeeping genes

Gene Name	NO.	Upstream Primer (5'-3')	NO.	Downstream Primer (5'-3')
<i>ACTB</i>	SEQ ID NO:43	AGGCACCAGGGCGTGA T	SEQ ID NO:44	CTCGTCGCCCACATAGGAA
<i>GAPDH</i>	SEQ ID NO:45	CAACGGATTTGGTCGT ATTGG	SEQ ID NO:46	TGATGGCAACAATATCCACT TTACC
<i>GUSB</i>	SEQ ID NO:47	TGAAGATGGTGATCGC TCACA	SEQ ID NO:48	GTTAGAGTTGCTCACAAAGG TCACA
<i>NUP214</i>	SEQ ID NO:49	ATCAGGCACAGGAGGG TTCA	SEQ ID NO:50	CCACCAAAACCCTGGACAGA
<i>VCAN</i>	SEQ ID NO:51	ACTACAGGGCGGCAAC CA	SEQ ID NO:52	AACGTCCAAACAAGCCTTCT GA

Table 5 qRT-PCR probe sequences for housekeeping genes

Gene Name	Sequence	Probe Sequence
-----------	----------	----------------

Number		
<i>ACTB</i>	SEQ ID NO:53	TGGGCATGGGTCAGAA
<i>GAPDH</i>	SEQ ID NO:54	AGGGCTGCTTTTAACTC
<i>GUSB</i>	SEQ ID NO:55	ATCCTTGGACCCCTCC
<i>NUP214</i>	SEQ ID NO:56	CTTTGGGTCAAATAAC
<i>VCAN</i>	SEQ ID NO:57	CACACTGAATTTTGAGGCTG

Example 4 Establish model for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery

[0071] Random forest algorithm in the machine learning method is used to evaluate and test the risk value of recurrence or mortality of the samples in 3-10 years after surgery, and to establish a gene assay model for breast cancer prognosis evaluation. The random forest is composed of many decision trees. The decision tree is constructed by the method of double randomness of attributes and samples, so it is also called random decision tree. In the random forest, there is no correlation between decision trees. When the test data enter the random forest, it is classified by each decision tree. Finally, the category with the most classification in all the decision trees is the final result, i.e., the “voting” result of the decision tree. In other words, the random forest is a classifier including multiple decision trees, and the category of its output is determined by the mode of the category of the output of the individual tree. In the present disclosure, we have optimized the traditional random forest algorithm. We randomly sample for 1000 times and establish 1000 models. 39 optimal models having both higher accuracy and specificity from the 1000 model are selected as submodels of the final model, and the median of the 39 submodels are set as the final prediction result.

Example 5 Further verify by independent clinical samples

[0072] 19 FFPE samples from ER or PR positive early or intermediate stage breast cancer patients with known clinical follow-up data: PT stage was stage 1 and 2, surgeries were operated between 2004 and 2008, followed-up was carried out from 2011 to 2015 and the follow-up period was over 3-10 years.

[0073] Total RNA of the above 19 FFPE samples was extracted using High Pure FFPE RNA Isolation Kit (Roche). After quality control as being qualified, the RNA was subjected to reverse transcription reaction to obtain cDNA sample. The cDNA samples were subjected to qRT-PCR reaction to detect the internal reference genes *ACTB*, *GAPDH*, *GUSB*, *NUP214*, *VCAN*, and genes *BCL2*, *PGR*, *SCUBE2*, *ESR1*, *MKi67*, *CCNB1*, *MYBL2*, *GRB7*, *ERBB2*, *MMP11*, *CD68*, *BAG1*, *MAPT* and *MS4A1*. The Ct values were introduced into the model for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery constructed by random forest algorithm and the risk values of recurrence or mortality in 3-10 years were obtained. The results of being a good prognosis or a poor prognosis were predicted based on the risk threshold. The concordance rate between the predicted results and the known follow-up information was 73.6%. The detailed results were shown in Table 6.

Table 6 Results of prognosis evaluation of 19 breast cancer patients

Sample number	Surgery date	Follow-up date	Time of recurrence and metastasis	Prediction result
43200	2004-6-1	2015-11-6	No recurrence during the follow-up	Good prognosis
43346	2004-8-1	2010-5-1	No recurrence during the follow-up	Poor prognosis
45110	2004-12-1	2011-1-1	2008/06: Liver metastasis	Poor prognosis
50170	2005-1-13	2014-11-1	No recurrence during the follow-up	Poor prognosis
53552	2005-8-1	2010-2-13	2006/9: Left chest wall recurrent poorly differentiated adenocarcinoma; 2006/10: right lung metastasis; 2009/02: double lung metastasis	Poor prognosis
53827	2005-8-18	2014-11-1	No recurrence during the follow-up	Good prognosis

54564	2005-9-28	2009-08-28	2006/12: Right axilla recurrent poorly differentiated adenocarcinoma; 2007/12: double neck, right lung, mediastinal metastasis after right breast cancer surgery	Poor prognosis
54674	2005-10-10	2014-11-1	No recurrence during the follow-up	Good prognosis
54775	2005-10-13	2007-7-27	2006/6/6: Recurrent poorly differentiated adenocarcinoma; 2006/11: right axillary, double interclavicle, left pleural	Poor prognosis
54938	2005-10-20	2014-11-1	No recurrence during the follow-up	Poor prognosis
55335	2005-11-11	2008-11-19	2007/12: Liver metastasis	Poor prognosis
55475	2005-11-22	2012-4-12	2009/12/10: Liver metastasis; 2010/10/28: liver and right cervical lymph node metastasis; 2011/11/11: bone metastasis	Good prognosis
55706	2005-12-8	2014-11-1	No recurrence during the follow-up	Good prognosis
55709	2005-12-12	2015-11-6	2009/2: Recurrent invasive ductal carcinoma involving the chest muscle; 2009/2: bone metastasis; 2009/6: right axillary lymph node metastasis	poor prognosis
55792	2005-12-15	2009-4-1	2008/01: multiple bone metastases	Poor prognosis
60546	2006-2-16	2013-6-1	No recurrence during the follow-up	Good prognosis

70466	2007-1-29	2014-11-1	No recurrence during the follow-up	Good prognosis
76032	2007-11-16	2014-11-1	No recurrence during the follow-up	Good prognosis
86967	2008-12-25	2015-3-29	No recurrence during the follow-up	Good prognosis

Table 7

Verification results from 19 independent samples						
Sample	Pathological prediction			Evaluation test system for breast cancer prognosis		
Type	Good prognosis	Poor prognosis	Predictive outcome	Good prognosis	Poor prognosis	Predictive outcome
Good prognosis	7	4	72.70%	8	3	72.70%
(follow-up)			(specificity)			(specificity)
Poor prognosis	1	7	87.50%	1	7	87.50%
(follow-up)			(sensitivity)			(sensitivity)

Example 6

- 5 At present, the decision of clinical treatment and treatment plan for breast cancer ultimately depends on the results of pathological examination. At the same time, the results of pathological examination are the most important objective basis for the prediction of prognosis. The test system used 289 FFPE samples from newly diagnosed breast cancer patients with known clinical follow-up data collected by Tianjin Medical University Cancer Hospital and Henan Cancer
- 10 Hospital were tested by the evaluation system of the present disclosure on 5 internal reference genes and 14 molecular markers.

[0074] The test results are shown in Table 8.

Table 8

Verification results from 289 samples (large sample size)						
Sample	Pathological prediction			Evaluation test system for breast cancer prognosis		
Type	Good prognosis	Poor prognosis	Predictive outcome	Good prognosis	Poor prognosis	Predictive outcome
Good prognosis	128	50	71.90%	142	33	81.14%
(follow-up)			(specificity)			(specificity)
Poor prognosis	48	63	56.77%	52	62	54.39%
(follow-up)			(sensitivity)			(sensitivity)

[0075] The kit of the present disclosure provided an accuracy rate of 81.1% for newly diagnosed breast cancer patients with low risk of recurrence or mortality in 3-10 years, while the accuracy rate provided by the pathological test was 71.9%. The sensitivity of the accuracy rate provided by the kit was 54.4% for newly diagnosed breast cancer patients with high risk of recurrence or mortality in 3-10 years, which was close to the accuracy rate of the corresponding pathological prediction of 56.8%. Compared with the clinical follow-up information, the kit achieves a concordance rate of 70%. Except for age, pT stage, and LN number of the patients, there is no need to rely on other clinical pathological information.

[0076] The test system and kit has a superior prognosis evaluation test performance for breast cancer as compared with the clinical pathological evaluation results. It can to a certain extent reduce the over-treatment and improper treatment caused by pathological diagnosis errors, meeting the requirements of individualized and precise treatment of breast cancer patients and further improving the technical methods for the prognosis prediction of breast cancer in China.

[0077] The molecular markers and application thereof, test kit and method for constructing test model provided by the present disclosure are described in detail above. The principles and embodiments of the present disclosure have been described with reference to specific examples, and the description of the above embodiments is only to assist in understanding the method of the

present disclosure and its core idea. It should be noted that those skilled in the art can make various modifications and changes to the present disclosure without departing from the principle of the present disclosure, which fall within the protection scope of the claims of the present disclosure.

What is claimed is,

1. A gene composition, comprising molecular marker(s) *MAPT* and/or *MS4A1*.

5 2. A gene composition, comprising molecular markers *BCL2*, *PGR*, *SCUBE2*, *ESR1*, *MKi67*, *CCNB1*, *MYBL2*, *GRB7*, *ERBB2*, *MMP11*, *CD68*, *BAG1*, *MAPT* and *MS4A1*.

3. The gene composition according to claim 2, further comprising internal reference genes *ACTB*, *GAPDH*, *GUSB*, *NUP214* and *VCAN*.

10

4. Use of the gene composition according to claims 1 or 2 or 3 for the manufacture of a device for risk prediction of recurrence and/or mortality in 3-10 years after breast cancer surgery.

15 5. The use according to claim 4, wherein the risk prediction of recurrence and/or mortality in 3-10 years after breast cancer surgery is performed by:

obtaining total RNA of a test sample, obtaining cDNA by reverse transcription, obtaining Ct values of the molecular marker and the internal reference genes by a fluorescent quantitative PCR, calculating average Ct value of the internal reference genes to obtain an average Ct value (Ct') of internal reference gene combination, normalizing Ct value of the molecular marker respectively
20 by subtracting Ct' value of the internal reference gene combination to obtain Δ Ct, and then analyzing Δ Ct value, and age, pT value and LN value of the subject by a model for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery constructed by random forest algorithm to obtain a result.

25 6. The use according to claim 5, wherein the constructing of the model for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery is performed by:

constructing a matrix by using Δ Ct value of the molecular marker of the test sample, and age,

pT value and LN value of the subject, randomly selecting half as a training set and another half as a verification set, establishing a prediction model containing 10,000 decision trees by random forest algorithm, randomly sampling for ≥ 1000 times and establishing ≥ 1000 prediction models, selecting ≥ 39 optimal models having the highest concordance rate with follow-up information from the ≥ 1000 prediction models as submodels of final model, and setting the median of the ≥ 39 submodels as a final prediction value of prognostic risk.

7. The use according to claims 5 or 6, wherein the test sample is a FFPE sample from an untreated early or intermediate stage ER or PR positive breast cancer newly diagnosed patient.

8. A primer set for amplifying the gene composition according to claims 1 or 2, having sequences represented by SEQ ID NO: 1 to SEQ ID NO: 28.

9. A probe set for amplifying the gene composition according to claims 1 or 2, having sequences represented by SEQ ID NO: 29 to SEQ ID NO: 42.

10. A primer set for amplifying the internal reference genes of the gene composition according to claim 3, having sequences represented by SEQ ID NO: 43 to SEQ ID NO: 47.

11. A probe set for amplifying the internal reference genes of the gene composition according to claim 3, having sequences represented by SEQ ID NO: 48 to SEQ ID NO: 52.

12. A test kit for risk prediction of recurrence and/or mortality in 3-10 years after breast cancer surgery, comprising the primer set according to claims 8 and/or 10 and/or the probe set according to claims 9 and/or 11, and reagents commonly used in a kit.

13. A method for constructing the model for risk prediction of recurrence or mortality in 3-10

years after breast cancer surgery, comprising:

constructing a matrix by using ΔCt value of the molecular marker of the test sample, and age, pT value and LN value of the subject, randomly selecting half as a training set and another half as a verification set, establishing a prediction model containing 10,000 decision trees by random forest algorithm, randomly sampling for ≥ 1000 times and establishing ≥ 1000 prediction models, selecting ≥ 39 optimal models having the highest concordance rate with follow-up information from the ≥ 1000 prediction models as submodels of final model, and setting the median of the ≥ 39 submodels as a final prediction value of prognostic risk.

14. A method for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery, comprising:

obtaining total RNA of a test sample, obtaining cDNA by reverse transcription, obtaining Ct values of the molecular marker and the internal reference genes by a fluorescent quantitative PCR, calculating average Ct value of the internal reference genes to obtain an average Ct value (Ct') of internal reference gene combination, normalizing Ct value of the molecular marker respectively by subtracting Ct' value of the internal reference gene combination to obtain ΔCt , and then analyzing ΔCt value, and age, pT value and LN value of the subject by a model for risk prediction of recurrence or mortality in 3-10 years after breast cancer surgery constructed by random forest algorithm to obtain the result.

15. The method according to claims 13 or 14, wherein the test sample is a FFPE sample from an untreated early or intermediate stage ER or PR positive breast cancer newly diagnosed patient.