

DESCRIPTION

Title of Invention

MACHINE LEARNING-BASED METHOD FOR PREDICTION OF BREAST CANCER PROGNOSIS USING NEXT-GENERATION SEQUENCING, AND PREDICTION SYSTEM THEREFOR

Technical Field

The technique as described below relates to a technique for predicting breast cancer prognosis using gene expression data.

Background Art

Various studies have been conducted to predict tumor prognosis.

For example, in the breast cancer field, a technique for predicting breast cancer prognosis through genetic analysis has been developed. Typically, Oncotype DX[®], Mammaprint[®], and the like, which are used as standard tools for predicting breast cancer prognosis, are all real-time PCR (RT-PCR)-based tests.

Technical Problem

An RT-PCR-based test has a limitation in that it is difficult, in terms of cost and efficiency, to analyze many genes at the same time. The technique as described below is intended to provide a technique that allows for prediction of breast cancer prognosis through analysis of gene expression levels with a next-generation sequencing (NGS) technique.

Solution to Problem

A machine learning-based method for prediction of breast cancer prognosis using next-

generation sequencing comprises the steps of: allowing a computer device to use RNA sequencing data of a subject's tissue, to measure an expression level of a target gene; allowing the computer device to input the expression level of the target gene into an artificial neural network adapted beforehand; and allowing the computer device to estimate breast cancer prognosis for the subject based on an output value from the artificial neural network.

A machine learning-based system for prediction of breast cancer prognosis using next-generation sequencing comprises a client device for storing RNA gene data of a subject's tissue; and an analysis server in which RNA sequencing data, which is obtained by performing next-generation sequencing (NGS)-based sequencing on the gene data, is used to measure an expression level of a target gene, the expression level of the target gene is input into an artificial neural network adapted beforehand, and breast cancer prognosis for the subject is estimated based on an output value from the artificial neural network.

The artificial neural network is adapted beforehand to have, as input values, expression levels of target genes for a plurality of samples and to output a result depending on Oncotype DX recurrence scores for the plurality of samples.

Advantageous Effects of Invention

The technique as described below uses an NGS technique, which makes it possible to rapidly analyze a target gene sequence at a lower cost than an RT-PCR-based technique. The technique as described below uses a machine learning model that has been trained with expression levels of target genes and Oncotype DX recurrence scores, which makes it possible to predict breast cancer prognosis in a rapid and accurate way.

Brief Description of Drawings

FIG. 1 illustrates an example of a flow chart for a machine learning-based method for prediction of breast cancer prognosis using next-generation sequencing.

FIG. 2 illustrates an example showing target genes.

FIG. 3 illustrates an example of quality of NGS-based data for target genes.

FIG. 4 illustrates an example of using gene expression data of a specific target exon region.

FIG. 5 illustrates an example of verifying stability of a method in which specific target exon regions are utilized.

FIG. 6 illustrates an example of a process for normalizing RNA sequencing data.

FIG. 7 illustrates an example of an artificial neural network structure for machine learning.

FIG. 8 illustrates an example of a machine learning-based system for prediction of breast cancer prognosis using next-generation sequencing.

Detailed Description of Invention

The technique as described below relates to a technique for predicting breast cancer prognosis using gene expression data. A machine learning model is used as a tool for predicting breast cancer prognosis. The machine learning model is trained using gene expression levels and Oncotype DX[®] (hereinafter referred to as Oncotype DX) recurrence scores (hereinafter referred to as RS). First, Oncotype DX is briefly described.

Oncotype DX is an analysis tool developed by Genomic Health, USA. Oncotype DX is a test that measures and analyzes activity of 21 different genes in breast cancer tissue to predict the likelihood of breast cancer recurrence and how effective chemotherapy will be.

In Oncotype DX, RS is calculated based on 16 genes and 5 reference genes. The expression used to calculate RS is as follows. $RS = +0.47 \times \text{HER2 Group Score} - 0.34 \times \text{Estrogen (ER) Group Score} + 1.04 \times \text{Proliferation Group Score} + 0.10 \times \text{Invasion Group Score} + 0.05 \times \text{CD68} - 0.08 \times \text{GSTM1} - 0.07 \times \text{BAG1}$. Here, each item indicates a gene group classified according to functional classification. The HER2 Group (2 genes), ER Group (4 genes), Proliferation Group (5 genes), and Invasion Group (2 genes) include a plurality of genes. CD68, GSTM1, and BAG1 are each one individual gene. The RS score has a value ranging from 0 to 100. For example, in a case where the RS score is low, it is highly likely that recurrence rate is low and an effect of chemotherapy is also low. On the contrary, in a case where the RS score is high, it is highly likely that recurrence rate is high and chemotherapy is

effective.

As described above, Oncotype DX detects activity of genes based on RT-PCR. However, the technique as described below uses an NGS technique. In the technique as described below, RNA sequencing is performed based on an NGS technique, to determine a gene expression level. The NGS technique is also briefly described.

The NGS technique has multiplexing capacity to perform hundreds of thousands of reactions at the same time, and enables sequencing with a small amount of sample. In NGS techniques, their specific application method is somewhat different depending on commercially available techniques; however, in general, clonal amplification, massively parallel sequencing, and a new sequencing method with a different mechanism of action from the Sanger's method are used. Commercially available techniques are briefly described as follows. Roche merged with 454 Life Sciences in 2007 and launched the 454 GS improved FLX model sequencer. Illumina, Inc. launched Genome Analyzer HiSeq in 2006, and then Applied Biosystems launched SOLiD in 2007. The three platforms have the following in common: they abandoned a complicated library construction and cloning process; adopted a clonal amplification technique; adopted a massively parallel sequencing technique that enables bulk processing at once; and performed sequencing with a sequencing by synthesis approach through cyclic sequencing, thereby eliminating a complex electrophoresis process. In addition, such platforms use an algorithm in which shorter reads that have been read using the shotgun method are aligned by a computer and overlapped portions are searched for to complete the whole.

The technique as described below predicts breast cancer prognosis based on a gene expression level (RNA expression level) for a sample extracted from specific tissue (breast tissue). First, an example of a process, in which a sample is prepared and RNA is extracted in a specific operation, will be described below.

Selection of target breast cancer patient and preparation of tissue for examination

1) Select a representative formalin-fixed paraffin-embedded (FFPE) block among tissues from surgery of stage 1-2 breast cancer which is hormone receptor-positive and lymph node metastasis-negative.

- The block is selected under identification of H&E-stained slide by a pathologist.

Here, a block in which a target tumor is definitely present should be selected; and if possible, it is preferred to select a block that has the largest tumor area in one cross-section and has little or no necrotic parts in tumor tissue.

2) Prepare 10 unstained slides with a thickness of 10 μm .

Protocol for extracting RNA from FFPE tissue

1) RNA extraction kit: One of the two commercially available kits below may be used.

① Ambion RecoverAll™ Total Nucleic Acid Isolation Kit for FFPE

② QIAGEN RNeasy FFPE Kit

2) Preparation of wash solution

– Mix 42 mL of 100 mL ethanol with Wash 1 -> Wash 1

– Mix 48 mL of 100 mL ethanol with Wash 2/3 -> Wash 2/3

3) Deparaffinization (preparation: tissue, 100% xylene, 100% ethanol, heating block at 50°C, pipette, vortex mixer, centrifuge)

① Tissue preparation: Prepare 4 to 8 paraffin sections with a thickness of 10 μm , which are 40 to 80 μm in total, cut from a paraffin block. Here, in a case where the tumor size in the section is smaller than 40 mm^2 , use all 8 sections. Place the prepared section in a 1.5 mL tube with caution so that if possible, only a tumor portion of the section is introduced thereinto.

② Add 1 mL of 100% xylene to the tissue, perform mixing with a vortex mixer, and perform centrifugation briefly. Then, allow it to stand at 50°C for 3 minutes so that the paraffin melts (if the paraffin does not melt, repeat this process)

③ Perform centrifugation at maximum speed for 2 minutes so that a mass is formed. If the mass is not tightly packed, add 2-minute centrifugation. Remove xylene without breaking the mass.

④ Washing of xylene

a. Add 1 mL of 100% ethanol to the sample and perform mixing with a vortex mixer

(the sample becomes cloudy).

b. Perform centrifugation at maximum speed for 1 minute at room temperature so that a mass is formed.

c. Remove ethanol without breaking the mass.

d. Repeat processes a to c once.

e. Perform centrifugation briefly, and then remove the remaining ethanol as much as possible without touching the mass.

⑤ Perform drying at room temperature for 15 to 45 minutes.

4) Protein degradation (preparation: heating blocks at 50°C & 80°C, protease is taken out of the freezer and allowed to melt at room temperature)

① Add 200 µl of digestion buffer and 4 µl of protease to each sample. Here, shake it gently to be mixed well.

② Place the sample in the heating block of 50°C (protease activation temperature) for 15 minutes or longer until it becomes completely clear.

③ Then, place the sample in the heating block at 80°C (protease inactivation temperature) for 15 minutes. Here, keep the time exactly.

④ If melting does not occur, repeat the above processes (② & ③) with addition of 4 µl of protease only.

5) Isolation of nucleic acid (preparation: isolation additive/ethanol mixture, all other reagents)

① Preparation of isolation additive/ethanol mixture

- 240 µl of isolation additive + 500 µl of 100% ethanol = 790 µl in total

- After preparation, store it in a 50 mL tube.

(In a case where multiple samples are intended, prepare approximately 5% more than the indicated amount.)

② Dispense 790 µl of the prepared isolation additive/ethanol mixture into a tube containing each sample while performing mixing with a pipette.

③ Filtration of mixture

- a. Place a filter cartridge in a tube provided in the kit.
- b. Place 700 µl of the mixture obtained in process ② on the filter and close the lid.
- c. Perform centrifugation at 10,000 rpm for 30 seconds.
- d. Discard the filtered solution and place the filter in the same tube.
- e. If necessary (in a case where the mixture is not sufficiently filtered), perform centrifugation once again so that the mixture is filtered through the filter.

④ Wash 1

- a. Add 700 µl of Wash 1 to the filter cartridge
- b. Perform centrifugation at 10,000 rpm for 30 seconds.
- c. Discard the filtered solution and place the filter in the same tube.

⑤ Wash 2/3

- a. Add 500 µl of Wash 2/3 to the filter cartridge
- b. Perform centrifugation at 10,000 rpm for 30 seconds.
- c. Discard the filtered solution and place the filter in the same tube.
- d. Perform centrifugation at 10,000 rpm once again to remove the remaining solution.

6) RNA isolation and purification (preparation: DNase (DNA degrading enzyme) and nuclease (nucleic acid degrading enzyme) are taken out of the freezer and allowed to melt)

① RNA isolation

- a. Preparation of DNase mixture: 6 µl of 10X DNase buffer + 4 µl of DNase + 50 µl of nuclease-free water = 60 µl in total
- b. Add 60 µl of DNase mixture to the center of each filter cartridge.

c. Close the lid and allow it to stand at room temperature of 22°C to 25°C for 30 minutes.

② Wash 1

a. Add 700 µl of Wash 1 to the filter cartridge and allow it to stand at room temperature for 30 to 60 seconds.

b. Perform centrifugation at 10,000 rpm for 30 seconds.

c. Discard the filtered solution and place the filter in the same tube.

③ Wash 2/3

a. Add 500 µl of Wash 2/3 to the filter cartridge.

b. Perform centrifugation at 10,000 rpm for 30 seconds.

c. Discard the filtered solution and place the filter in the same tube.

d. Repeat steps a to c once again.

e. Perform centrifugation at 10,000 rpm for 1 minute.

④ Addition of elution solution and storage

a. Place the filter cartridge in a new tube.

b. Add 60 µl of elution solution to the center of the filter.

c. Close the lid and allow it to stand for 1 minute.

d. Perform centrifugation at maximum speed for 1 minute; then discard the filter and store the filtered solution at -20°C or lower.

Hereinafter, a process for predicting breast cancer prognosis using RNA extracted from the sample tissue will be described. FIG. 1 illustrates an example of a flow chart for a machine learning-based method for prediction of breast cancer prognosis using next-generation sequencing. First, sequencing of the RNA sample is performed (110). The RNA sequencing may be performed in a variety of ways. The RNA sequencing may be performed using a variety of commercially available kits and commercially available solutions. One example is described.

Targeted RNA-sequencing

1) Using the KAPA Stranded RNA-Seq Kit with RiboErase (KK8483, Kapa Biosystems) kit, ribosomal RNA is removed from total RNA.

2) cDNA is prepared from mRNA and cDNA NGS library is completed through additional processes. Using the cDNA Library, Hybridization solution, and Target Capture Probe, the solution-based hybridization capture method is performed.

3) Library amplification is performed by partial amplification of the obtained product.

4) RNA expression level can be predicted based on sequencing depth data of the target region generated by sequencing of the final product (using a kit of Illumina, Inc.).

In a case where NGS-based RNA sequencing is performed, the RNA sequence is converted into certain digital data in a commercially available program. In addition, the commercially available program is capable of calculating the expression level of each RNA gene using the sequencing result. Thus, the process of analyzing the RNA sample is performed by a computer device and the analysis result is obtained accordingly. Therefore, hereinafter, it is assumed that the computer device performs prediction of breast cancer prognosis using an RNA expression level. The computer device generates expression level data for the sample RNA (120).

Meanwhile, without analyzing the entire gene of the sample RNA, a gene associated with breast cancer prognosis (hereinafter referred to as target gene) may be selected and analyzed. Publicly available genetic data were utilized to determine the target gene. The genetic data of subjects who are estrogen receptor-positive and have no lymph node metastasis were utilized. The publicly available data utilized are GSE2034, GSE2990, GSE3494, GSE4922, GSE6532, GSE7390, and GSE12093.

Based on the publicly available data, a correlation between each gene and Oncotype DX RS (recurrence score) was analyzed. Pearson and Spearman techniques, which are representative correlation analysis techniques, were utilized. Among respective publicly available data, genes with an average correlation coefficient of 0.5 or higher with respect to Oncotype DX RS were selected. A total of 135 genes were selected as shown in the table below.

[Table 1]

	GSE2034	GSE2990	GSE3494	GSE4922	GSE6532	GSE7390	GSE12093	Total (>0.5)
Oncotype DX RS	184	298	267	292	168	77	56	135

Further, 16 genes used for calculation of Oncotype DX RS were additionally selected. As a result, 149 genes were utilized as target genes. Of course, experimentally, a combination of different genes may be selected as the target gene. However, a gene having a high correlation with respect to Oncotype DX RS is determined as the target gene.

FIG. 2 illustrates an example showing target genes. In FIG. 2, a total of 149 genes are indicated. In FIG. 2, the genes indicated in shade are genes associated with a mechanism (cell cycle) that regulates cell division. In FIG. 2, the genes indicated in solid circle are genes that are involved in the p53 signaling pathway and thus associated with a mechanism that regulates cell division. In FIG. 2, the genes indicated in dotted circle are genes that regulate DNA replication. In FIG. 2, the genes indicated in solid square are genes involved in cell cycle and p53 signaling pathway. In FIG. 2, the genes indicated in dotted square are genes involved in cell cycle and DNA replication.

As described above, the computer device measures an expression level of the target gene based on NGS. FIG. 3 illustrates an example of quality of NGS-based data for the target genes. FIG. 3 illustrates an example of a graph showing quality of data obtained by targeted RNA-sequencing. The results are obtained by identifying deviation between measured values for gene expression level and values measured by whole-transcript sequencing, to verify data quality. For a total of 84 panel genes, expression levels of 10 RNA samples were calculated by targeted sequencing and whole-transcript sequencing, respectively, and the Pearson's correlation coefficient was measured. As a result, a high correlation of 0.85 or higher was identified. Therefore, from the viewpoint that a high correlation is observed between the targeted RNA-sequencing results and the whole-transcript RNA sequencing results, it is interpreted that the results of the targeted RNA-sequencing used in the experiment have data quality similar to the results of the whole-transcript RNA sequencing.

A specific exon site that is commonly expressed in respective samples may be used instead of calculating the entire gene expression level for respective samples. In this case, a faster gene expression level calculation process may be performed. That is, this method uses a panel of commonly conserved exon for the samples (hereinafter referred to as CE method).

FIG. 4 illustrates an example of using gene expression data of a specific target exon region. FIG. 4 illustrates an expression state of gene exons for a plurality of samples (A to D).

Regions (target regions) that are commonly expressed with respect to a plurality of samples may be determined, and gene expression levels may be determined based on the regions. In the method (hereinafter referred to as WG method) using the whole-genome region, difference in expression levels of the individual's gene subtype transcripts is not considered, and thus deviation may occur at the time of measuring gene expression levels due to difference in subtype transcript expression rates among respective patients. The CE method uses only the regions shared by subtype transcripts, and thus no measurement deviation occurs unlike the WG method, which enables more stable expression level measurement. Stability refers to the degree to which measured values vary during repeated measurements.

Furthermore, stability of the CE method was verified using sample data. FIG. 5 illustrates an example of verifying stability of a method in which specific target exon regions are utilized. FIG. 5(a) illustrates an example showing a process for verifying stability of the CE method. Replicate data are values obtained by performing repeated measurements for a single sample. Thus, errors between the data obtained by performing repeated measurements correspond to measurement errors. A total of 18 target RNA sample sequencing data were generated by performing 2 repeated experiments on each of 9 RNA samples extracted from tumor cell lines (targeted RNA-sequencing I and targeted RNA-sequencing II). FIG. 5(b) illustrates experimental results showing stability between the measured values according to the WG method and between the measured values according to the CE method. In a case where the replicate data are compared, 8 out of a total of 9 samples showed that the CE method is more stable (showing higher Pearson's coefficient) than the WG method. The average Pearson's coefficient for 9 samples is also higher in the CE method than in the WG method, and the CE method shows less standard deviation between the samples. Therefore, it is interpreted that the CE method is more stable than the WG method.

Referring back to FIG. 1, the computer device may regularly perform post-processing and normalization on the expression level data generated for target genes (130). First, an applicable example of a post-processing process will be described. Post-processing and normalization correspond to a process of regularly processing digital data (files).

Post-processing of targeted RNA-sequencing results

1) Read removal

Reads that do not meet the read quality criteria are removed (for example, processing is performed so that bases having an average quality of 20 or higher and having an average quality of 2 or lower are less than 5%).

The Trimmomatic (0.33) program is used to remove index sequences inserted during the sequencing process.

2) Alignment of sequenced reads to reference genome

The STAR aligner program is used to find locations of the sequenced reads based on the reference genome (hg19), and simultaneous alignment is performed with the SortedByCoordinate option.

3) Calculation of gene expression level

The Cufflinks program is used to calculate the expression level for each gene and the expression level for each transcript from the information of aligned reads. The expression level may be calculated as an FPKM (Fragments Per Kilobase of exon per Million fragments mapped) value. The expression level calculated for each gene may be generated as the file `gene.fpkm_tracking`, and the expression level calculated for each transcript may be generated as the file `isoforms.fpkm_tracking`.

Meanwhile, for accurate sequence alignment and expression level measurement, pre-analysis processing may be performed so that among the reads resulting from sequencing, reads with poor base quality are removed, and index sequences (inserted during the sequencing process) which may remain at the ends of each read are removed. For the pre-processed reads, the STAR program is used to identify locations of the respective reads on the reference genome. The identified location information is generated in the BAM file format, and this BAM file may be used to calculate the respective expression levels for each gene and each transcript using the Cufflinks program.

The computer device may perform normalization on the generated data (130).

Normalization of expression information from targeted RNA-sequencing

Among conventional normalization techniques, the "Trimmed Mean of M-value (TMM)" technique used in R package edgeR (Robinson et al. Bioinformatics 2010) is known to have the highest stability. In the computer device, the edgeR package, a commercially available package, may be installed to design a pipeline that automatically extracts normalized gene expression information from targeted RNA-sequencing data generated.

Sequencing data generated using an NGS technique is mapped to a reference genome using an alignment software (for example, RNA-STAR). The number of sequences coming from each gene may be counted through the mapping result, which is a direct estimate for the gene expression level.

The normalization pipeline receives data processed in the BAM file format after completion of the mapping. The mapped data may be calculated as normalized expression level values, which allow for comparison between samples, by a series of software packages, HTseq-count and edgeR, embedded in the pipeline.

FIG. 6 illustrates an example of a process for normalizing RNA sequencing data. FIG. 6 illustrates an example of a process for normalizing data for two different samples. First, the computer device receives, as sample data, the mapped data representing gene expression levels. The computer device calculates the gene expression levels by applying HTseq-count. Each sample has a different library size (100 sequences for Sample 1, and 300 sequences for Sample 2). The computer device then applies edgeR to normalize the expression levels. FIG. 6 illustrates an example of normalizing expression levels of a sample based on a library size of 100.

The computer device inputs the gene expression level data into a machine learning model adapted beforehand (140). The gene expression data input into the machine learning model is normalized data. The machine learning model has been trained beforehand with expression levels of target genes and Oncotype DX RS. For example, the machine learning model may be trained to determine, as high risk (high likelihood of recurrence), a case where the Oncotype DX RS is 25 or higher, and determine, as low risk, a case where the Oncotype DX RS is lower than 25.

FIG. 7 illustrates an example of an artificial neural network structure for machine learning. FIG. 7 illustrates an example of an artificial neural network in the machine learning

model. For the artificial neural network, any one of a variety of types may be utilized. The artificial neural network in FIG. 7 is a fully connected network. A hidden node is used in stages. In addition, in order to avoid over-fitting of learning data, batch normalization may be applied for each layer. The batch normalization is a process of causing activation values to be distributed, in a forcible and appropriate manner, over the results obtained by passing through a layer.

As described in the machine learning model, the artificial neural network is also trained beforehand using data on a plurality of samples. That is, an artificial neural network model is generated using, as input values, expression levels of target genes for a plurality of samples and Oncotype DX RS of the target genes (samples).

It is now possible to predict breast cancer prognosis for a given subject by inputting a subject's data into an artificial neural network. The values input into the artificial neural network, which has been trained beforehand, are expression levels of the subject's target genes. The output value from the artificial neural network may be a result obtained by predicting breast cancer prognosis based on Oncotype DX RS. The output value from the artificial neural network may be output as a quantitative result such as a specific score. Alternatively, the output value from the artificial neural network may be a direct qualitative evaluation result such as high risk or low risk.

Although description has been made by focusing on the artificial neural network in FIG. 7, it is possible to estimate breast cancer prognosis using various other machine learning models. For example, in the machine learning model, various methods such as genetic algorithm, support vector machine, and Bayesian network may be utilized.

The computer device predicts breast cancer prognosis for the currently input sample (subject), based on a value resulting from the machine learning model (150). For example, in the computer device, the value resulting from the machine learning model may be a specific score. Here, the computer device may determine that the current sample (subject) is at high risk when the output value from the artificial neural network is equal to or higher than a specific value in a case of being compared with a reference value provided beforehand. On the contrary, the computer device may determine that the current sample (subject) is at low risk when the output value from the artificial neural network is lower than the reference value.

FIG. 8 illustrates an example of a machine learning-based system for prediction of breast cancer prognosis using next-generation sequencing. FIG. 8(a) illustrates an example of a system 200 implemented in a network. The system 200 for prediction of breast cancer prognosis includes a client device 210 and an analysis server 220. Furthermore, the system 200 for prediction of breast cancer prognosis may include a model DB 230. The analysis server 220 corresponds to the above-described computer device.

The client device 210 is a device that provides data for a subject. The client device 210 transmits, to the analysis server 220, data (genetic data) on the subject's gene expression. The data transmitted by the client device 210 may be data representing gene sequences.

The analysis server 220 measures an expression level of a target gene using RNA sequencing data obtained by performing next-generation sequencing (NGS)-based sequencing on the genetic data, and estimates breast cancer prognosis for the subject based on the output value that is output in a case where the expression level of the target gene is input into a machine learning model adapted beforehand. The analysis server 220 may also perform the above-described post-data processing and normalization process. The analysis server 220 uses a machine learning model adapted beforehand. The machine learning model may be stored in a separate model DB 230. The process, in which the analysis server 220 processes data and predicts breast cancer prognosis, is as described above.

FIG. 8(b) illustrates an example of a computer device 300 for predicting breast cancer prognosis. The computer device 300 illustrated in FIG. 8(b) may be the above-described analysis server 220. The computer device 300 refers to a device such as personal computer (PC), laptop, smart device, or server. The computer device 300 includes an input device 310, a calculation device 320, a storage device 330, and an output device 340.

The input device 310 receives the subject's genetic data. The genetic data refers to data associated with expression of the target gene or a gene sequence thereof. The input device 310 is a device that inputs the subject's data into the computer device 300 via communication or a separate storage device. Furthermore, the input device 310 may be an interface device (keyboard, mouse, touch screen, or the like) through which the subject's clinical data is directly input into the computer device 300.

The storage device 330 is a device that stores the above-described machine learning

model. The storage device 330 may store the subject's genetic data transmitted from the input device 310. The storage device 330 may be a device in which software for predicting breast cancer prognosis through analysis of the genetic data is installed.

The calculation device 320 measures gene expression levels for the input genetic data, inputs the gene expression levels into the machine learning model, and estimates breast cancer prognosis for the subject based on the output value from the machine learning model.

The output device 340 is a device that outputs information on prognosis in a certain form. The output device 340 includes at least one of a display device, an output device for outputting a document, and a communication device for transmitting the information on prognosis to another device.

In addition, the above-described machine learning-based method for prediction of breast cancer prognosis using next-generation sequencing may be implemented as a program (or application) including an executable algorithm that can be executed on a computer. The program may be provided in a form stored in a non-transitory computer readable medium.

The non-transitory readable medium refers to a medium that stores data semi-permanently and is readable by an apparatus, rather than a medium that stores data temporarily such as register, cache, and memory. Specifically, various applications or programs as described above may be provided in a form stored in a non-transitory readable medium such as CD, DVD, hard disk, Blu-ray disk, USB memory stick, memory card, and ROM.

It should be noted that the embodiments and the accompanying drawings are merely intended to clearly describe a part of the technical spirit included in the present invention, and it is obvious that all variations and specific embodiments readily deducible by those skilled in the art within the technical spirit included in the specification and drawings are included in the scope of the present invention.

CLAIMS

1. A method for prediction of breast cancer prognosis, comprising:
 - allowing a computer device to use RNA sequencing data of a subject's tissue, to measure expression levels of target genes;
 - wherein the computer device has, as input values, expression levels of target genes for a plurality of samples and has output values by inputting the expression levels of the target genes to an artificial neural network depending on Oncotype DX recurrence scores for the plurality of samples,
 - wherein the target genes include all of the following 149 genes:

ACTB, APOBEC3B, ASF1B, ASPM, AURKA, AURKB, BAG1, BCL2, BIRC5, BLM, BUB1, BUB1B, C14orf45, C16orf61, C7orf63, CACNA1D, CCNA2, CCNB1, CCNB2, CCNE1, CCNE2, CCT5, CD68, CDC20, CDC25A, CDC45, CDC6, CDCA3, CDCA8, CDK1, CDKN3, CENPA, CENPE, CENPF, CENPM, CENPN, CEP55, CHEK1, CIRBP, CKS2, CRIM1, CTSL2, CX3CR1, CYBRD1, DBF4, DDX39, DLGAP5, DNMT3B, DONSON, DTL, E2F1, E2F8, ECHDC2, ERBB2, ERCC6L, ESPL1, ESR1, EXO1, EZH2, FAM64A, FANCI, FBXO5, FEN1, FOXM1, GAPDH, GINS1, GRB7, GSTM1, GTSE1, GUSB, HJURP, HMMR, HN1, IFT46, KIF11, KIF13B, KIF14, KIF15, KIF18A, KIF18B, KIF20A, KIF23, KIF2C, KIF4A, KIFC1, KPNA2, LMNB1, LMNB2, LRIG1, LRRC48, LRRC59, MAD2L1, MARCH8, MCM10, MCM2, MCM6, MELK, MKI67, MLF1IP, MMP11, MYBL2, NCAPG, NCAPG2, NCAPH, NDC80, NEK2, NUP93, NUSAP1, OIP5, PBK, PDSS1, PGR, PKMYT1, PLK1, PLK4, PRC1, PTTG1, RACGAP1, RAD51, RAD51AP1, RAI2, RFC4, RPLP0, RRM2, SCUBE2, SETBP, SF3B3, SHCBP1, SHMT2, SLC25A1, SLC7A5, SPAG5, SPC25, SQLE, STARD13, STIL, STMN1, SYNC, TACC3, TFRC, TK1, TOP2A, TPX2, TRIP13, TROAP, TTK, UBE2C, UBE2S, and ZWINT,
 - wherein the subject is estrogen receptor positive.

2. The method according to claim 1,
 - wherein the computer device measures the expression levels of the target genes with a next-generation sequencing (NGS) technique.

3. The method according to claim 1,
wherein the target genes include 16 genes used for calculating the Oncotype DX recurrence scores and additional genes of which a correlation with respect to the Oncotype DX recurrence scores is equal to or higher than a reference value, and
the genes are selected from the genes with a correlation coefficient equal to or higher than the reference value by performing correlation analysis with the Oncotype DX recurrence scores for each of the sample genes, which have no lymph node metastasis.
4. The method according to claim 1,
wherein the RNA sequencing data includes a conserved exon region that is commonly expressed in the plurality of sample data.
5. The method according to claim 1,
wherein the computer device normalizes the expression levels of the target genes using Trimmed Mean of M-value (TMM) technique.
6. The method according to claim 1,
wherein the computer device estimates the prognosis that the subject is at high risk in a case where the output value is equal to or higher than a reference value, and estimates the prognosis that the subject is at low risk in a case where the output value is lower than the reference value.
7. A computer-readable recording medium having recorded therein a program for executing, on a computer, the method for prediction of breast cancer prognosis according to any one of claims 1 to 6.
8. A system for prediction of breast cancer prognosis, comprising:
a client device for storing RNA gene data of a subject's tissue; and
an analysis server in which RNA sequencing data, which is obtained by performing next-generation sequencing (NGS)-based sequencing on the gene data, is used to measure expression levels of target genes, and breast cancer prognosis for the

subject is estimated; and

wherein the analysis server has, as input values, expression levels of target genes for a plurality of samples and has output values by inputting the expression levels of the target genes to an artificial neural network depending on Oncotype DX recurrence scores for the plurality of samples,

wherein the target genes include all of the following 149 genes:

ACTB, APOBEC3B, ASF1B, ASPM, AURKA, AURKB, BAG1, BCL2, BIRC5, BLM, BUB1, BUB1B, C14orf45, C16orf61, C7orf63, CACNA1D, CCNA2, CCNB1, CCNB2, CCNE1, CCNE2, CCT5, CD68, CDC20, CDC25A, CDC45, CDC6, CDCA3, CDCA8, CDK1, CDKN3, CENPA, CENPE, CENPF, CENPM, CENPN, CEP55, CHEK1, CIRBP, CKS2, CRIM1, CTSL2, CX3CR1, CYBRD1, DBF4, DDX39, DLGAP5, DNMT3B, DONSON, DTL, E2F1, E2F8, ECHDC2, ERBB2, ERCC6L, ESPL1, ESR1, EXO1, EZH2, FAM64A, FANCI, FBXO5, FEN1, FOXM1, GAPDH, GINS1, GRB7, GSTM1, GTSE1, GUSB, HJURP, HMMR, HN1, IFT46, KIF11, KIF13B, KIF14, KIF15, KIF18A, KIF18B, KIF20A, KIF23, KIF2C, KIF4A, KIFC1, KPNA2, LMNB1, LMNB2, LRIG1, LRRC48, LRRC59, MAD2L1, MARCH8, MCM10, MCM2, MCM6, MELK, MKI67, MLF1IP, MMP11, MYBL2, NCAPG, NCAPG2, NCAPH, NDC80, NEK2, NUP93, NUSAP1, OIP5, PBK, PDSS1, PGR, PKMYT1, PLK1, PLK4, PRC1, PTTG1, RACGAP1, RAD51, RAD51AP1, RAI2, RFC4, RPLP0, RRM2, SCUBE2, SETBP, SF3B3, SHCBP1, SHMT2, SLC25A1, SLC7A5, SPAG5, SPC25, SQLE, STARD13, STIL, STMN1, SYNC, TACC3, TFRC, TK1, TOP2A, TPX2, TRIP13, TROAP, TTK, UBE2C, UBE2S, and ZWINT,

wherein the subject is estrogen receptor positive.

9. The system according to claim 8,

wherein the target genes include genes that have a correlation coefficient equal to or higher than a reference value in a case where correlation analysis with the Oncotype DX recurrence scores is performed for each of the sample genes which have no lymph node metastasis.

10. The system according to claim 8,

wherein the RNA sequencing data includes a conserved exon region that is

commonly expressed in the plurality of sample data.

11. The system according to claim 8,
wherein the analysis server normalizes the expression levels of the target genes
using Trimmed Mean of M-value (TMM) technique.