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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 September 2015 (24.09.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/139652 A1

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

(21) International Application Number:
PCT/CN2015/074606

(22) International Filing Date:
19 March 2015 (19.03.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/968,140	20 March 2014 (20.03.2014)	US
61/990,389	8 May 2014 (08.05.2014)	US

(71) Applicant: PHARMACOGENETICS LIMITED [CN/CN]; Unit C9, 5/F, Wah Lok Industrial Centre, Phase II, 31-35 Shan Mei Street, Shatin, N.T., Hong Kong (CN).

(72) Inventors: **XUE, Hong**; Flat D, 57/F, Tower 3, Residence Oasis, 15 Pui Shing Road, Tseung Kwan O, N. T., Hong Kong (CN). **DING, Xiaofan**; Flat B, 4/F Wing Ming House, 35-37 Yuk Wah Crescent, Tsz Wan Shan, Kowloon, Hong Kong (CN). **TSANG, Shui-Ying**; Flat E, 12/F South Hillcrest, Tuen Mun, N.T., Hong Kong (CN).

(74) Agent: **WAN HUI DA INTELLECTUAL PROPERTY AGENCY**; Yiyuan Office Building, Friendship Hotel, No. 1 Zhongguancun Street South, Haidian District, Beijing 100873 (CN).

(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, BN, BR, BW, BY, BZ, CA, CH, CL, 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, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ,

[Continued on next page]

(54) Title: USE OF RECURRENT COPY NUMBER VARIATIONS IN CONSTITUTIONAL HUMAN GENOME FOR PREDICTION OF PREDISPOSITION TO CANCER

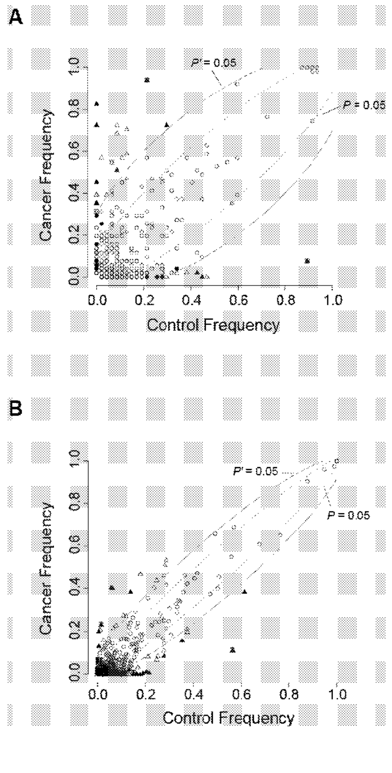

Figure 4

(57) **Abstract**: In the present application, prediction on the predisposition of a human test subject to cancer is made based on machine learning-assisted comparison of the copy number variations ("CNV") found in the constitutional DNA of the test subject with a set of diagnostic recurrent CNV features (viz. markers) selected from a collection of constitutional DNA samples from noncancer subjects (designated as "Noncancer DNA" samples) plus constitutional DNA samples from cancer patients (designated as "Cancer DNA" samples), all from the same ethnic group as the test subject. Selection and testing of the set of diagnostic recurrent CNV features is performed using a machine learning procedure, exemplified by the CFS-based method, the Frequency-based method and the Classifier-based method, together with the Naive Bayes classification method. Prediction of the test subject's predisposition to cancer is also performed with the Naive Bayes classification method. The cancer patients from whom the constitutional "Cancer DNA" samples are prepared, for the purpose of selection of the diagnostic recurrent CNV features, can consist of patients inflicted with one type of cancer or more than one types of cancers.

TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

USE OF RECURRENT COPY NUMBER VARIATIONS IN CONSTITUTIONAL HUMAN GENOME FOR PREDICTION OF PREDISPOSITION TO CANCER

BACKGROUND

[0001] The present invention relates to a method of using recurrent copy number variations ("CNV") in the constitutional, viz. germline, genome of a human subject to predict the subject's predisposition to cancer. This method identifies the recurrent constitutional CNVs in a collection of DNA samples comprising both the DNA of noncancerous tissues of individuals without experience of cancer (referred to as "Noncancer DNA" samples) and the DNA of noncancerous tissues of cancer patients (referred to as "Cancer DNA" samples"), and selects from this collection using machine learning procedures a set of diagnostic recurrent CNV features comprising some of the CNVs that are enriched in individuals without experience of cancer relative to cancer patients, along with some of the CNVs that are enriched in cancer patients relative to individuals without experience of cancer, all of the same ethnic group. The usefulness of such a set of diagnostic recurrent CNV features as classifier between known "Noncancer DNA" samples and "Cancer DNA" samples is tested. Upon confirmation of usefulness, the CNVs found in the constitutional DNA of any test subject from the same ethnic group as the sources of the "Noncancer DNA" and "Cancer DNA" can be analyzed to determine the presence or absence of the various CNVs contained in the set of diagnostic recurrent CNV features, and thereby arrive at a prediction of the level of predisposition of the test subject to cancer.

[0002] The CNVs present in the DNA of the constitutional genome in noncancerous tissues of any noncancer individual, cancer patient or test subject can be determined from single nucleotide polymorphism (SNP) microarrays of human genomic DNA, qPCR, whole-genome sequencing of the person's genome, or from DNA sequencing of a subset of sequences amplified from the genome exemplified by an "AluScan" sequence subset containing inter-Alu and/or Alu-proximal genomic sequences that have been amplified by polymerase chain reaction (“PCR”) employing PCR primers the sequences of which are based on the consensus sequences of Alu-insertion elements in the human genome. The CNVs that are found in any collection of DNA samples can be identified as “recurrent” CNVs or “rare” CNVs based on their frequencies and statistical criteria. Hitherto although various “rare” CNVs have been correlated with different specific types of cancer, no correlation between recurrent constitutional CNV and cancer has been obtained and employed as a basis for the prediction of predisposition to cancer.

[0003] In the present method, the prediction of the predisposition to cancer of test subjects requires a set of diagnostic recurrent CNV features selected from the recurrent CNVs that are present in a collection of “Noncancer DNA” samples and “Cancer DNA” samples from the constitutional genomes in the noncancerous tissues of individuals without experience of cancer and cancer patients respectively. For this purpose, machine learning-assisted selection is performed using statistical selection methods exemplified by, and not limited to, the following: (I) Correlation-based Feature Selection (CSF) Method; this can be used to generate CFS-based CNV-features that are highly correlated with the recurrent CNVs in either the “Noncancer DNA” class or the “Cancer DNA” class yet uncorrelated with one another, for example using CfsSubsetEval from the Weka machine learning package together with the BestFirst method (Hall MA and Smith LA, Feature subset selection: A correlation based filter approach. *International Conference on Neural Information Processing and Intelligent Information Systems*. New Zealand; 1997: 8555-858; Dagliyan O et al., Optimization based tumor classification from microarray gene expression data. *PLoS One* 2011, 6:e14579); (II) Frequency-based Method; in this method, a CNV-feature is selected by virtue of its frequency in the “Noncancer DNA” samples being significantly different from its frequency in the “Cancer DNA” samples; and (III) Classifier-based Method; in this method, CNV-features are selected by use of a classifier, for example the ClassifierSubsetEval attribute evaluator in the Weka machine learning package together with the BestFirst method (Hall MA, et al.: The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 2009; 11: 10-18.)

[0004] The usefulness of a diagnostic set of recurrent CNV features as a classification tool to classify DNA samples between the “Noncancer DNA” and “Cancer DNA” classes can be assessed by machine learning implementation of the Naïve Bayes classification method, and receiver operating characteristic (ROC) analysis which was originally introduced to distinguish between meaningful radar signals and noise, and has since found important application in diverse fields of clinical medicine (Zweig MH and Campbell G: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 1993, 39:561-577; Zhou X *Statistical Methods in Diagnostic Medicine.* New York, USA: Wiley & Sons; 2002).

[0005] Once a set of diagnostic recurrent CNV features selected from the recurrent CNVs found in a collection of “Noncancer DNA” and “Cancer DNA” samples from an ethnic population is found to yield an ROC-AUC (ROC-area under the curve) greater than 0.5, and therefore useful as a classification tool for classifying DNA samples between the “Noncancer DNA” and “Cancer DNA” classes, it can be employed to predict the predisposition to cancer of the constitutional DNA samples from test subjects belonging to the same ethnic population.

[0006] The principle of the prediction method referred to in [0005] consists of the assembly of a Learning Band of labeled DNA samples (viz. wherein the identities of the DNA samples are known to belong to either the “Noncancer DNA” or the “Cancer DNA” class), selection of a set of diagnostic recurrent CNV-features from all the DNA samples in the Learning Band, and confirming that the set of diagnostic recurrent CNV-features selected is useful as a classifier tool for classifying unlabeled DNA samples (viz. wherein it is not known which DNA samples belong to the “Noncancer DNA” class and which to the “Cancer DNA” class) into the “Noncancer DNA” and “Cancer DNA” classes. Once usefulness is confirmed, the CNVs occurring in each constituent DNA sample in the Learning Band are examined to determine the presence or absence of the different CNVs of the set of diagnostic recurrent CNV features in that constituent sample. The results obtained enable the estimation of the B-value for that constituent sample on the basis of Eqn.

1, and the relative B-values of all the labeled constituent samples in the Learning Band can be ranked on a B-value scale:

$$\mathrm{Pr(cancer|features)} = \sum_{i=1}^{n} \frac{\mathrm{Pr\ (feature_i|cancer)} \times \mathrm{Pr(cancer)}}{\mathrm{Pr(feature_i)}}$$

$$\mathrm{Pr(noncancer|features)} = \sum_{i=1}^{n} \frac{\mathrm{Pr\ (feature_i|noncancer)} \times \mathrm{Pr(noncancer)}}{\mathrm{Pr(feature_i)}}$$

$$B = \log\left(\frac{\mathrm{Pr(cancer|features)}}{\mathrm{Pr(noncancer|features)}}\right)$$
$$= \log\left(\frac{(\sum_{i=1}^{n} \mathrm{Pr\ (feature_i|cancer)}) \times \mathrm{Pr(cancer)}}{(\sum_{i=1}^{n} \mathrm{Pr\ (feature_i|noncancer)}) \times \mathrm{Pr(noncancer)}}\right)$$

Eqn 1

in which B is the log of the ratio between Pr(cancer|feature) viz. the Bayesian posterior probability of membership in the Cancer class given the CNV data of the constituent sample, and Pr(noncancerl|feature) viz. the Bayesian posterior probability of membership in the Noncancer class given the CNV data of the constituent sample; Pr(features|cancer) is the likelihood function of the CNV data given membership in the Cancer class; Pr(features|noncancer) is the likelihood function of the CNV data given membership in the Noncancer class; Pr(cancer) and Pr(Noncancer) are the prior distributions of Cancer and Noncancer samples respectively in the Learning Band. The expected classification for any test sample is 'Cancer' if B >0, 'Noncancer' if B <0, or indeterminate if B = 0. Accordingly, when the different samples in the Learning Band are ranked according to their B-values, the "Noncancer DNA" samples will tend to have low rankings, whereas the "Cancer DNA" samples will tend to have high rankings, on the B-value scale.

The B-value scale constructed from all the labeled Learning Band samples provides a standard B-value scale for DNA samples for the ethnic population from which the "Noncancer DNA" samples and "Cancer DNA" samples are derived. Having this standard B-value scale, the CNVs detected in the constitutional DNA of any test subject from the same ethnic population can be analyzed to determine the presence or absence of various CNV features contained in the set of diagnostic recurrent CNV features employed to construct the B-value scale, and thereupon a B-value for the test subject on the basis of Eqn. 1. By comparing the B-value of the test subject to the B-values for various constituent "Noncancer DNA" and "Cancer DNA" samples in the Learning Band, the subject's predisposition to cancer will be revealed as high (i.e. if the subject's B-value is high on the B-value scale), intermediate (i.e. if the subject's B-value is intermediate-positioned on the B-value scale), or low (i.e. if the subject's B-value is low on the B-value scale).

SUMMARY

[0007] The present invention relates to a method using the copy number variations ("CNV") in the constitutional genome of a human subject to predict the subject's predisposition to cancer. This method identifies the recurrent constitutional CNVs in a collection of DNA samples comprising both the DNA of noncancerous tissues of individuals without cancer or previous experience of cancer (referred to as "Noncancer DNA" samples) and the DNA of noncancerous tissues of cancer patients (referred to as "Cancer DNA" samples"), and selects from this collection by means of machine learning procedures a set of diagnostic recurrent CNV features comprising some of the recurrent CNVs that are enriched in individuals without any experience of cancer relative to cancer patients, along with some of the CNVs that are enriched in cancer patients relative to individuals without any experience of cancer, all from the same ethnic group. The usefulness of such a set of diagnostic recurrent CNV features as classifier between "Noncancer DNA" samples and the "Cancer DNA" samples is tested. Upon confirmation of usefulness, the CNVs found in the constitutional DNA of any test subject from the same ethnic group as the sources of the “Noncancer DNA:” and “Cancer DNA” samples can be analyzed to determine the presence or absence of the various CNVs contained in the set of diagnostic recurrent CNV features, and thereby arrive at a prediction of the level of predisposition of the test subject to cancer.

[0008] The selection of a set of diagnostic recurrent CNV features comprising recurrent CNVs referred to in [0007] is performed employing machine learning methods exemplified by, but not limited to, the following methods: (I) Correlation-based Feature Selection (CSF) Method; (II) Frequency-based Method; and (III) Classifier-based Method. The usefulness of the set of diagnostic recurrent CNV features selected is tested by employing the set of features as classification tool to classify known “Noncancer DNA” and “Cancer DNA” samples into the “Noncancer DNA” and “Cancer DNA” classes using the Naïve Bayes classification method, and evaluating the accuracy of the classification achieved by means of . receiver-operating characteristic (ROC) analysis.

[0009] Once a set of diagnostic recurrent CNV features is found to be useful, yielding an ROC-AUC (ROC area under the curve) value greater than 0.5, the set of features can be employed to predict the predisposition to cancer of any test subject from the same ethnic population as the sources of the “Noncancer DNA” and “Cancer DNA” samples that give rise to the set of diagnostic recurrent CNV features on the basis of Bayesian posterior probability analysis.

[0010] Because the CNV features in a set of diagnostic recurrent constitutional CNV features are typically distributed with different frequencies among the “Cancer DNA” samples from patients bearing different types of cancer, the present invention can be employed not only to identify test subjects with enhanced predisposition to cancer in general, but also subjects with enhanced predispositions to specific types of cancer.

BRIEF DESCRIPTION OF DRAWINGS

[0011] The following drawings form part of the present specification and are included to further demonstrate certain aspects of the present invention. The invention may be better understood by reference to one or more of these drawings in combination with the detailed description of specific embodiments presented herein.

[0012] Figure 1 shows recurrent CNVs identified from noncancerous white blood cell DNAs, using Affymetrix SNP6.0 arrays, of (A) a Caucasian cohort of Noncancer subjects and Cancer patients; and (B) a Korean cohort of Noncancer subjects and Cancer patients. In these examples, only significantly recurrent regions with lengths greater than 1kb and less than 10Mb, and with a q-value <0.25 were included in the analysis. Upper panel of the figure shows q values of copy number gains ("CNV-gains"), and lower panel shows q values of copy number losses ("CNV-losses"). The q values were generated by GISTIC2.0 such that a high "-log q-value" indicated a highly non-random event. The CNV-gains (marked as A-series) and CNV-losses (marked as D-series) selected for inclusion in the CFS-based diagnostic CNV-features for the Caucasian and Korean cohorts are shown in Figure 2 and Figure 3 respectively.

[0013] Figure 2 shows a set of CFS-based diagnostic recurrent CNV-features selected from the noncancerous white blood cell DNAs of a Caucasian cohort of Noncancer and Cancer subjects analyzed by Affymetrix SNP6.0 array. "Cancer Freq" indicates frequency of the CNV-feature among "Cancer DNA" samples, "Control Freq" indicates frequency of the CNV-feature among control "Noncancer DNA" samples, and "Can/Con ratio" refers to their ratios. CNVG = CNV-gain; CNVL = CNV-loss. The A-series and D-series ID numbers are added to facilitate location of the various CNV features in Figure 1(A).

[0014] Figure 3 shows a set of CFS-based diagnostic recurrent CNV-features selected from the noncancerous white blood cell DNAs of a Korean cohort of Noncancer and Cancer subjects analyzed by Affymetrix SNP6.0 array. "Cancer Freq" indicates frequency of the CNV-feature among "Cancer DNA" samples, "Control Freq" indicates frequency of the CNV-feature among control "Noncancer DNA" samples, and "Can/Con ratio" refers to their ratios. CNVG = CNV-gain;

CNVL = CNV-loss. The A-series and D-series ID numbers are added to facilitate location of the various CNV features in Figure 1(B).

[0015] Figure 4 shows the frequencies of occurrence of recurrent CNV-features selected by the CFS-, Frequency- and Classifier-based methods among the cancer patients and noncancer controls of (A) Caucasian cohort and (B) Korean cohort. Solid triangle, CNV-feature selected by both CFS and Frequency methods; solid circle, ones selected only by CFS method; open triangle, selected only by Frequency method; solid triangle plus solid inverted triangle, selected by CFS method, Frequency method and Classifier method; open triangle plus open inverted triangle, selected by Frequency method and Classifier method; open circle, not selected by any of the three methods. Chi-square based probability *P* of Cancer and Control frequencies being equal is >0.05 between the two dashed lines representing *P* = 0.05, and <0.05 outside these two dashed lines. The two solid lines representing *P*'= 0.05, where *P'* stands for *P* value after Bonferroni correction, likewise separate the in-between region of P'>0.05 and the outer regions of P'<0.05.

[0016] Figure 5 shows a table of ROC-AUC values for Caucasian and Korean samples attained with the sets of recurrent CNV-features obtained using three different CNV feature-selection methods.

[0017] Figure 6 shows the prediction accuracies of cancer occurrence in (A) Caucasian cohort, and (B) Korean cohort, using CFS-based CNV-features. For each of the cohorts, the DNA samples were randomly separated into a Learning Band and a Test Band containing the same or approximately the same number of Noncancer DNA samples, as well as the same or approximately the same number of Cancer DNA samples. CFS-based CNV-features were selected from the Learning Band, and employed to predict the classification of each sample in the Test Band into the Noncancer and Cancer classes based on the value of *B* in Eqn. 1 as given in [0006]. The classification would be 'Cancer' if *B* >0, 'Noncancer' if *B* <0, or indeterminate if *B* = 0. By repeating the random separation of samples into Learning Band and Test Band 1,000 times, and each time making predictions on every sample in the Test Band and estimating the Accuracy of prediction using Eqn. 2, 1,000 Accuracy estimates were obtained from the 1,000 runs:

$$\text{Accuracy} = \frac{[\text{True prediction of non-cancer}] + [\text{True prediction of cancer}]}{[\text{Total prediction of non-cancer}] + [\text{Total prediction of cancer}]} \text{x}100\% \quad \text{Eqn. 2}$$

The distributions of the 1,000 Accuracy estimates obtained for the Caucasian and Korean cohorts together with the Average accuracy in each case for the 1,000 runs, are indicated on graphs (A) and (B) respectively.

[0018] Figure 7 shows the distribution of CFS-based diagnostic recurrent CNV-features in the non-tumor white blood cell DNA of (A) Caucasian cancer patients, where the CFS-based diagnostic recurrent CNV-features are those described in Figure 2; and (B) Korean cancer patients, bearing different types of cancers, where the CFS-based diagnostic recurrent CNV-features are those described in Figure 3. In each instance, K-means clustering was employed to cluster the different types of cancer-patient DNAs according to their contents of CFS-based CNV-features using the kmean package in R (Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 2006, **22**:1540-1542). Since the number of CFS-based CNV-features was greater than two, the CLUSPLOT function in the cluster package in R (Pison G, et al. Dispalying a clustering with CLUSPLOT. *Comput Stat Data An* 1999, 30:381-392) was used to reduce the dimensions of the data by principal component analysis (PCA) to produce the graphical output in terms of only the first two principal components. Different types of cancer patients included patients of colorectal cancers (green circles), gliomas (blue triangles), myelomas (red squares), gastric cancers (blue squares) and hepatocellular carcinomas (HCC, red triangles).

[0019] Figure 8 shows a Table of CFS-based recurrent CNV-features selected from the noncancerous white blood cell DNAs of a Chinese cohort of noncancer controls and cancer patients analyzed by AluScan sequencing. “Cancer Freq” indicates frequency of the CNV-feature among “Cancer DNA” samples, “Control Freq” indicates frequency of the CNV-feature among control “Noncancer DNA” samples, and “Can/Con ratio” refers to their ratios. CNVG = CNV-gain; CNVL = CNV-loss.

[0020] Figure 9 shows the frequencies of occurrence of the recurrent CNV-features selected by CFS-based method among the noncancer controls and cancer patients of a Chinese cohort. The selected recurrent CNV-features, as indicated in Figure 8, are represented by solid triangles. The unselected recurrent CNVs are represented by open circles.

[0021] Figure 10 shows the prediction accuracies of cancer occurrence in the Chinese cohort determined through random separation of the Noncancer and Cancer DNA samples into a Learning Band a Test Band; thereupon the CFS-based method was used to select recurrent CNV-features from the Learning Band for predicting the classification of each sample in the Test Band into the Noncancer and Cancer classes, as described in Figure 6. The distributions of the Accuracy estimates obtained from 100 rounds of this procedure of randomized Learning-Test Band separation, selection of diagnostic recurrent CNV features from the Learning Band, and making prediction of cancer predisposition on the samples in the Test Band, together with the Average accuracy for the 100 runs, are indicated on the graph.

[0022] Figure 11 shows a summary of the procedure in the present invention for predicting predisposition to cancer. N represents constitutional DNA samples from the noncancerous tissues of Noncancer subjects, and C represents constitutional DNA samples from the noncancerous tissues of Cancer patients.

DETAILED DESCRIPTION

[0023] It will be readily apparent to one skilled in the art that various substitutions and modifications may be made in the invention disclosed herein without departing from the scope and spirit of the invention.

Terms:

[0024] The term "a" or "an" as used herein in the specification may mean one or more. As used herein in the claim(s) the words "a" or "an" may mean one or more than one. As used herein “another” may mean at least a second or more.

[0025] The term “copy number variation”, or CNV, refers to variation from the standard human genome where the DNAs in the autosomal chromosomes, and in the X chromosome in females, are present in two copies (viz. “diploidal”), such that any DNA segment present in more than or less than two copies represents a CNV. The standard DNAs in the X and Y chromosomes in males are present in a single copy (viz. “haploidal”), such that any DNA segment present in more or less than one copy represents a CNV. Any CNV containing more than the standard number of copies constitutes a CNV-gain, and any CNV containing less than the standard number of copies constitutes a CNV-loss.

[0026] The term “recurrent CNV” refers to CNVs that are not too rare in occurrence, so that they can provide a useful basis for prediction purpose. Methods for identifying recurrent CNVs may be obtained from standard reviews such as Rueda, O.M. & Diaz-Uriarte, R. Finding Recurrent Regions of Copy Number Variation, Collection of Biostatistics Research Archive 2008, Paper 42, The Berkeley Electronic Press, which lists the MSA, GISTIC, RAE, MAR, CMAR, cghMCR, CGHregions, Master HMMs, STAC, Interval Scores, CoCoA, KC SMART, SIRAC, GEAR and Markers methods and their associated softwares.

[0027] The term “diagnostic recurrent CNV features” in the present invention refers to constitutional recurrent CNVs selected from the recurrent CNVs identified from a collection of genomic DNAs of both the noncancerous tissue samples of Noncancer (viz. noncancer individuals) subjects and the noncancerous tissue samples of Cancer (viz. cancer patients) subjects belonging to the same ethnic group. These CNV features are typically enriched in Noncancer DNAs relative to Cancer DNAs, or enriched in Cancer DNAs relative to Noncancer DNAs, such that a prediction regarding the extent of predisposition toward cancer of any test subject of the same ethnic population can be made based on the presence or absence of the various constituent diagnostic recurrent CNV features in the test subject’s constitutional DNA. Selection of CNV features can be conducted using various statistical methods including but not limited to the following methods: (I) Correlation-based Feature Selection (CSF) Method, (II) Frequency-based Method, and (III) Classifier-based Method. Each of the methods gives rise to a set of diagnostic recurrent CNV features, and the utility of any set of diagnostic recurrent CNV features can be tested by employing it to classify individual samples in a sample collection comprising both labeled Noncancer DNA samples and labeled Cancer DNA samples using a probabilistic classifier such as Fisher’s linear discriminant, Logistic regression, Naïve Bayes classifier, decision trees, neural networks etc. Once a set of diagnostic recurrent CNV features is found to be diagnostically useful, i.e. yielding an ROC-AUC value in excess of 0.5, it can be employed as the basis for predicting the extent of predisposition to cancer of test genomes belonging to the same ethnic population as the Noncancer and Cancer DNA samples that generated the particular set of CNV features.

[0028] In one embodiment of the present invention, single nucleotide polymorphism (SNP) array data on whole blood samples from 51 Caucasian cancer patients and 47 ethnically-matched noncancer controls obtained using the high resolution Affymetrix SNP6.0 array platform were retrieved from the Gene Expression Omnibus (GEO) [http://www.ncbi.nlm.nih.gov/geo/] database. The program apt-copynumber-workflow with default settings from Affymetrix Power Tools (http://www.affymetrix.com/partners_programs/programs/developer /tools/powertools.affx) was employed to generate CNV callings for these Cancer and Noncancer samples using a reference template generated from the averaged microarray data for 270 HapMap samples acquired using the Affymetrix SNP6.0 platform and processed with apt-copynumber-workflow. Segmentation of neighboring copy number variations into CNV-gain segments and CNV-loss segments was performed based on the copy number values using Circular Binary Segmentation (CBS) with default parameters in DNACopy in R program (Olshen AB et al. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 2004, 5:557-572). The genomic coordinates employed in the present study referred to human reference genome version hg19/GRCh37, and the annotation file used with the SNP6.0 platform was release version 32. To identify significantly recurrent CNVs, the GISTIC2.0 method (Mermel C.H. et al, Genome Biol. 12(4):R41, 2011) was employed with the options "-smallmem 1–broad 1 –brlen 0.5–conf 0.9–ta 0.2–td 0.2–twosides 1–genegistic 1". CNVs with a log2 ratio change of either > 0.2 or < -0.2 are regarded as recurrent CNVs (Ding, X. et al. Application of machine learning to development of copy number variation-based prediction of cancer risk. *Genomics Insights* 2014:7, 1-10). The recurrent CNVs identified are shown in Figure 1(A).

[0029] In this embodiment of the present invention, each of the Correlation-based Feature Selection (CSF) Method, Frequency-based Method, and Classifier-based Method was employed to generate three sets of diagnostic recurrent CNV features from the Caucasian Cancer and Noncancer DNA microarray data described in [0028]. To assess the capability of each of these three sets of diagnostic recurrent CNV features as a basis for classifying samples between the Cancer and Noncancer classes, the Naïve Bayes classification method from the Weka package was employed to generate a training model incorporating one of the CNV-feature sets, which was tested with 1,000 iterations of twofold cross validation. To test the robustness of the model, 10,000 permutated datasets were generated by randomly shuffling the group labels ('Noncancer' vs. 'Cancer') for each sample within the original dataset, and the whole classification process was repeated for each permutated dataset. The significance of the original classification was calculated based on the distribution of correct prediction percentage from the 10,000 permutations. The results of Naïve Bayes classification obtained using the three training models incorporating the three different CNV-feature sets to make decisions on sample classification into the 'Noncancer' and 'Cancer' classes are shown in Figure 5, where the CNV-feature sets based on the CFS method, Frequency-based method and Classifier-based method yielded ROC-AUV values of 0.996±0.001, 0.991±0.007, and 0.986±0.014 respectively for the Caucasian samples. These high ROC-AUC values showed that all three CNV-feature ensembles are capable of classifying samples into the "Noncancer DNA" and "Cancer DNA" classes with a high level of accuracy. Each of these CNV-feature sets therefore provide a useful basis for predicting the predisposition of Caucasian test subjects to cancer. The basis for the usefulness of the sets of selected CNV-features as classifiers for the Caucasian samples is demonstrated in Figure 4(A). The CNV features selected all displayed a highly biased distribution, occurring either frequently in the Cancer DNA samples but infrequently in the control Noncancer DNA samples, or frequently in the control Nnoncancer DNA samples but infrequently in the Cancer DNA samples. As a result, they are endowed with the ability to serve as markers for Cancer DNA, or as markers for control Noncancer DNA.

[0030] To confirm the expectation that CNV-feature sets can provide a valid basis for predicting predisposition to cancer, the Noncancer control DNA samples (N) in the Caucasian cohort were randomly divided in a trial run into two groupings that were equal in number when there were an even number of samples; or, when there were an odd number of samples, an extra sample was randomly allocated to one of the two groupings so that they differed in size by only a single sample. One of the groupings was randomly assigned to the Learning Band, and the other grouping to the Test Band. Similarly, for the cancer patients (C), the DNA samples from the colorectal cancer patients were randomly divided into two groupings that were either equal in size or different by only one sample; again one grouping was randomly assigned to the Learning Band, and the other to the Test Band. The glioma patient samples and the myeloma patient samples were treated the same way to finally yield an [N + C] Learning Band and an [N + C] Test Band containing an equal or near-equal number of N and C samples. Thereupon a set of CFS-based CNV-features were derived from the CNVs included in the Learning Band. Applying this set of learnt CFS-based CNV-features to each and every individual sample in the Test Band using Eqn. 1 yielded either a ‘true’ or ‘not true’ allocation of the individual into the Noncancer or Cancer class; altogether the predictions pertaining to all the individuals in the Test Band would yield an Accuracy estimate for this trial run based on Eqn. 2:

$$\text{Accuracy} = \frac{[\text{True prediction of non-cancer}]+[\text{True prediction of cancer}]}{[\text{Total prediction of non-cancer}]+[\text{Total prediction of cancer}]}\text{x}100\% \quad \text{Eqn. 2}$$

By repeating this random partition of the sample into Learning Band and Test Band 1,000 times, 1,000 estimates of accuracy were obtained. The distribution of these 1,000 accuracy estimates is shown in Figure 6(A), and their Average value of 93.6% confirmed the valid use of diagnostic recurrent CNV features to predict the predisposition of Caucasian test subjects to cancer..

[0031] In another embodiment of the present invention, single nucleotide polymorphism array data on whole blood samples from 347 Korean cancer patients and 195 ethnically-matched Noncancer controls obtained using the high resolution Affymetrix SNP6.0 platform were retrieved from the Gene Expression Omnibus (GEO) [http://www.ncbi.nlm.nih.gov/geo/] and caArray databases [https://array.nci.nih.gov/caarray/]. Using the same procedures as those described in [0028] and [0029], recurrent CNVs comprising both CNV-gains and CNV-losses were called from the Noncancer and Cancer samples, and the Correlation-based Feature Selection (CSF) Method, Frequency-based Method, and Classifier-based Method were employed to generate three different CNV feature sets from the Noncancer and Cancer and DNA array data. The Naïve Bayes classification method was employed to generate three training model incorporating the three different CNV-feature sets, making decisions in each case on sample classification into the “Noncancer DNA” or “Cancer DNA” classes. As shown in Figure 5, the CNV-feature sets using the CFS method, Frequency-based method and Classifier-based method yielded ROC-AUV values of 0.975±0.002, 0.958±0.009, and 0.867±0.016 respectively for the Korean samples. These high

ROC-AUC values showed that all three CNV-feature ensembles are capable of classifying samples into the Noncancer and Cancer classes with a high level of accuracy, and therefore provide a useful basis for predicting the predisposition of Korean test subjects to cancer. The basis for the usefulness of the sets of selected CNV-features as classifiers for the Korean samples is demonstrated in Figure 4(B). The CNV features selected all displayed a highly biased distribution, occurring either frequently in the Cancer DNA samples but infrequently in the control Noncancer DNA samples, or frequently in the control Nnoncancer DNA samples but infrequently in the Cancer DNA samples. As a result, they are endowed with the ability to serve as markers for Cancer DNA, or as markers for control Noncancer DNA.

[0032] In addition, when the various Noncancer control subjects and cancer subjects in the Korean cohort were randomly partitioned into a Learning Band and a Test Band 1,000 times as described in [0030] for the Caucasian cohort, followed by estimation of the accuracy of predictions made each time on samples in the Test Band using recurrent CNV features selected from the Learning Band by means of the CSF-based method, the distribution of the 1,000 accuracy estimates is shown in Figure 6(B), and their Average value of 86.5% confirmed the valid use of recurrent CNV features to predict the predisposition of Korean test subjects to cancer.

[0033] The Caucasian cancer patient samples described in [0028] came from patients inflicted variously with three types of cancers: glioma, myeloma and colorectal cancer. Figure 7A shows that the CNV-feature contents in the three types of cancer-patient constituent genomes were dissimilar. It follows that, when carrying out the selection of diagnostic recurrent CNV features, one can employ DNAs from the noncancerous tissues of noncancer subjects, together with DNAs from the noncancerous tissues of cancer patients inflicted with one (or a restricted number of) cancer type instead of multiple cancer types, in order to focus prediction on cancer predisposition to that one (or a restricted number of) cancer type instead of predisposition to cancer in general. Likewise, the Korean cancer patient samples described in [0031] also came from patients inflicted variously with three types of cancer: gastric cancer, hepatocellular carcinoma (HCC) and colorectal cancer. As shown in Figure 7B, the CNV-feature contents in the three types of cancer-patient constituent genomes were also dissimilar. Therefore, again one can employ DNA samples from the noncancerous tissues of noncancer subjects, together with DNA samples from the noncancerous tissues of patients inflicted with one (or a restricted number of) cancer type instead of multiple cancer types for selection of diagnostic recurrent CNV features, in order to focus prediction on cancer predisposition to that one (or a restricted number of) cancer type instead of predisposition to cancer in general. These examples indicate that the use of diagnostic recurrent CNV features to predict predisposition to cancer applies to either predisposition to cancer in general, or predisposition to one (or a restricted number of) type of cancer in particular.

[0034] In the preceding embodiments of the present invention, recurrent CNVs comprising both CNV-gains and CNV-losses were called from human genomic data from the high resolution Affymetrix SNP6.0 platform. In another embodiment of the present invention, recurrent CNVs comprising both CNV-gains and CNV-losses were called from genomic data on a cohort of 28 Chinese cancer patients inflicted with 14 liver cancers, 4 gastric cancers, 3 lung cancers, 4 gliomas and 3 leukemias, and 22 ethnically-matched noncancer controls analyzed using the AluScan next generation sequencing platform (Mei L, Ding X, Tsang SY, Pun FW, Ng SK, Yang J, Zhao C, Li D, Wan W, Yu CH *et al*: AluScan: a method for genome-wide scanning of sequence and structure variations in the human genome. *BMC genomics* 2011, **12**:564). From the recurrent CNVs called from the AluScan sequence data using the AluScanCNV algortithm with 350 kb windows (Yang, J.F. et al. Copy number variation analysis based on AluScan sequences. *J Clin Bioinformatics* **4**, 15, 2014), a set of diagnostic recurrent CNV features were selected by means of the CFS-based method (Figure 8).

[0035] As shown in Figure 9, the recurrent CNVs called from the 28 Cancer DNA samples and the 22 Noncancer DNA samples in the Chinese cohort were found to occur in various Cancer and Noncancer DNA samples with a wide spectrum of frequencies (open circles in Figure 9). In contrast, the set of diagnostic recurrent CNV features selected by the CFS-based method from all the recurrent CNVs, as shown in Figure 8, displayed strongly biased frequencies that were either enriched in the Cancer DNA samples relative to the Noncancer DNA samples, or enriched in the Noncancer DNA samples relative to the Cancer DNA samples (solid triangles in Figure 9). When this selected CNV-feature set was employed to classify the 28 Cancer DNA samples and 22 Noncancer DNA samples in the Chinese cohort into the "Cancer" and "Noncancer" classes based on Eqn. 1, the ROC-AUC value obtained was 0.993± 0.001, showing that the selected CNV-feature set is capable of classifying samples into the "Cancer" and "Noncancer" classes with a high level of accuracy. This CNV-feature set therefore provides a useful basis for predicting the predisposition of Chinese test subjects to cancer. The basis for the usefulness of the sets of selected CNV-features as classifiers for the Chinese samples is demonstrated in Figure 9. The CNV features selected all displayed a highly biased distribution, occurring either frequently in the Cancer DNA samples but infrequently in the control Noncancer DNA samples, or frequently in the control Noncancer DNA samples but infrequently in the Cancer DNA samples. As a result, they are endowed with the ability to serve as markers for Cancer DNA, or as markers for control Noncancer DNA.

[0036] When the 28 Cancer and 22 Noncancer samples in the Chinese cohort were randomly separated into a Learning Band and a Test Band for 100 times by means of the procedure described in [0030] for the Caucasian cohort, followed by estimation of the accuracy of predictions made each time on samples in the Test Band using diagnostic recurrent CNV features selected from the Learning Band using the CSF-based method, the distribution of the 100 accuracy estimates is shown in Figure 10, and their Average value of 83.7% confirmed the valid use of diagnostic recurrent CNV features to predict the predisposition of Chinese test subjects to cancer.

What is claimed is:

1. A method of using the copy number variations ("CNV") in the constitutional (i.e. germline) genomic DNA of a human subject for predicting the predisposition of the subject to cancer, based on a comparison between the CNVs in his/her DNA with a set of diagnostic recurrent CNV features (or markers) that have been selected from the recurrent copy number variations found in a collection of constitutional DNA samples from the noncancerous tissues of noncancer subjects plus constitutional DNA samples from the noncancerous tissues of cancer patients, and comprising the steps of:

(a) Identify the recurrent copy number variations (CNV) in a collection of constitutional DNA samples from the noncancerous tissues of subjects without experience of cancer (designated as "Noncancer DNA" samples) plus constitutional DNA samples from the noncancerous tissues of cancer patients (designated as "Cancer DNA" samples), all from the same ethnic group.

(b) Select, from the recurrent CNVs in a collection of "Noncancer DNA" samples plus "Cancer DNA" samples, one or more sets of recurrent CNV features (or, markers) with the capability of serving as a classifier tool to classify DNA samples between the "Noncancer DNA" and "Cancer DNA" classes.

(c) Testing the capability of the selected set or sets of recurrent CNV features for their capability of serving as a classifier tool to classify DNA samples between the "Noncancer DNA" and "Cancer DNA" classes. Once a set of recurrent CNV features is found to be useful as a classifier tool to classify DNA samples between the "Noncancer DNA" and "Cancer DNA" classes, it can be regarded and employed as a diagnostic set of recurrent CNV features.

(d) Analyze the CNVs found in the constitutional genomic DNA in the noncancerous tissues of a test subjects belonging to the same ethnic group as the sources of the

"Noncancer DNA" samples and "Cancer DNA" samples from which a diagnostic set of recurrent CNV features is derived, in order to determine the presence or absence of each and every recurrent CNV contained in the diagnostic set of recurrent CNV features. Based on the data regarding the presence or absence of the different recurrent CNVs contained in the diagnostic set of recurrent CNV features, prediction on the level of the predisposition of the test subject to cancer can then be made.

2. The method of claim 1, wherein CNVs are identified from genomic DNA based on the use of high resolution Affymetrix SNP array.

3. The method of claim 1, wherein CNVs are identified from genomic DNA based on whole genome DNA sequencing.

4. The method of claim 3, wherein the whole genome sequencing is performed with a next generation sequencing method.

5. The method of claim 1, wherein CNVs are identified from next generation sequencing of a subset of genomic DNA sequences.

6. The method of claim 5, wherein the subset of genomic DNA sequences is obtained with the use of an AluScan sequencing platform.

7. The method of claim 1, where recurrent CNVs are identified based on statistical procedures exemplified by, and not limited to, the GISTIC2.0 algorithm.

8. The method of claim 1, where recurrent CNVs are identified based on statistical procedures exemplified by, and not limited to, the AluScan algorithm.

9. The method of claim 6, wherein recurrent CNVs are identified based on the AluScanCNV algorithm.

10. The method of claim 1 wherein a set of recurrent CNV features is selected from the recurrent CNVs identified in the a collection of DNA comprising both “Noncancer DNA” samples and “Cancer DNA” samples by use of a Correlation-based feature selection (CFS) method, where features are selected by virtue of their being highly correlated with either the “Noncancer DNA” class or the “Cancer DNA” class but not with one another.

11. The method of claim 1 wherein a set of recurrent CNV features is selected from the recurrent CNVs identified in the a collection of DNA comprising both “Noncancer DNA” samples and “Cancer DNA” samples by use of a Frequency-based method, where a recurrent CNV feature is selected by virtue of its frequency in “Noncancer DNA” samples being significantly different from its frequency in “Cancer DNA” samples.

12. The method of claim 1 wherein a set of recurrent CNV features is selected from the recurrent CNVs identified in the a collection of DNA comprising both “Noncancer DNA” samples and “Cancer DNA” samples by use of a Classiifier-based method, where recurrent CNV features are selected by use a classifier, for example the ClassifierSubsetEval attribute evaluator from the Weka machine learning package together with the BestFirst search method.

13. The method of claim 1 wherein testing the usefulness of a set of diagnostic recurrent CNV features is performed with the use of Bayesian posterior probability analysis.

14. The method of claim 1 wherein estimation of the predisposition of a test subject to canc3er is performed with the use of Bayesian posterior probability analysis.

15. The method of claim 1 wherein the “Cancer DNA” samples employed consist of the constitutional genomic DNAs of patients inflicted with more than one types of cancer.

16. The method of claim 1 wherein the “Cancer DNA” samples employed consist of the constitutional genomic DNAs of patients inflicted with a single type of cancer.

17. The method of claim 1 wherein the following recurrent CNVs are found to be individually useful as members of a set of diagnostic recurrent CNV features for predisposition to cancer testing for Caucasian test subjects (CNVG = CNV-gain; CNVL = CNV-loss):

GENOMIC REGION	TYPE
chr 1: 17082580-17093244	CNVG
chr 1: 196790519-196801642	CNVG
chr 2: 91774012-91778756	CNVG
chr 3: 155483565-155492176	CNVG
chr 3: 178883723-178885918	CNVG
chr 7: 76303499-76309667	CNVG
chr 8: 1360723-1362790	CNVG
chr 9: 686583-694566	CNVG
chr 9: 68713481-68753608	CNVG
chr 10: 46918173-46989538	CNVG
chr 11: 1961189-2022483	CNVG
chr 12: 34467864-34523670	CNVG
chr 13: 19319636-19400859	CNVG
chr 19: 41365625-41375784	CNVG
chr 21: 11123429-11126187	CNVG
chr 21: 48069120-48129895	CNVG
chr 22: 16102481-16395149	CNVG
chr 22: 22447034-22453683	CNVG
chr 1: 152768559-152776742	CNVL
chr 3: 195422280-195429688	CNVL
chr 11: 4967240-4970264	CNVL
chr 11: 73581673-73590246	CNVL

18. The method of claim 1 wherein the following recurrent CNVs are found to be individually useful as members of a set of diagnostic recurrent CNV features for predisposition to cancer testing for Korean test subjects (CNVG = CNV-gain; CNVL = CNV-loss):

GENOMIC REGION	TYPE
chr1:144008324-144013581	CNVG
chr2:132366274-132452986	CNVG
chr6:161032508-161068029	CNVG
chr7:76303499-76308210	CNVG
chr7:97405580-97420636	CNVG
chr7:110175088-110177523	CNVG
chr8:140566271-140583019	CNVG
chr9:16911092-16913776	CNVG
chr11:58833238-58835701	CNVG
chr11:69329675-69351720	CNVG
chr14:101515428-101529413	CNVG
chr14:106980636-107003597	CNVG
chr15:20180946-20186638	CNVG
chr17:12894795-12900382	CNVG
chr18:2262552-2263726	CNVG
chr19:40783234-40786732	CNVG
chr21:11123429-11126187	CNVG
chr1:179078208-179203917	CNVL
chr1:196741305-196770682	CNVL
chr2:219313355-219433596	CNVL
chr5:788049-863796	CNVL
chr5:125932873-125966005	CNVL
chr5:180329360-180380190	CNVL
chr6:74221700-74234042	CNVL
chr6:150042816-150075171	CNVL
chr7:38297824-38319338	CNVL
chr11:7813449-7829919	CNVL
chr16:11912686-11927917	CNVL
chr19:15983972-16013337	CNVL
chr19:53603953-53641568	CNVL

19. The method of claim 1 wherein the following recurrent CNVs are found to be individually useful as members of a set of diagnostic recurrent CNV features for predisposition to cancer testing for Chinese test subjects test subjects (CNVG = CNV-gain; CNVL = CNV-loss):

GENOMIC REGION	Type
chr2:38150001-38500000	CNVG
chr5:167300001-167650000	CNVG
chr6:170800001-171115067	CNVG
chr12:106050001-106400000	CNVG
chr14:101850001-102200000	CNVG
chr15:92050001-92400000	CNVG
chr19:29400001-29750000	CNVG
chr1:117950001-118300000	CNVL
chr1:175000001-175350000	CNVL
chr1:71400001-71750000	CNVL
chr3:64400001-64750000	CNVL
chr5:167300001-167650000	CNVL
chr5:168000001-168350000	CNVL
chr6:5250001-5600000	CNVL
chr6:85400001-85750000	CNVL
chr7:80850001-81200000	CNVL
chr10:64400001-64750000	CNVL
chr15:92050001-92400000	CNVL
chr17:34300001-34650000	CNVL
chr18:73500001-73850000	CNVL

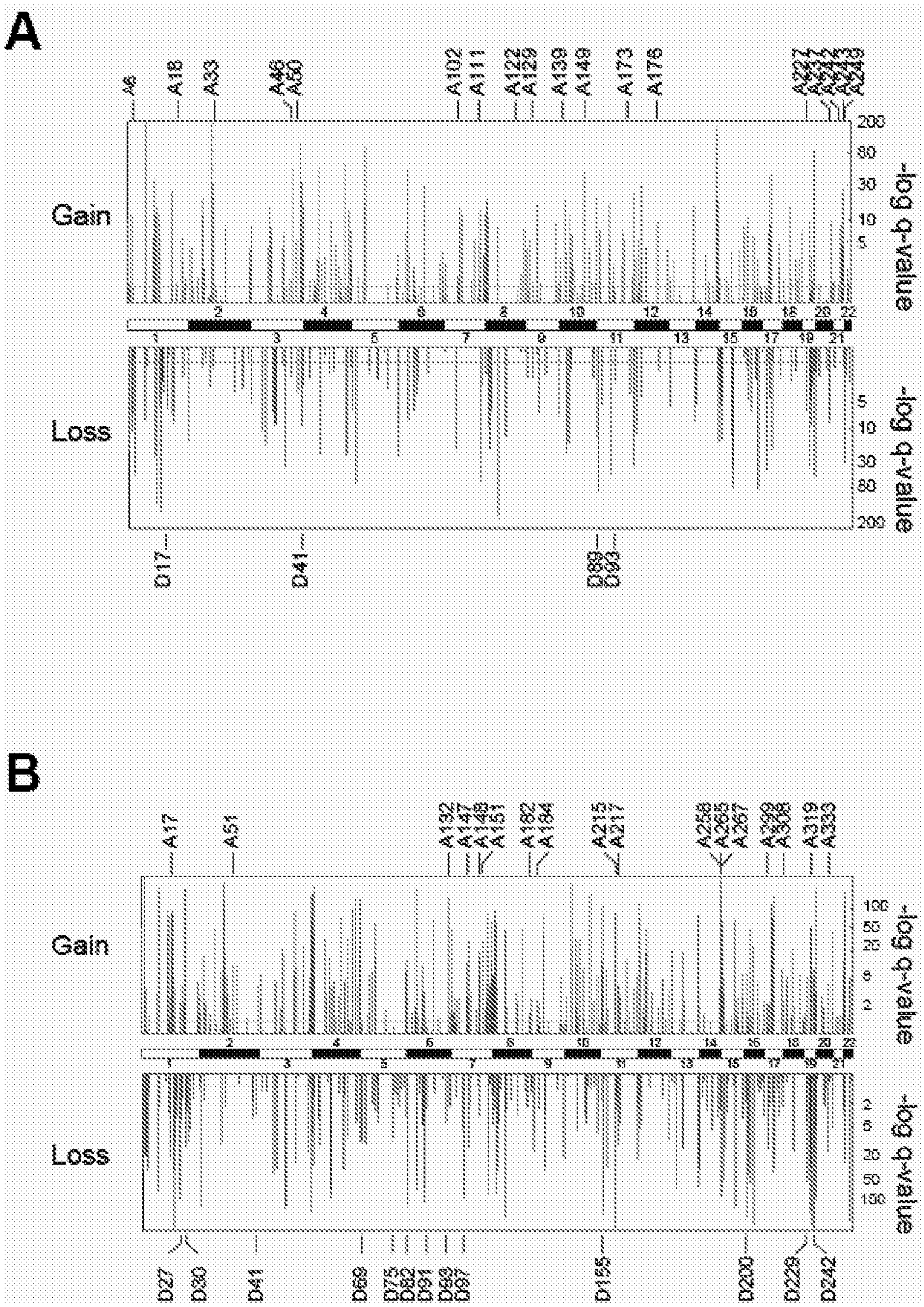
A
Gain
Loss
-log q-value
200
80
30
10
5
A6
A18
A33
A46
A50
A102
A111
A122
A129
A139
A149
A173
A176
A227
A237
A238
A242
A243
A249
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
D17
D41
D89
D93
B
100
50
20
6
2
A17
A51
A132
A147
A148
A151
A182
A184
A215
A217
A258
A265
A267
A299
A308
A319
A333
23
D27
D30
D41
D69
D75
D82
D91
D89
D97
D155
D200
D229
D242

Figure 1

CNV ID	GENOMIC REGION	TYPE	CANCER FREQ.	CONTROL FREQ	CAN/CON RATIO
A6	chr 1: 17082580-17093244	CNVG	0.51	0.09	5.67
A18	chr 1: 196790519-196801642	CNVG	0.73	0.30	2.43
A33	chr 2: 91774012-91778756	CNVG	0.94	0.21	4.48
A46	chr 3: 155483565-155492176	CNVG	0.06	0	NA
A50	chr 3: 178883723-178885918	CNVG	0	0.21	0
A102	chr 7: 76303499-76309667	CNVG	0.02	0.43	0.05
A111	chr 8: 1360723-1362790	CNVG	0	0.28	0
A122	chr 9: 686583-694566	CNVG	0	0.26	0
A129	chr 9: 68713481-68753608	CNVG	0.73	0	NA
A139	chr 10: 46918173-46989538	CNVG	0.25	0.02	12.5
A149	chr 11: 1961189-2022483	CNVG	0.16	0	NA
A173	chr 12: 34467864-34523670	CNVG	0.08	0.89	0.09
A176	chr 13: 19319636-19400859	CNVG	0.35	0	NA
A227	chr 19: 41365625-41375784	CNVG	0.04	0	NA
A237	chr 21: 11123429-11126187	CNVG	0.82	0	NA
A242	chr 21: 48069120-48129895	CNVG	0.45	0	NA
A243	chr 22: 16102481-16395149	CNVG	0.29	0	NA
A249	chr 22: 22447034-22453683	CNVG	0	0.45	0
D17	chr 1: 152768559-152776742	CNVL	0.04	0.34	0.12
D41	chr 3: 195422280-195429688	CNVL	0.16	0	NA
D89	chr 11: 4967240-4970264	CNVL	0.08	0	NA
D93	chr 11: 73581673-73590246	CNVL	0	0.26	0

Figure 2

CNV ID	GENOMIC REGION	TYPE	CANCER FREQ.	CONTROL FREQ	CAN/CON RATIO
A17	chr1:144008324-144013581	CNVG	0.23	0.02	11.5
A51	chr2:132366274-132452986	CNVG	0.2	0.01	20
A132	chr6:161032508-161068029	CNVG	0.16	0.35	0.46
A147	chr7:76303499-76308210	CNVG	0	0.05	0
A148	chr7:97405580-97420636	CNVG	0.01	0.07	0.14
A151	chr7:110175088-110177523	CNVG	0.01	0.11	0.09
A182	chr8:140566271-140583019	CNVG	0.01	0.21	0.05
A184	chr9:16911092-16913776	CNVG	0.02	0	NA
A215	chr11:58833238-58835701	CNVG	0.08	0.28	0.29
A217	chr11:69329675-69351720	CNVG	0.03	0	NA
A258	chr14:101515428-101529413	CNVG	0.01	0.09	0.11
A265	chr14:106980636-107003597	CNVG	0.38	0.62	0.61
A267	chr15:20180946-20186638	CNVG	0.4	0.06	6.67
A299	chr17:12894795-12900382	CNVG	0	0.04	0
A308	chr18:2262552-2263726	CNVG	0	0.05	0
A319	chr19:40783234-40786732	CNVG	0.13	0.01	13.00
A333	chr21:11123429-11126187	CNVG	0.4	0.06	6.67
D27	chr1:179078208-179203917	CNVL	0.02	0.13	0.15
D30	chr1:196741305-196770682	CNVL	0.02	0	NA
D41	chr2:219313355-219433596	CNVL	0	0.19	0
D69	chr5:788049-863796	CNVL	0.02	0	NA
D75	chr5:125932873-125966005	CNVL	0.01	0.22	0.05
D82	chr5:180329360-180380190	CNVL	0	0.02	0
D91	chr6:74221700-74234042	CNVL	0	0.18	0
D93	chr6:150042816-150075171	CNVL	0	0.16	0
D97	chr7:38297824-38319338	CNVL	0.11	0.56	0.20
D155	chr11:7813449-7829919	CNVL	0.01	0	NA
D200	chr16:11912686-11927917	CNVL	0	0.16	0
D229	chr19:15983972-16013337	CNVL	0.02	0	NA
D242	chr19:53603953-53641568	CNVL	0.01	0	NA

Figure 3

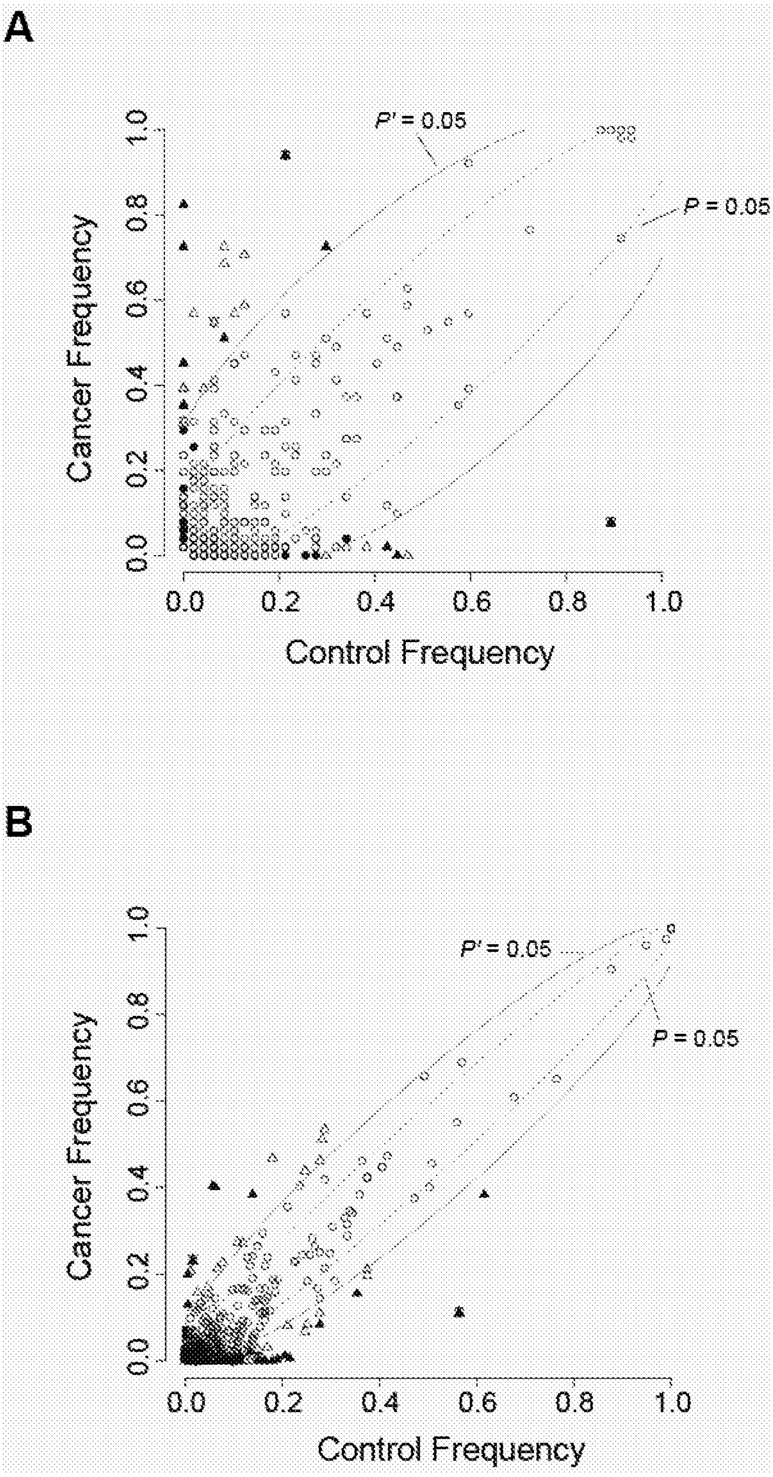
A
P′ = 0.05
P = 0.05
Cancer Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Control Frequency
0.0
0.2
0.4
0.6
0.8
1.0
B
P′ = 0.05
P = 0.05
Cancer Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Control Frequency
0.0
0.2
0.4
0.6
0.8
1.0

Figure 4

Basis of CNV-features	**Caucasian**	**Korean**
	n = 98	**n = 542**
CFS	0.996 ± 0.001	0.975 ± 0.002
Frequency	0.991 ± 0.007	0.958 ± 0.009
Classifier	0.986 ± 0.014	0.867 ± 0.016

Figure 5

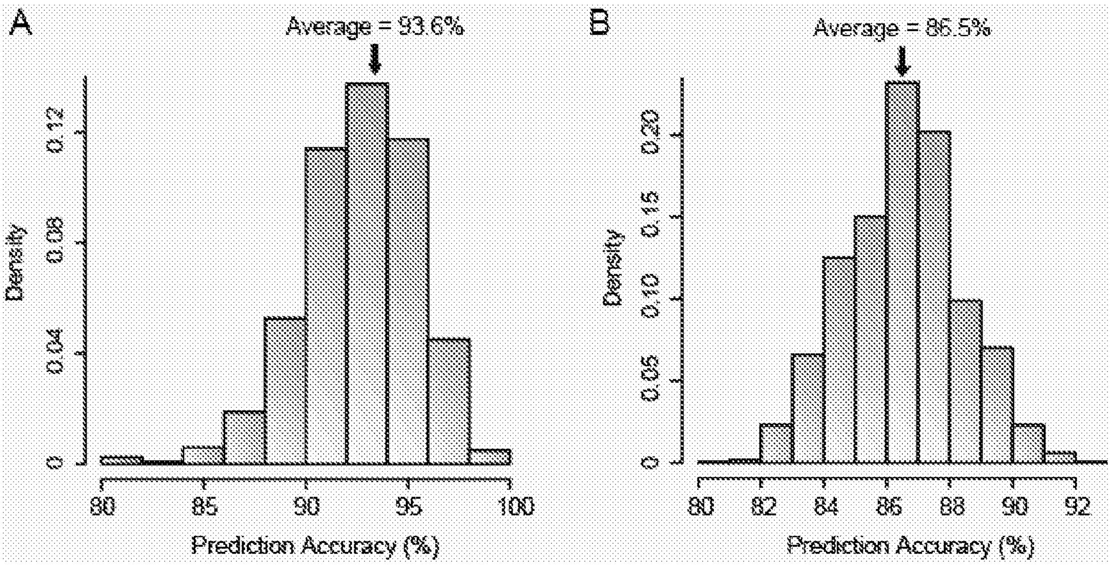
A
Average = 93.6%
Density
0
0.04
0.08
0.12
80
85
90
95
100
Prediction Accuracy (%)
B
Average = 86.5%
Density
0
0.05
0.10
0.15
0.20
80
82
84
86
88
90
92
Prediction Accuracy (%)

Figure 6

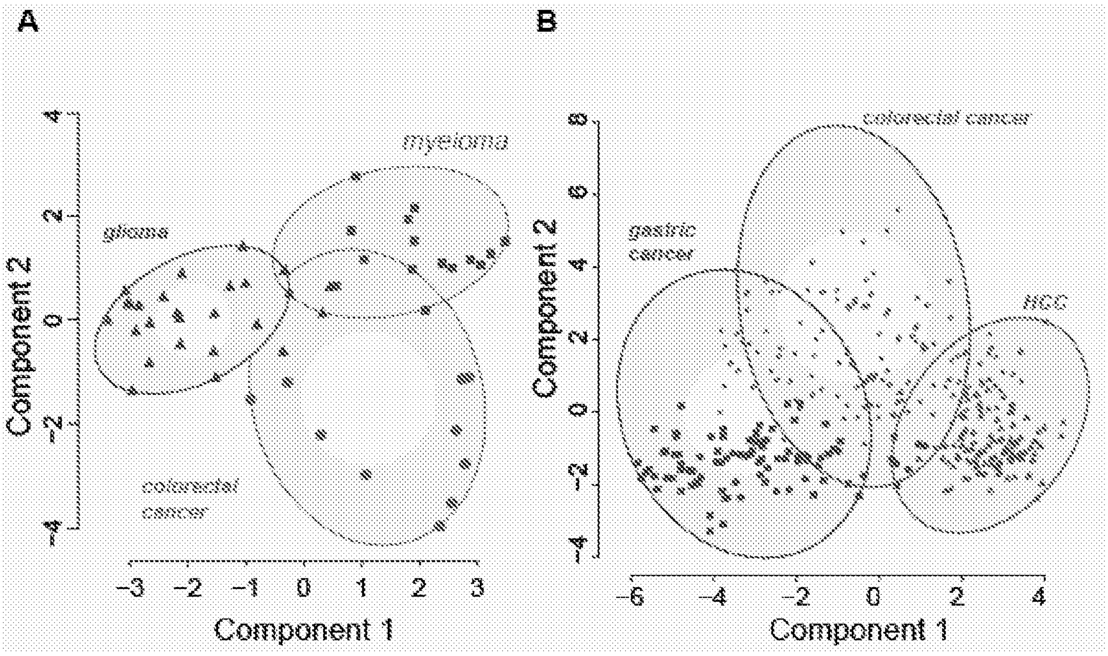
A
B
myeloma
glioma
colorectal cancer
Component 2
Component 1
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
colorectal cancer
gastric cancer
HCC
8
6
4
2
0
−2
−4
−6
−4
−2
0
2
4

Figure 7

GENOMIC REGION	type	CANCER FREQ.	CONTROL FREQ	CAN/CON RATIO
chr2:38150001-38500000	CNVG	0.30	0	NA
chr5:167300001-167650000	CNVG	0.36	0	NA
chr6:170800001-171115067	CNVG	0.36	0	NA
chr12:106050001-106400000	CNVG	0.42	0	NA
chr14:101850001-102200000	CNVG	0.55	0	NA
chr15:92050001-92400000	CNVG	0.33	0	NA
chr19:29400001-29750000	CNVG	0.42	0	NA
chr1:117950001-118300000	CNVL	0.09	0.57	0.16
chr1:175000001-175350000	CNVL	0	0.39	0
chr1:71400001-71750000	CNVL	0	0.39	0
chr3:64400001-64750000	CNVL	0.09	0.52	0.17
chr5:167300001-167650000	CNVL	0.18	0.74	0.25
chr5:168000001-168350000	CNVL	0.06	0.70	0.09
chr6:5250001-5600000	CNVL	0	0.39	0
chr6:85400001-85750000	CNVL	0	0.39	0
chr7:80850001-81200000	CNVL	0	0.44	0
chr10:64400001-64750000	CNVL	0	0.39	0
chr15:92050001-92400000	CNVL	0.06	0.65	0.09
chr17:34300001-34650000	CNVL	0.42	0	NA
chr18:73500001-73850000	CNVL	0.03	0.61	0.05

Figure 8

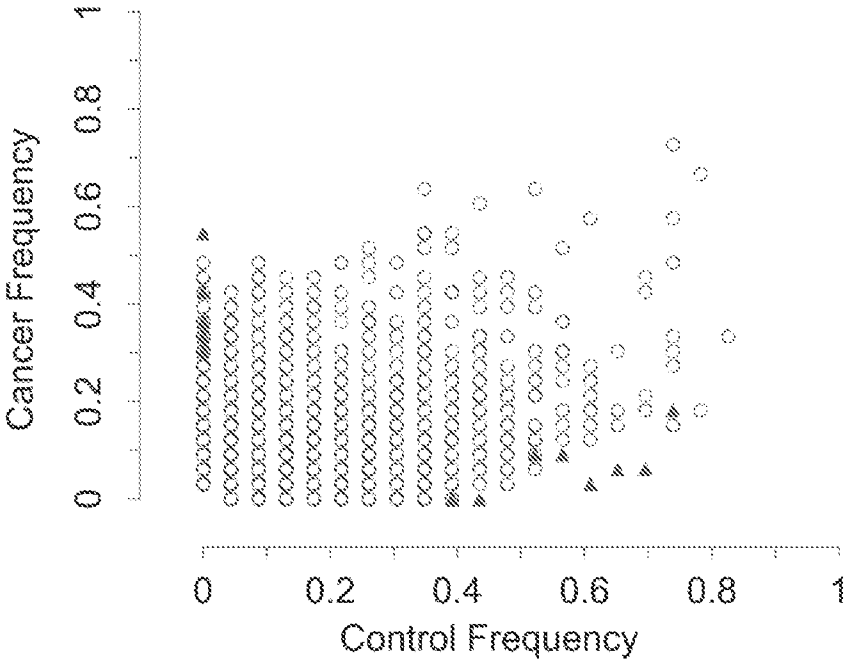
Cancer Frequency
0
0.2
0.4
0.6
0.8
1
Control Frequency
0
0.2
0.4
0.6
0.8
1

Figure 9

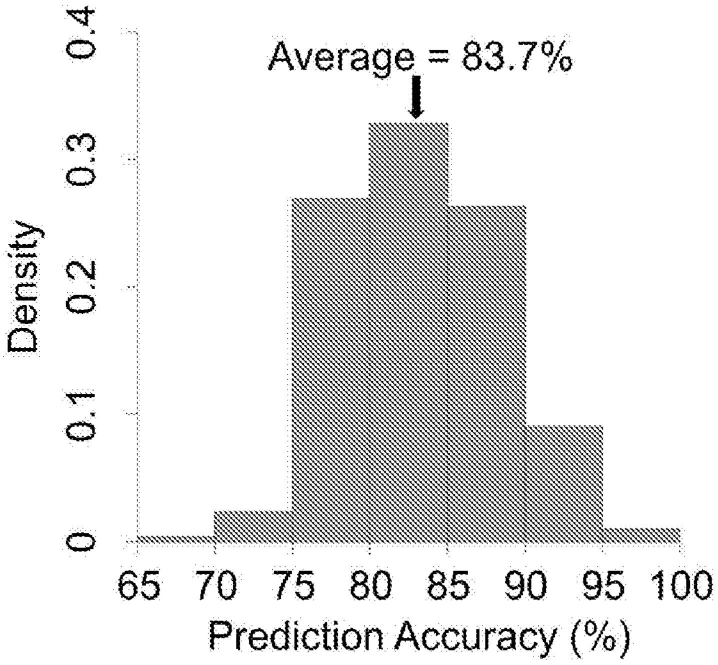
Average = 83.7%
Density
0
0.1
0.2
0.3
0.4
65
70
75
80
85
90
95
100
Prediction Accuracy (%)

Figure 10

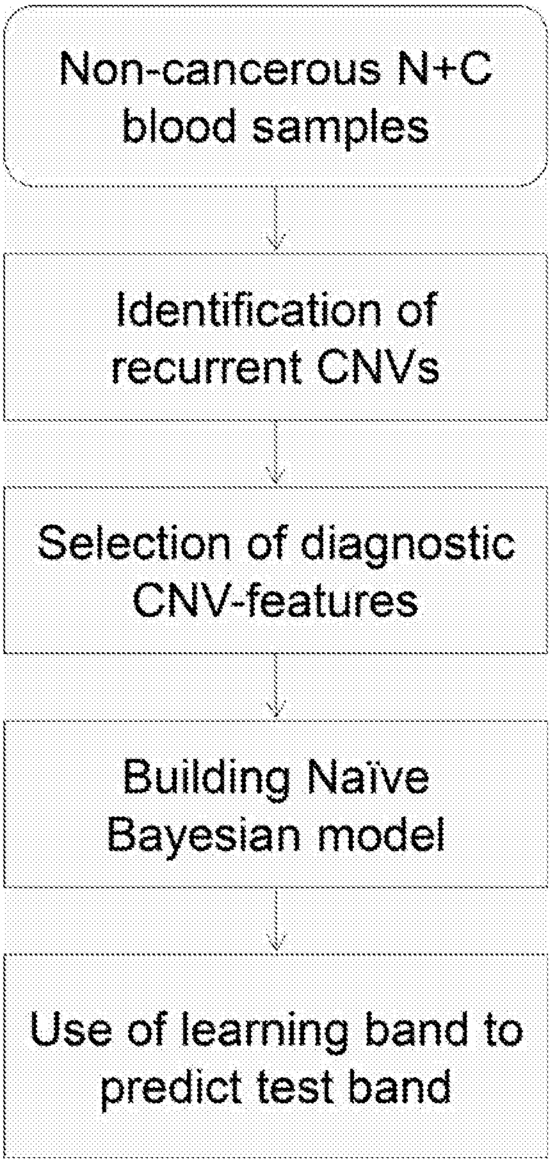
Non-cancerous N+C blood samples
Identification of recurrent CNVs
Selection of diagnostic CNV-features
Building Naïve Bayesian model
Use of learning band to predict test band

Figure 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/074606

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C12Q1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI; EPODOC; CNPAT; CNKI; PubMed; ISI; copy number variations, CNV, recurrent, common, cancer, carcinoma, neoplasm, tumor, tumour, predisposition, susceptibility, predict, machine learning, Correlation-based Feature Selection, CSF, CFS, Frequency based, Classifier based, CfsSubsetEval, BestFirst, ClassifierSubsetEval, AluScan

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	DING， X. et al. "Application of Machine Learning to Development of Copy Number Variation-based Prediction of Cancer Risk." *Genomics Insights.*, Vol. vol. 7， , 26 June 2014 (2014-06-26), pages 1-11	1-18
PX	YANG， J. F. et al. "Copy number variation analysis based on AluScan sequences." *Journal of Clinical Bioinformatics.*, Vol. vol. 4， , No. No. 15， , 05 December 2014 (2014-12-05), pages 1-14	1-16
A	CLIFFORD， R. J. et al. "Genetic Variations at Loci Involved in the Immune Response Are Risk Factors for Hepatocellular Carcinoma." *HEPATOLOGY.*, Vol. vol. 52， , No. No. 6， , 31 December 2012 (2012-12-31), pages 2034-2043	1-19
A	LONG， J. et al. "A Common Deletion in the APOBEC3 Genes and Breast Cancer Risk." *JNCI.*, Vol. vol. 105， , No. No. 8， , 17 April 2013 (2013-04-17), pages 573-579	1-19

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search **14 May 2015**	Date of mailing of the international search report **17 June 2015**
Name and mailing address of the ISA/CN **STATE INTELLECTUAL PROPERTY OFFICE OF THE P.R.CHINA** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer **PAN,Hao** Telephone No. **(86-10)82245455**

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KREPISCHI， A. C. V. et al. "Germline copy number variations and cancer predisposition." *Future Oncol.,* Vol. vol. 8， , No. No. 4， , 31 December 2012 (2012-12-31), pages 441-450	1-19
A	DISKIN， S. J. et al. "Copy number variation at 1q21.1 associated with neuroblastoma." *NATURE.,* Vol. vol. 459， , 18 June 2009 (2009-06-18), pages 987-991	1-19
A	WO 2011/048498 A2 (STICHTING HET NEDERLANDS KANKER INSTITUUT) 28 April 2011 (2011-04-28) the whole document	1-19

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2015/074606

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
WO 2011/048498 A2	28 April 2011	EP 2491141 A2	29 August 2012
		US 2012316080 A1	13 December 2012