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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 January 2019 (24.01.2019)

W!PO | PCT

(10) International Publication Number
WO 2019/016353 Al

(51) International Patent Classification:
G06F 19/22 (201 1.01) *C12Q 1/6869* (2018.01)

(21) International Application Number:
PCT/EP20 18/069720

(22) International Filing Date:
20 July 2018 (20.07.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/535,727 2 1 July 2017 (21 .07.2017) US

(71) Applicant *(for all designated States except DE, US):* **F, HOFFMANN-LA ROCHE AG** [CH/CH]; Grenzacherstrasse 124, 4070 Basel (CH).

(71) Applicant *(for DE only):* **ROCHE DIAGNOSTICS GMBH** [DE/DE]; Sandhofer StraBe 116, 68305 Mannheim (DE).

(71) Applicant *(for US only):* **ROCHE SEQUENCING SOLUTIONS, INC.** [US/US]; 4300 Hacienda Drive, Pleasanton, California 94588 (US).

(72) Inventors: **LAL, Preeti;** c/o Roche Sequencing Solutions, Inc., 4300 Hacienda Drive, Pleasanton, California 94588 (US). **LAM Y. K., Hugo;** c/o Roche Sequencing Solutions, Inc., 4300 Hacienda Drive, Pleasanton, California 94588 (US). **LEE, John;** c/o Roche Sequencing Solutions, Inc., 4300 Hacienda Drive, Pleasanton, California 94588 (US). **LOVEJOY, Alex;** c/o Roche Sequencing Solutions, Inc., 4300 Hacienda Drive, Pleasanton, California 94588 (US). **PALMA F., Johm;** c/o Roche Sequencing Solutions, Inc., 4300 Hacienda Drive, Pleasanton, California 94588 (US). **ROSENTHAL, Andre;** Groebener Dorfstr. 25, 14 194 Ludwigsfelde (DE). **YAO, Lijing;** c/o Roche Sequencing Solutions, Inc., 4300 Hacienda Drive, Pleasanton, California 94588 (US).

(74) Agent: **HILDEBRANDT, Martin** et al; Roche Diagnostics GmbH, Patent Department (LPP.....6164), P.O.Box 11 52, 82372 Penzberg (DE).

(54) **Title:** CLASSIFYING SOMATIC MUTATIONS FROM HETEROGENEOUS SAMPLE

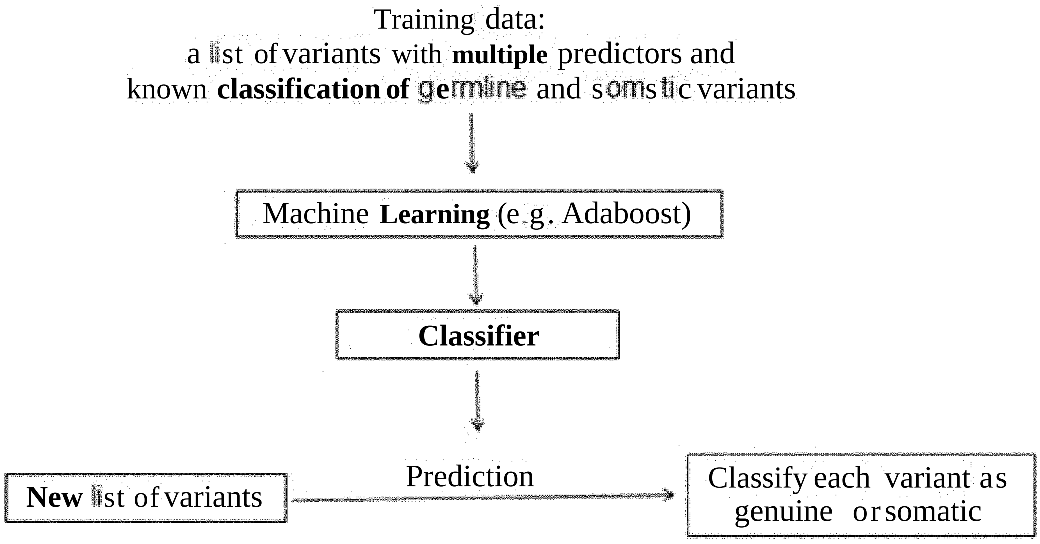

Training data:
a list of variants with **multiple** predictors and
known **classification of** germline and somatic variants
Machine **Learning** (e.g. Adaboost)
Classifier
New list of variants
Prediction
Classify each variant as
genuine or somatic

FIG. 4

(57) **Abstract:** In order to determine somatic mutations from germline mutations when matched normal sequences may not be available, adaboost machine learning algorithms are developed to classify germline mutations and somatic mutations. Three models were built based on different types of samples. The types samples can include, for example, fresh frozen samples, formalin-fixed paraffin-embedded (FFPE) samples, and plasma samples. The performances of the algorithms are evaluated with either ten-fold cross-validation or tested on independent set of samples.

[Continued on nextpage]

(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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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, 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))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

CLASSIFYING SOMATIC MUTATIONS FROM HETEROGENEOUS SAMPLE

BACKGROUND

[0001] Next generation sequencing (NGS) of solid tumors has been widely implemented in the last decade. These technologies are starting to play an essential role in our understanding of altered genes and pathways important in the development of cancer. Additionally, NGS is increasingly being used as the primary testing method in molecular diagnostics [1]. While current non-NGS methods, for example overexpression of Human epidermal growth factor receptor-2 (HER2) by immunohistochemistry (IHC) or fluorescence in situ hybridization (FISH), mostly detect only one variant type, NGS technology allows the tumor to be tested for multiple types of variations like single nucleotide variations (SNV), insertions/deletions (InDels), duplications, copy number variations (CNV), and translocations [2]. NGS also has the potential to detect genetic alterations which may be missed by traditional non-NGS methods [3]. NGS can be applied to solid tumors (tissue biopsy) or blood (liquid biopsy) from patients with malignancies. Most personalized medicine strategies rely on single tumor biopsy sample to determine specific mutations or aberrations. However, single tumor biopsy may lead to underestimation of the tumor genomic landscape, since extensive intratumor and intertumor heterogeneity is well documented [4] [5]. In contrast, NGS of blood has the potential to examine many aberrations in a non-invasive manner and thus deliver individualized health care for patients [6]. Liquid biopsy also potentially provides the advantage to minimize inter and intra tumor heterogeneity which may be underestimated by single sampling of tissue biopsy. Since liquid biopsy is non-invasive, multiple samples over the course of the disease can be easily obtained allowing for monitoring of changes in tumor in response to therapy. While liquid biopsy has many advantages over tissue biopsy, the main limitations are very low concentration of tumor derived material and uncertain nature of the tissue of origin.

[0002] Sequencing of both tissue and blood will typically identify mutations that were inherited (germline mutations) as well as those which accumulated during a person's lifetime (somatic mutations). Germline mutations occur in germ cells and are passed onto the next generation while somatic mutations occur in any cell and are typically not passed on to the next generation. A majority of somatic mutations are considered harmless (passenger mutations) since they do not have noticeable effects. However, some mutations may occur in genes or regions of the genome and may lead to a selective advantage to the cell. These mutations are called "driver mutations" and can lead to cell proliferation and eventual cancer[7]. An important goal of cancer sequencing is to identify driver somatic mutations so as to be able to, for example, target personalized therapies, monitor disease, determine possible resistance to therapy, and understand the mechanism of oncogenesis.

[0003] Sequencing and comparison of matched normal samples to cancer genome can allow accurate identification of somatic variants from germline variants. However, in routine clinical practice, it is difficult to always sequence matched normal genetic material mainly because of difficulty in obtaining matched blood or adjacent normal sample and increased cost.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] FIG. 1 illustrates an exemplary lineage of mitotic cell divisions from a fertilized egg to a single cell within a cancer, showing the timing of somatic mutations acquired by the cancer cell and the processes that may contribute to them[8].

[0005] FIG. 2 illustrates schematically how a patient's blood sample (or tissue sample) may show both germline variants inherited from the patient's father and mother and somatic variants.

[0006] FIG. 3 shows some exemplary databases and the predictors that may be used in a machine learning algorithm according to some embodiments of the present invention.

[0007] FIG. 4 illustrates schematically a method of classifying somatic variants from germline variants using machine learning according to some embodiments of the present invention.

[0008] FIG. 5 illustrates schematically an exemplary machine learning (ML) module according to an embodiment.

[0009] FIG. 6 illustrates an exemplary tree classifier according to an embodiment of the present invention.

[0010] FIG. 7 shows a confusion matrix for an exemplary model trained with The Cancer Genome Atlas (TCGA) tissue samples according to an embodiment of the present invention.

[0011] FIG. 8 illustrates the relative importance of various predictors used in the model trained with TCGA tissue samples according to an embodiment of the present invention.

[0012] FIG. 9 shows a confusion matrix for the model trained with TCGA tissue samples as tested on an independent dataset according to an embodiment of the present invention.

[0013] FIG. 10 illustrates a Receiver Operator characteristics (ROC) curve (i.e., AUC) of the model trained with TCGA tissue samples according to an embodiment of the present invention.

[0014] FIG. 11 shows a confusion matrix for an exemplary model trained with plasma samples according to an embodiment of the present invention.

[0015] FIG. 12 illustrates the relative importance of various predictors used in the model trained with plasma samples according to an embodiment of the present invention.

[0016] FIG. 13 shows a flowchart illustrating a method for classifying genomic variants in a heterogeneous sample including DNA from various cells of an organism according to an embodiment of the present invention.

[0017] FIG. 14 illustrates an example computer system that may be utilized to implement techniques disclosed herein.

[0018] FIG. 15 illustrates an example sequence analytical system according to an embodiment of the present invention.

DETAILED DESCRIPTION

[0019] In order to classify somatic mutations from germline mutations when matched normal sequences may not be available, machine learning algorithms are developed. Three models are built based on different types of samples. The types of samples can include, for example, fresh frozen samples, formalin-fixed paraffin-embedded (FFPE) samples, and plasma samples. The performances of the algorithms are evaluated either with ten-fold cross-validation or tested on independent set of samples.

[0020] In some embodiments, fresh frozen 569 samples from lung adenocarcinoma from TCGA are used to train a model, which can achieve a receiver operating characteristic (ROC) area under curve (AUC) of 0.9968 as tested on an independent data set of 490 lung squamous cell carcinoma samples available in TCGA. In some other embodiments, 9 in-house FFPE samples are used to train a model, which can achieve an ROC AUC of 0.991 by ten-fold cross-validation. In another embodiment, a model trained on 48 plasma samples can achieve an ROC AUC of 0.991 with accuracy of 0.998.

[0021] While sequencing data from matched normal samples may be an ideal approach to identify somatic mutations with high confidence, embodiments of the present invention demonstrate that machine learning can be utilized to identify somatic mutations in tissue samples or blood samples with high accuracy in absence of normal DNAs. Accordingly, the machine learning techniques can help to reduce the amount of sequencing, e.g., by removing the step of sequencing healthy cells of the subject. Thus, embodiments can provide cheaper and quicker techniques for identifying somatic mutations.

I. TUMOR MUTATION PROFILING

A. Germline mutations and somatic mutations

[0022] A cancer cell carries a copy of diploid genome; it can also carry set of differences from the progenitor fertilized egg. These differences from the original are called somatic mutations. Somatic mutations that do not occur in germ cells are not passed onto offspring, while the germline mutations are inherited from parents and are transmitted to offspring.

[0023] Cancer may arise due to somatically acquired mutations acquired during one's lifetime, for example as a result of errors which occur during cell replication or from exposure to carcinogens (e.g., tobacco smoke) or radiation (e.g., UV light from sun), or due to mutations in the germline of an individual. Cancers arising from germline inherited mutations are called inherited cancers. It is estimated that inherited genetic mutations play a major role in about 5 to 10 percent of all cancers[9]. For most cancers, multiple somatic mutations can be detected by sequencing. However, most of the mutations are not involved in the development of cancer. Such mutations are referred to as "passenger mutations." On the other hand, certain mutations in genes or regions of the genes, for example mutations that are involved in tumor suppression or oncogenesis may confer growth advantage on the cancer cell and may be positively selected, thus leading to cancer growth. Mutations that may lead to cancer are referred to as "driver mutations." A passenger mutation typically does not confer any growth advantage and is not considered to contribute to cancer. Many passenger mutations are found in cancer genome because these mutations often occur during cell division[8].

[0024] FIG. 1 illustrates an exemplary lineage of mitotic cell divisions from a fertilized egg to a single cell within a cancer, showing the timing of somatic mutations acquired by the cancer cell and the processes that may contribute to them. As illustrated, somatic mutations can be caused by environmental and lifestyle exposures during a person's lifetime[8].

B. *Single nucleotide polymorphisms (SNPs) and allele frequency (AF)*

[0025] Single nucleotide polymorphism (SNP) is a phenomenon that two or more than two nucleotides can occur at a specific locus in the genome among the population. For example, a majority of population may have A at a given locus while some individuals have C or T at that locus. While a majority of SNPs may not have phenotypic consequences, some of them may lead to changes, such as eye color, height and disease predisposition. Multiple consortiums are dedicated to build repositories of common SNPs by sequencing thousands of individuals and investigate disease predisposition using SNPs' linkage disequilibrium mapping[10].

[0026] A single-nucleotide variant (SNV) is a single nucleotide substitution that can occur in any frequency. It may be inherited from parents (e.g. SNPs) or arise in somatic cells. Somatic single nucleotide variations may contribute to cancer.

[0027] Allele frequency (AF) is the frequency of an allele at a particular locus. In DNA sequencing, allele frequency of a given variant is percentage of reads supporting that variant at that locus.

C. *Tissue biopsy and liquid biopsy*

[0028] Identifying genes involved in the development of cancer can be important for understanding cancer biology, for developing novel therapeutics for cancer treatment and monitoring, and for providing methods for cancer prevention and early diagnosis. The use of polymorphic markers, such as somatic mutations, can help identifying driver genes and genomic regions contributing to the cancer phenotype, as well as monitoring progression of cancer.

[0029] A tissue biopsy is a procedure to remove a piece of tissue or a sample of cells from a tumor so that it can be analyzed in a laboratory. Liquid biopsy is a test done on a sample of blood to look for cancer cells from a tumor that are circulating in the blood or for pieces of DNA from tumor cells that are in the blood. Liquid biopsy can be useful for comprehensive screening of cancer, diagnosing tumor staging, primary tumor profiling, and treatment monitoring. A liquid biopsy may be used to help find cancer at an early stage. It may also be used to help plan treatments or to find out how well the treatment is working or if a patient has relapsed. Being able to take multiple samples of blood over time may also help clinicians understand the kind of molecular changes taking place in a tumor, for example resistance to therapy and recurrence of the disease. Thus, liquid biopsy has the potential to enable clinicians make personalized decisions for a given therapy or change in therapy based on the tumor sequences.

D. Tumor mutation profiling for cancer diagnosis and treatments

[0030] Tumor molecular profiling by NGS from tissue or blood (liquid biopsy) holds tremendous potential to guide cancer treatments. Therapies targeting specific genetic alterations can be more effective than traditional chemotherapies when used in an appropriate patient population[11]. One of the challenges of liquid biopsies has been ensuring adequate accuracy of determining somatic variations. In a blood specimen (also referred to as a plasma sample), there may be very little tumor DNA present in the blood specimen.

[0031] Cancer genome sequencing may determine both germline variants and somatic variants. FIG. 2 illustrates schematically how a patient's blood sample (or tissue sample) may show both germline variants inherited from the patient's father and mother (e.g., variants A and B, respectively) and somatic variants (e.g., variant C). Since somatic variants in oncogenes and tumor suppressors mostly drive the development and growth of the tumor, it may be important to separate somatic variants from germline variants.

[0032] Sequencing of tumor DNA and comparing it to sequences obtained from the same person's normal samples will allow one to classify germline mutations and somatic mutations. However, often it is not possible to sequence matched normal tissue because it may be difficult to obtain a blood or adjacent normal samples, or it may result in increased cost. In some instances, patient may not provide consent to examine germline variants since these variants are shared with many family members and can be passed on to offspring[12].

[0033] Therefore, one of the challenges on using tumor mutation profiling for cancer diagnosis and treatment is how to determine whether a given mutation is a somatic mutation or a germline mutation in the absence of matched sequences from normal DNA.

II. MACHINE LEARNING AND CLASSIFICATION

[0034] Classification is a problem of categorization, identifying to which categories a new case belongs to. A classic example of classification is determining whether or not a given email is "spam". Two main algorithms for classification problem are statistical modeling and machine learning[13].

[0035] Machine learning algorithms for classification learn from a set of training data whose categories are known, and predict the category membership of a new observation. An algorithm that implements classification is referred as a classifier.

[0036] Each observation possesses a set of quantifiable properties, known as predictors, explanatory variables, or features. These properties may be categorical (e.g. "Female," "Male," for gender), ordinal (e.g. "tall," "medium" or "short"), continuous values (e.g. the blood pressure measures). Machine learning algorithms utilize a selected set of properties to learn from training data and then assign each new observation to different classes.

A. Binary classification

[0037] Binary classification is a type of classification problems, in which a given set of new observations will be assigned to only two classes based on the classification rules. Some typical binary classification tasks include medical testing to diagnose if a patient has certain disease or not, and quality check if a product passes or fails.

B. Decision tree learning

[0038] Decision tree learning is one of several successful approaches in machine learning and data mining. It uses a decision tree to build a prediction model, which classify an observation by sorting it down from the root node to the leaf node. Each node in the tree is a test of a particular feature and each branch descending from the node represents the outcome of the test. The leaf node indicates the class label of a given observation. The target class value in the decision tree can be both discrete and continuous. The advantages of decision tree learning include the ability to handle missing value, the ability to deal with heavily skewed data, and robustness to outliers[14].

C. *Ensemble method and boosting*

[0039] Ensemble methods of machine learning build a predictive model by integrating multiple learning algorithms (e.g. neural networks or decision trees) to improve predictive performance of a single unstable or weak classifier. Multiple studies showed that ensemble approach can have more accurate prediction than any single classifier in the ensemble[15].

[0040] Boosting is one of the popular ensemble approaches. It combines iteratively built simple classifiers with respect to emphasizing the training instances that are mis-classified in the previous classifier[16].

[0041] Adaboost, short for "adaptive boosting," is a common implementation of boosting. The adaptive aspect lies in the sense that each iterative simple learner is weighted in favor of cases misclassified in previous classifier. In each step of the sequence, Adaboost attempts to find a new classifier according to the current weight of each observation. The cases that are misclassified in the classifier will receive more weight in the next iteration and the cases that are correctly classified will receive less weight. In the final model, the classifiers with highest accuracy will weigh more than the less accurate classifier[16].

III. CLASSIFYING SOMATIC MUTATIONS FROM GERMLINE MUTATIONS USING MACHINE LEARNING

[0042] According to some embodiments of the present invention, machine learning algorithms are developed to classify somatic variants and germline variants. A set of rules may be learned from a set of training data.

A. Predictors and databases

[0043] A set of features or predictors may be selected for the machine learning. The predictors may relate to the genome databases that are available now. FIG. 3 shows some exemplary databases from which data and the predictors that may be used in a machine learning algorithm according to some embodiments of the present invention. Exemplary databases for germline variants may include Exome Aggregation Consortium (ExAC), Single Nucleotide Polymorphism Database (dbSNP), and 1000 Genomes Project. Exemplary databases for somatic variants may include The Cancer Genome Atlas (TCGA) and Catalogue Of Somatic Mutations In Cancer (COSMIC). Other databases may also be used.

[0044] The ExAC database and the 1000 Genomes Project database include germline SNPs identified from DNA sequences of thousands of people. If a variant identified in a sample exists in the ExAC database and/or the 1000 Genomes Project database, it may mean that the variant is common among the population, and therefore may have a higher likelihood of being a germline variant. The TCGA database and the COSMIC database include many somatic SNPs identified from DNA sequences of many tumor tissues. If a variant identified in a sample exists in the TCGA database and/or the COSMIC database, the variant may have a higher likelihood of being a somatic variant that occurs only in tumors.

[0045] Predictors or features may be generated based on the databases. For example, the population allele frequency (AF) of a variant in a database may indicate how frequent the variant shows up in a population. Whether a variant exists in a database and the number of counts a variant appears in a database can also be used as predictors.

[0046] Allele frequency can be a good predictor for classifying somatic variants from germline variants. Somatic variants have distinguished features as compared to germline variants. The allele frequency values for germline variants tend to cluster around 50% or 100% depending on if the variant is inherited from father, mother or both of the parents, whereas the allele frequency values for somatic variants are expected to be randomly distributed between 1% to *100%* depending on percentage of tumor cells that have the variation as well as percentage of tumor cells in a given specimen[17].

[0047] The consistency of allele frequencies across multiple time point samples (for example in liquid biopsy where it is easy to obtain multiple samples) can be a good predictor to classify somatic variants from germline variants. Because germline variants would show up in every sample, the allele frequencies of germline variants tend to be consistent across matched samples. In contrast, for somatic variants, as the patient goes through different stages of therapy and treatment, their allele frequencies can vary across matched samples. According to some embodiments, a statistic measure of consistency is used as one of the predictors in the machine learning model.

[0048] According to embodiments of the present invention, multiple predictors can be used in the machine learning algorithm. A set of rules based on the multiple predictors may be learned from a set of training data. The rationale for using multiple databases and multiple predictors may include: (1) usually no single database can be perfect or can include all germline SNPs and somatic SNVs; (2) simple cutoffs based on allele frequencies (AF) may not yield the best results because of the chances of missing low AF or not considering high AF; and (3) it may be difficult to come up with a set of rules to maximize the accuracy of the classification of somatic variants from germline variants.

1. Databases

[0049] The Exome Aggregation Consortium (ExAC) consolidated exome sequencing data from a variety of large-scale sequencing projects and made summary data public for the scientific community. ExAC data set contains sequencing data from 60,706 unrelated individuals. This data set can serve as a useful reference set of human genetic variation[18].

[0050] The Single Nucleotide Polymorphism Database (dbSNP) is a free public database for single nucleotide polymorphisms (SNPs) and multiple small-scale variations including insertion, deletion and microsatellites in different species. It is developed and hosted by the National Center for Biotechnology Information

(NCBI) in collaboration with the National Human Genome Research Institute (NHGRI). It also can serve as a reference set of human genetic variation[19].

[0051] 1000 Genomes Project is launched in 2008, aiming to comprehensively understand common human genetic variations. Thousands of individuals from 26 populations have been sequenced by whole-genome sequencing and a repository of human common genetic variants are made available for public scientific community[20].

[0052] The Cancer Genome Atlas (TCGA) consortium is collaboration of multiple institutions to understand cancer genome by integration of multi-omics data, including large-scale DNA/RNA sequencing, epigenome profile, transcriptome and proteome and clinic information. It has generated comprehensive cancer genome profiles for 33 cancer types. TCGA is publicly available, becoming a reservoir of information for cancer research[21].

[0053] The Catalogue Of Somatic Mutations In Cancer (COSMIC) is the world's largest database to store somatic mutations related to human cancer which are curated by expert scientists and supported by numbers of scientific publications. COSMIC collected somatic mutations for all cancer types and is available in a number of ways (e.g. graphic visualization of data and programmatic API). COSMIC provides useful resource and tools for cancer research [22].

2. Predictors

[0054] Exemplary predictors may include:

[0055] AF: allele frequency of variants in the sample (e.g., tissue sample or plasma sample);

[0056] CV: coefficient of variations of variants' AFs across multiple time points samples from the same patient (CV can be a consistency measure);

[0057] EXAC_AF: population frequency of variants in the ExAC database;

[0058] DBSNP_AF: population frequency of variants in the dbSNP database;

[0059] KG_AF: population frequency of variants in the 1000 Genomes Project database;

[0060] IN_EXAC: existence of variants in the ExAC database;

[0061] IN_DBSNP: existence of variants in the dbSNP database;

[0062] IN_KG: existence of variants in the 1000 Genomes Project database;

[0063] IN_TCGA: existence of variants in the TCGA database;

[0064] IN_COSMIC: existence of variants in the COSMIC database;

[0065] DBSNP_AQ: alignment quality in the dbSNP (l=unique mapping, 2=non-unique, 3=many matches);

[0066] DBSNP_COMMON: common variant in the dbSNP database;

[0067] COSMIC_COUNT: number of times variant appears in the COSMIC database;

[0068] TCGA_COUNT: number of times variant appears in the TCGA database;

[0069] WHITELIST: check if variants are white list variants that are known driver mutations, for example whitelist variants in AVENIO ctDNA assay. The AVENIO ctDNA Kit is a next-generation sequencing (NGS) liquid biopsy assay that can detect all four mutation classes (SNVs, inDels, fusions, and CNVs) in a single assay with high sensitivity and specificity[23].

[0070] Among the above-listed predictors, some of them are real-valued predictors. For example, AF, CV, EXAC_AF, DBSNP_AF, and KG_AF are real-valued predictors. Some of them are categorical predictors. For example, IN_EXAC, IN_DBSNP, IN_KG, IN_TCGA, IN_COSMIC, DBSNP_AQ, and DBSNP_COMMON (common variant in the dbSNP database) are categorical predictors. Some of them are integer-valued predictors. For example, COSMIC_COUNT and TCGA_COUNT are integer-valued predictors.

[0071] Other predictors can also be used, including but not limited to copy number variation (CNV), linkage disequilibrium database, tumor content in the specimen, and the like.

3. Consistency measure

[0072] According to some embodiments, consistency measure may be used as one of the predictors if multiple samples are available from the same patient. Multiple methods can be used to measure the consistency of the data. For example, range, interquartile range, variance, standard deviation, and coefficient of variation can be used as consistency measures.

[0073] Range (R) is a measure of the distance between the maximum number and the minimum number in a given dataset, calculated as follows,

$$R = maximum — minimum.$$

[0074] Interquartile range (IQR) is the difference between the upper quartile (Q3 : 75th percentile) and the lower quartile (Ql: 25th percentile) in a given dataset, calculated as follows,

$$IQR = Q3 - Ql.$$

[0075] Variance (σ^2) is the average squared distance from the mean in a given dataset, calculated as follows,

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{\mathbf{n\text{-}l}},$$

where x is the value of AF, n is the number of samples in the dataset, and μ is the mean AF calculated as follows,

$$\mu = \frac{\Sigma x}{n}.$$

[0076] Standard deviation (σ) is the square root of variance in a given dataset.

[0077] The coefficient of variation (CV) can be calculated using the values of μ and σ as,

$$CV = \frac{\sigma}{\mu}.$$

[0078] For example, assuming that variant A's AF values are 14% and 15% at Sample 1 and Sample 2, respectively, the values of μ and σ can be calculated as follows,

$$\mu = \frac{(14\%+15\%)}{2} = 14.5\%\text{; and}$$

$$\sigma = \sqrt{\frac{(\text{ï}\,4\%-\mu)^{2}+(15\%-\mu)^{2}}{2-1}} = .00707\ .$$

B. Machine learning algorithms

[0079] FIG. 4 illustrates schematically a method of classifying somatic variants from germline variants using machine learning according to some embodiments of the present invention. A training dataset may include a list of variants with multiple predictors and known classifications. A machine learning (ML) algorithm, such as adaboost, is used to build a classifier model trained on the training dataset. Once the classifier model has been trained, it can be used to classify a new list of variants.

1. Adaboost ML algorithms

[0080] FIG. 5 illustrates schematically an exemplary adaboost machine learning (ML) module according to an embodiment. Adaboost is an ensemble method of machine learning that can create a strong classifier from a number of weak classifiers. The ML module may include a plurality of classifiers. For example, the ML module can include three classifiers as illustrated in FIG. 5. Each classifier may use a respective set of predictors 502, 504, or 506 to produce a respective prediction. The predictions from the plurality of classifiers can then be combined into a weighted sum to provide a final prediction. In other embodiments, a ML module can include more or fewer number of classifiers.

2. Tree classifier with multiple predictors

[0081] Each classifier can be a tree classifier. FIG. 6 illustrates an exemplary tree classifier according to an embodiment. The leaves of the tree represent classification labels (i.e., "somatic" or "germline"). The branches of the tree represent conjunctions of features, referred herein as "predictors," that lead to those classification labels.

[0082] For instance, in the example illustrated in FIG. 6, the tree classifier can evaluate an input variant at a first predictor 602 relating to EXAC_AF (population allele frequency of variants in the ExAC database). If the value of EXAC_AF is greater than 0.2%, the tree classifier evaluates the variant at a second predictor 604 relating to DBSNP_AF (population frequency of variants in the dbSNP database). If the value of DBSNP_AF is greater than 2%, the variant is classified as "germline." If the value of DBSNP_AF is not greater than 2%, the tree classifier evaluates the variant at a third predictor 606 relating to IN_COSMIC (existence of variants in the COSMIC database).

[0083] If the variant exists in the COSMIC database, the tree classifier evaluates the variant at a fourth predictor 608 relating to COSMIC_COUNT (number of times variant appears in the COSMIC database). If the number of times variant appears in the COSMIC database is greater than 3, the variant is classified as "somatic." If the number of times variant appears in the COSMIC database is not greater than 3, the tree classifier evaluates the variant at a fifth predictor 610 relating to AF (allele frequency of variants). If the allele frequency is greater than 80%, the variant is classified as "germline." If the allele frequency is not greater than 80%, the variant is classified as "somatic."

[0084] Referring back to the third predictor 606 relating to IN_COSMIC, If the variant does not exist in the COSMIC database, the tree classifier evaluates the variant at a sixth predictor 616 relating to AF (allele frequency of variants). If the allele frequency is less than 10%, the variant is classified as "somatic." If the allele frequency is not less than 10%, the variant is classified as "germline."

[0085] Referring back to the first predictor 602, if the value of EXAC_AF is not greater than 0.2%, the tree classifier evaluates the variant at a seventh predictor 612 relating to IN_TCGA (existence of variants in the TCGA database). If the variant exists in the TCGA database, the variant is classified as "somatic." If the variant does not exist in the TCGA database, the tree classifier evaluates the variant at an eighth predictor 614 relating to AF (allele frequency of variants). If the allele frequency is greater than 40%, the variant is classified as "germline." If the allele frequency is not greater than 40%, the variant is classified as "somatic."

[0086] It should be understood that the decision tree illustrated in FIG. 6 is only an exemplary decision tree according to one embodiment. Decision trees with more or fewer predictors and with different predictors may be used according to some other embodiments.

[0087] Referring to FIG. 5, different classifiers can have different decision trees, with different features and different number of features or predictors. For example, one classifier may include 5 features, and another classifier may include 10 features. In some embodiments, each classifier may include about 4 to 10 predictors, or about 4 to 8 predictors. Also, the threshold values for the real-valued or integer-valued predictors can be different in different classifiers.

[0088] Each classifier may produce a probability of whether a variant is a germline variant or a somatic variant. Referring again to FIG. 5, the probabilities produced by the plurality of classifiers may be combined to produce a weighted sum that represents a final probability. The adaboost machine learning (ML) module can include a cutoff or threshold value for the probability for classifying somatic or germline variants. For example, the cutoff can be 0.5. A variant may be classified as somatic if the probability is greater than 0.5, and germline otherwise. The cutoff value can be changed based on certain criteria.

[0089] One of the advantages of adaboost machine learning algorithm is that there is no need to normalize the training data, as suitable weights can be provided for each feature. Thus, raw numbers of counts can be used, as opposed to all features being normalized (e.g., the allele frequency AF). An example of a raw count is number of samples having the mutation in a database. The prediction can be robust whether a feature is positive or negative. In addition, the adaboost machine learning algorithm can prevent overfitting.

IV. PREDICTION PERFORMANCE

A. Cross-validation and test on independent dataset

[0090] Performing machine learning usually involves the following:

[0091] Training phase: train the machine learning model with a training dataset, by pairing the input with expected output;

[0092] Validation/test phase: estimate how well the model has been trained by estimating model performance characteristics (e.g., classification errors for classifiers, etc.) using a validation dataset and/or a test dataset; and

[0093] Application phase: apply the model to the real-world data and get the results.

[0094] The validation/test phase often has two parts: (1) validation - evaluating the model and tuning parameters using a validation dataset; and (2) test - estimating the accuracy of the selected classifier on a test dataset.

[0095] Cross-validation is a method of assessing predictive performance of a machine learning algorithm on an independent dataset that is not used to train the model. This involves partitioning a dataset into a training dataset and a validation dataset. One of the advantages of cross-validation is to prevent overfitting since it is tested on an independent dataset. There are multiple common cross-validation methods, including k-fold cross-validation, holdout, leave-one-out and so on[24].

[0096] In k-fold cross-validation, the original dataset is randomly split into k equal sized independent subgroups. In turns, a single subgroup is retained as the validation data for testing the model and the union of remaining subgroups are used to train the model. Each of these k subgroups will be selected once as validation data set. Then the average of k results will be the classifier performance estimation. The advantages of this method include: (1) the entire dataset is used for both training and validation for which each observation is exactly used once; and

(2) validation data in each round is independent from training data, thus preventing overfitting[25].

[0097] A test dataset is a set of data used to assess predictive performance of a classifier.

B. Model performance characteristics

1. ROC and AUC

[0098] In statistics, a receiver operating characteristic (ROC) curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) for different decision threshold. The true positive rate is also known as sensitivity or recall. The false positive rate is 1- specificity. An ROC curve can evaluate the discrimination ability of a classifier in terms of sensitivity and specificity and identify the optimal decision threshold. The ROC curve for a model with no discrimination capacity would be a 45-degree diagonal line. The area under the curve (AUC) is the probability that a classifier scores higher for randomly selected positive observation than a randomly selected negative observation. AUC can summarize the performance of a classifier with a single value[26].

2. Accuracy and confusion matrix

[0099] A confusion matrix in machine learning is a table that summarizes the prediction accuracy on a classification problem. Each row of the table represents the number of instances in each predicted class and each column represents the number of instances in each truth classes. A confusion matrix can also show the accuracy of predictions for each class.

C. Exemplary models and their performances

[0100] Three models are built with different types of samples according to some embodiments. The types of samples can include, for example, fresh frozen samples, formalin-fixed paraffin-embedded (FFPE) samples, and plasma samples. The models are validated either with ten-fold cross-validation method or tested on independent set of samples.

1. Data sources

[0101] Exemplary data sources can include the following TCGA lung cancer samples (whole exome-sequencing data on fresh frozen tissue): 490 lung squamous cell carcinoma (LUSC) and 569 lung adenocarcinoma (LUAD) samples.

2. Prediction performance in TCGA tissue samples

[0102] FIG. 7 shows a confusion matrix for an exemplary model trained with TCGA tissue samples according to some embodiments. The training dataset includes about 569 LUAD samples. A total of about 62913 variances are tested in the training dataset. A ten-fold cross-validation is used. As illustrated, the model can achieve an accuracy of about 0.9974, and an AUC value of about 0.9992.

[0103] FIG. 8 illustrates the relative importance of various predictors used in the model. As illustrated, EXAC_AF (population frequency of variants in the ExAC database), DBSNP_AF (population frequency of variants in the dbSNP database), and KG_AF (population frequency of variants in the 1000 Genomes Project database) can be more important predictors as compared to the other predictors.

[0104] FIG. 9 shows a confusion matrix for the model trained on LUAD samples as tested on an independent dataset. The test dataset includes about 490 LUSC samples. A total of about 53516 variants are tested in the test dataset. As illustrated, the model can achieve an accuracy of about 0.9968, and an AUC value of about 0.9989.

[0105] FIG. 10 illustrates an ROC curve (i.e., AUC) of the model.

3. Prediction performance in house plasma samples

[0106] FIG. 11 shows a confusion matrix for an exemplary model trained with 48 plasma samples according to some embodiments. A ten-fold cross-validation is used. As illustrated, the model can achieve an accuracy of about 0.991, and an AUC value of about 0.998.

[0107] FIG. 12 illustrates the relative importance of various predictors used in the model. As illustrated, AF (allele frequency of variants), EXAC_AF (population frequency of variants in the ExAC database), and DBSNP _AF (population frequency of variants in the dbSNP database) can be more important predictors as compared to the other predictors.

[0108] Note that the most important predictor is different between the plasma samples (as illustrated in FIG. 12) and the tissue samples (as illustrated in FIG. 8). In the case of plasma samples illustrated in FIG. 12, the AF of the sample can be the most important predictor. The AF of the sample can vary significantly in a plasma sample as the tumor cell free DNA concentration can vary significantly. Thus, there can be AFs that are around 10% for somatic mutations (corresponding to tumor concentration), and AFs that are around 50% or 100% for germline mutations where nearly all the cells would have such mutations since it is germline. The AF of 50%, would occur in cases where the variant is heterozygous, and AF of 100%, would occur in cases where the variant is homozygous.

[0109] Targeted sequencing can be performed by amplifying DNA fragments corresponding to certain parts of the genome (e.g., using certain primers) and/or capturing DNA fragments corresponding to certain parts of the genome (e.g., using capture probes).

V. METHOD OF CLASSIFYING SOMATIC MUTATIONS FROM GERMLINE MUTATIONS

[0110] FIG. 13 shows a flowchart illustrating a method for classifying cancer-specific variants in cancer genome sequencing data, including cancer-specific variants and germline variants, according to some embodiments of the present invention.

[0111] At 1302, sequencing is performed on the heterogeneous sample to obtain a plurality of sequence reads. The sequencing can be performed using techniques known to one skilled in the art. As examples, the sequencing can be performed using sequencing-by-synthesis, nanopore sequencing, or combinations thereof. The signals from the sequencing can be of various types, e.g., light signals from probes or electrical signals (e.g., voltage or current). The sequence reads can be of the entire DNA/RNA fragment or part(s) of it, e.g., the ends of the DNA/RNA fragment.

[0112] At 1304, the plurality of sequence reads is aligned to a reference genome corresponding to the organism to determine genomic locations for the plurality of sequences in the reference genome. Alignment can be performed using techniques known to one skilled in the art, such as exist in software programs like basic local alignment search tool (BLAST), BLAST-like alignment tool (BLAT), BWA, SOAP, and Bowtie, which can use a Burrows- Wheeler transform.

[0113] At 1306, the plurality of sequence reads is received at a computer system.

[0114] At 1308, a first genomic location in the reference genome exhibiting a first allele that is different than a reference allele of the reference genome is identified by the computer system based on an analysis of bases of sequence reads that align to the first genomic location. For example, the different bases on reads aligned to a location can be tracked. If a base is different than that of the reference genome, the location can be identified and the base be identified as part of a potential mutation.

[0115] At 1310, one or more first values for one or more first features are determined by the computer system using a first amount of sequence reads having the first allele at the first genomic location. An example of the first feature is an allele fraction (AF). The allele fraction of a particular allele at a location can be computed as the percentage (or fraction) of all reads aligned to the location that have the particular allele at the location. Other examples of first features are provided herein.

[0116] At 1312, one or more second values for one or more second features corresponding to population statistics of the first allele at the first genomic location (e.g., in one or more databases) corresponding to somatic mutations and/or germline mutations may be determined by the computer system. Examples of such second features are EXAC_AF, DBSNP_AF, and KG_AF. Step 1312 may be optional.

[0117] At 1314, a classification model is retrieved by the computer system from memory. The classification model is trained using other cancer genome sequencing data of the same type. For example, the samples can all be plasma, all be serum, all be fresh frozen tissue samples, or all be formalin-fixed paraffin-embedded (FFPE) tissue samples. The other heterogeneous samples include one or more genomic locations labeled as being a somatic mutation or a germline mutation. Some samples include the first genomic location, but not all of the samples need to include the first genomic location. The training uses values of the one or more first features and the one or more second features as determined for each of the one or more genomic location of the other cancer genome sequencing data.

[0118] Examples of classification models are provided herein, e.g., decision trees, support vector machines, neural networks, etc. The classification model may be composed of sub-models of an ensemble model, as is described herein.

[0119] At 1316, the one or more first values for the one or more first features and the one or more second values for the one or more second features are input by the computer system into the classification model. Once the features are determined, the function of the classification model can be invoked with the features as input variables.

[0120] At 1318, whether the first allele is somatic or germline is determined by the classification model executing on the computer system. The determination can be determined using a cutoff value, e.g., where the classification model provide a probability of being a somatic mutation or a germline mutation. The classification can be binary, with the cutoff value discriminating between the two classifications. In other embodiments, more classifications can exist (e.g., an indeterminate classification), where more than one cutoff can exist: each one discriminating between one pair of classifications.

VI. COMPUTER SYSTEM

[0121] Any of the computer systems mentioned herein may utilize any suitable number of subsystems. Examples of such subsystems are shown in FIG. 14 in computer system 10. In some embodiments, a computer system includes a single computer apparatus, where the subsystems can be the components of the computer apparatus. In other embodiments, a computer system can include multiple computer apparatuses, each being a subsystem, with internal components. A computer system can include desktop and laptop computers, tablets, mobile phones and other mobile devices.

[0122] The subsystems shown in FIG. 14 are interconnected via a system bus 75. Additional subsystems such as a printer 74, keyboard 78, storage device(s) 79, monitor 76, which is coupled to display adapter 82, and others are shown. Peripherals and input/output (I/O) devices, which couple to I/O controller 71, can be connected to the computer system by any number of means known in the art such as input/output (I/O) port 77 (e.g., USB, FireWire®). For example, I/O port 77 or external interface 81 (e.g. Ethernet, Wi-Fi, etc.) can be used to connect computer system 10 to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus 75 allows the central processor 73 to communicate with each subsystem and to control the execution of a plurality of instructions from system memory 72 or the storage device(s) 79 (e.g., a fixed disk, such as a hard drive, or optical disk), as well as the exchange of information between subsystems. The system memory 72 and/or the storage device(s) 79 may embody a computer readable medium. Another subsystem is a data collection device 85, such as a camera, microphone, accelerometer, and the like. Any of the data mentioned herein can be output from one component to another component and can be output to the user.

[0123] A computer system can include a plurality of the same components or subsystems, e.g., connected together by external interface 81 or by an internal interface. In some embodiments, computer systems, subsystem, or apparatuses can communicate over a network. In such instances, one computer can be considered a client and another computer a server, where each can be part of a same computer system. A client and a server can each include multiple systems, subsystems, or components.

[0124] Aspects of embodiments can be implemented in the form of control logic using hardware (e.g. an application specific integrated circuit or field programmable gate array) and/or using computer software with a generally programmable processor in a modular or integrated manner. As used herein, a processor includes a single-core processor, multi-core processor on a same integrated chip, or multiple processing units on a single circuit board or networked. Based on the disclosure and teachings provided herein, a person of ordinary skill in the art will know and appreciate other ways and/or methods to implement embodiments of the present invention using hardware and a combination of hardware and software.

[0125] Any of the software components or functions described in this application may be implemented as software code to be executed by a processor using any suitable computer language such as, for example, Java, C, C++, C#, Objective-C, Swift, or scripting language such as Perl or Python using, for example, conventional or object-oriented techniques. The software code may be stored as a series of instructions or commands on a computer readable medium for storage and/or transmission. A suitable non-transitory computer readable medium can include random access memory (RAM), a read only memory (ROM), a magnetic medium such as a hard-drive or a floppy disk, or an optical medium such as a compact disk (CD) or DVD (digital versatile disk), flash memory, and the like. The computer readable medium may be any combination of such storage or transmission devices.

[0126] Such programs may also be encoded and transmitted using carrier signals adapted for transmission via wired, optical, and/or wireless networks conforming to a variety of protocols, including the Internet. As such, a computer readable medium may be created using a data signal encoded with such programs. Computer readable media encoded with the program code may be packaged with a compatible device or provided separately from other devices (e.g., via Internet download). Any such computer readable medium may reside on or within a single computer product (e.g. a hard drive, a CD, or an entire computer system), and may be present on or within different computer products within a system or network. A computer system may include a monitor, printer, or other suitable display for providing any of the results mentioned herein to a user.

[0127] Any of the methods described herein may be totally or partially performed with a computer system including one or more processors, which can be configured to perform the steps. Thus, embodiments can be directed to computer systems configured to perform the steps of any of the methods described herein, potentially with different components performing a respective steps or a respective group of steps. Although presented as numbered steps, steps of methods herein can be performed at a same time or in a different order. Additionally, portions of these steps may be used with portions of other steps from other methods. Also, all or portions of a step may be optional. Additionally, any of the steps of any of the methods can be performed with modules, units, circuits, or other means for performing these steps.

VII. DNA SEQUENCE ANALYTICAL SYSTEM

[0128] FIG. 15 illustrates a sequence analytical system 1500 according to an embodiment of the present invention. System 1500 as shown includes a sample 1505, such as a DNA molecule, within a sample holder 1510, e.g., a flow cell or a tube containing droplets of DNA. A physical characteristic 1515, such as a fluorescence intensity value, from sample 1505 is detected by a detector 1520. A data signal 1525 from detector 1520 can be sent to analysis system 1530, which may include a processor 1550 and a memory 1535. Data signal 1525 may be stored locally in analysis system 1530 in memory 1535, or externally in an external memory 1540 or a storage device 1545.

[0129] Detector 1520 can detect a variety of physical signals, such as light (e.g., fluorescent light from different probes for different bases) or electrical signals (e.g., as created from a molecule traveling through a nanopore).

[0130] Analysis system 1530 may be, or may include, a computer system, ASIC, microprocessor, etc. It may also include or be coupled with a display (e.g., monitor, LED display, etc.) and a user input device (e.g., mouse, keyboard, buttons, etc.). Analysis system 1530 and the other components may be part of a stand-alone or network connected computer system, or they may be directly attached to or incorporated in a thermal cycler device. Analysis system 1530 may also include optimization software that executes in processor 1550.

[0131] The specific details of particular embodiments may be combined in any suitable manner without departing from the spirit and scope of embodiments of the invention. However, other embodiments of the invention may be directed to specific embodiments relating to each individual aspect, or specific combinations of these individual aspects.

[0132] The above description of example embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form described, and many modifications and variations are possible in light of the teaching above.

[0133] A recitation of "a," "an" or "the" is intended to mean "one or more" unless specifically indicated to the contrary. The use of "or" is intended to mean an "inclusive or," and not an "exclusive or" unless specifically indicated to the contrary. Reference to a "first" component does not necessarily require that a second component be provided. Moreover reference to a "first" or a "second" component does not limit the referenced component to a particular location unless expressly stated.

[0134] All patents, patent applications, publications, and descriptions mentioned herein are incorporated by reference in their entirety for all purposes. None is admitted to be prior art.

VIII. REFERENCES

[0135] 1. McCutcheon, J. N. & Giaccone, G. Next-Generation Sequencing: Targeting Targeted Therapies. *Clin. Cancer Res.* **21,** 3584-3585 (2015).

[0136] 2. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* **11,** 685-696 (2010).

[0137] 3. Drilon, A. *et al.* Broad, Hybrid Capture-Based Next-Generation Sequencing Identifies Actionable Genomic Alterations in Lung Adenocarcinomas Otherwise Negative for Such Alterations by Other Genomic Testing Approaches. *Clin. Cancer Res.* **21,** 3631-3639 (2015).

[0138] 4. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366,** 883-892 (2012).

[0139] 5. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376,** 2109-2121 (2017).

[0140] 6. FDA Approves First Liquid Biopsy CDx for Non-Small Cell Lung Cancer. *Clinical OMICs* **3,** 5, 14 (2016).

[0141] 7. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349,** 1483-1489 (2015).

[0142] 8. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719-724 (2009).

[0143] 9. Mangold, E., Friedl, W. & Propping, P. [Hereditary colorectal carcinoma: predictive diagnosis and genetic counseling]. *Praxis (Bern 1994)* **90,** 490-496 (2001).

[0144] 10. Syvanen, A. C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* **2,** 930-942 (2001).

[0145] 11. Stegmeier, F., Warmuth, M., Sellers, W. R. & Dorsch, M. Targeted cancer therapies in the twenty-first century: lessons from imatinib. *Clin. Pharmacol. Ther.* **87,** 543-552 (2010).

[0146] 12. Jones, S. *et al.* Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci TransI Med* **7,** 283ra53-283ra53 (2015).

[0147] 13. Harper, P. R. A review and comparison of classification algorithms for medical decision making. *Health Policy* **71,** 315-331 (2005).

[0148] 14. Song, Y.-Y. & Lu, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry* **27,** 130-135 (2015).

[0149] 15. Rokach, L. Ensemble-based classifiers. *Artificial Intelligence Review* (2010).

[0150] 16. Journal of Statistical Software. 1-27 (2006).

[0151] 17. Meric-Bernstam, F., Brusco, L. & Daniels, M. Incidental germline variants in 1000 advanced cancers on a prospective somatic genomic profiling protocol. *Annals of* ... (2016).

[0152] 18. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285-291 (2016).

[0153] 19. Smigielski, E. M., Sirotkin, K. & Ward, M. dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids* ... (2000).

[0154] 20. Consortium, T. 1. G. P. A global reference for human genetic variation. *Nature* **526,** 68-74 (2015).

[0155] 21. Tomczak, K., Czerwinska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19,** A68-77 (2015).

[0156] 22. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805-1 1 (2015).

[0157] 23. AVENIO ctDNA Surveillance Kit.

[0158] Available at: (Accessed: 22nd June 2017)

[0159] 24. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai* (1995).

[0160] 25. Anguita, D., Ghio, A., Ridella, S. & Sterpi, D. K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines. *DMIN* (2009).

[0161] 26. Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med* **4,** 627-635 (2013).

PATENT CLAIMS

1. A method of classifying genomic variants in a heterogeneous sample including DNA from various cells of an organism, the method comprising:
sequencing DNA from the heterogeneous sample to obtain a plurality of sequence reads;
aligning the plurality of sequence reads to a reference genome corresponding to the organism to determine genomic locations for the plurality of sequences in the reference genome;
receiving the plurality of sequence reads at a computer system;
identifying, by the computer system, a first genomic location in the reference genome exhibiting a first allele that is different than a reference allele of the reference genome based on an analysis of bases of sequence reads that align to the first genomic location;
determining, by the computer system, one or more first values for one or more first features using a first amount of sequence reads having the first allele at the first genomic location;
retrieving, by the computer system from memory, a classification model that is trained using other heterogeneous samples of a same type, the other heterogeneous samples including one or more genomic locations labeled as being a somatic mutation or a germline mutation, wherein the training uses first values of the one or more first features as determined for each of the one or more genomic location of the other heterogeneous samples;
inputting, by the computer system, the one or more first values for the one or more first features into the classification model; and
determining, by the classification model executing on the computer system, whether the first allele is somatic or germline.

2. The method of claim 1, further comprising:
determining, by the computer system, one or more second values for one or more second features corresponding to population statistics of the first allele at the first genomic location corresponding to somatic mutations and/or germline mutations, wherein the training also uses second values of the one

or more second features as determined for each of the one or more genomic location of the other heterogeneous samples; and
inputting, by the computer system, the one or more second values for the one or more second features into the classification model.

3. The method of claim 1, wherein the classification model includes:
a plurality of classifiers, each classifier providing a mutation classification of the first allele at the first genomic location;
assigning a weight to each of plurality of classifiers;
determining a weighted sum of the plurality of classifiers that all have the mutation classification being somatic mutation or that all have the mutation classification being germline mutation; and
comparing the weighted sum to a threshold value to determine whether the first allele is somatic or germline.

4. The method of claim 3, wherein the plurality of classifiers are decision trees.

5. The method of claim 1, wherein the amount of sequencing required to classify genomic variants is reduced as a result of the classifier model.

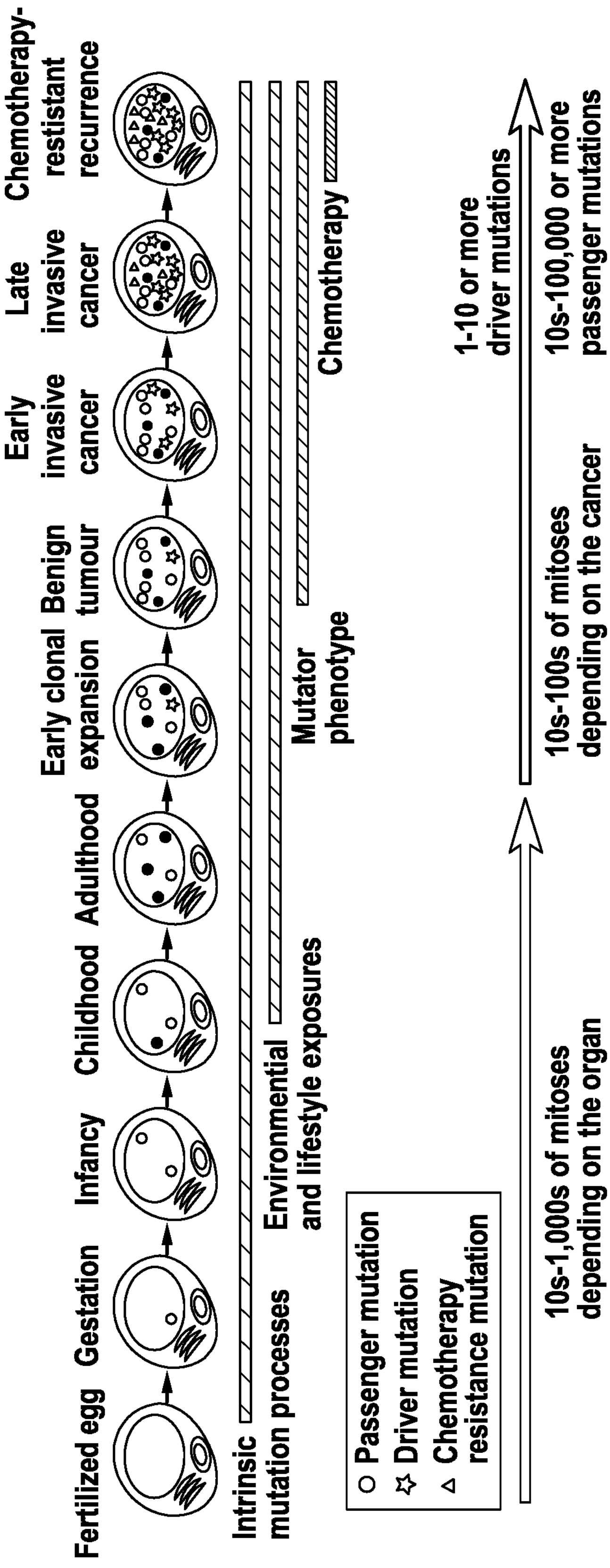
6. A computer product comprising a computer readable medium storing a plurality of instructions for controlling a computer system to perform operations of any of the methods above.

7. A system comprising:
the computer product of claim 6; and
one or more processors for executing instructions stored on the computer readable medium.

8. A system comprising means for performing any of the methods above.

9. A system configured to perform any of the above methods.

10. A system comprising modules that respectively perform the steps of any of the above methods.


Fertilized egg
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late invasive cancer
Chemotherapy-restistant recurrence
Intrinsic mutation processes
Environmental and lifestyle exposures
Mutator phenotype
Chemotherapy
○ Passenger mutation
☆ Driver mutation
△ Chemotherapy resistance mutation
10s-1,000s of mitoses depending on the organ
10s-100s of mitoses depending on the cancer
1-10 or more driver mutations
10s-100,000 or more passenger mutations

MR Stratton et al. Nature 458,719-724(2009)doi:10.1038/nature07943

Fig. 1

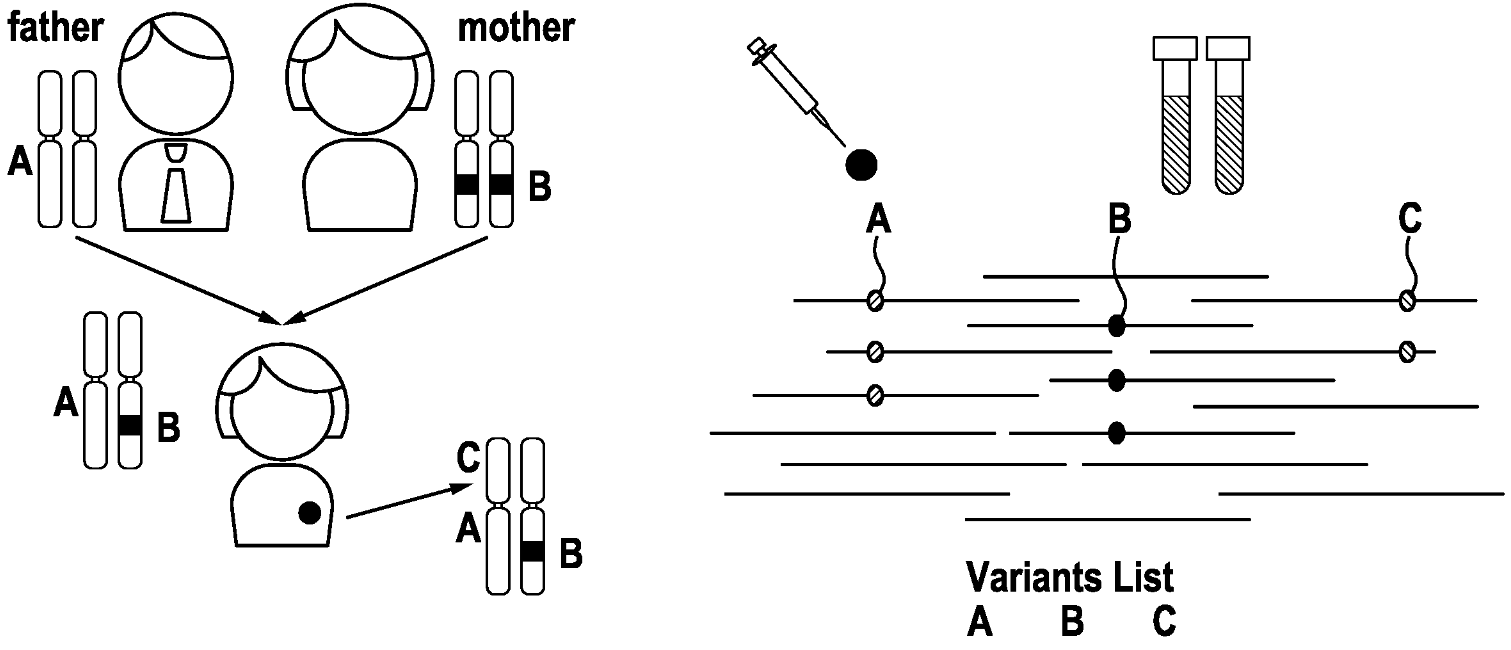


Fig. 2

	Germline	Somatic
Databases	EXAC, DbSNP and 1000 Genomes Project	TCGA, COSMIC
AF %	50% or 100%	Wide distribution a number between 0-100%
AF across matched samples	Consistent	Vary

FIG. 3

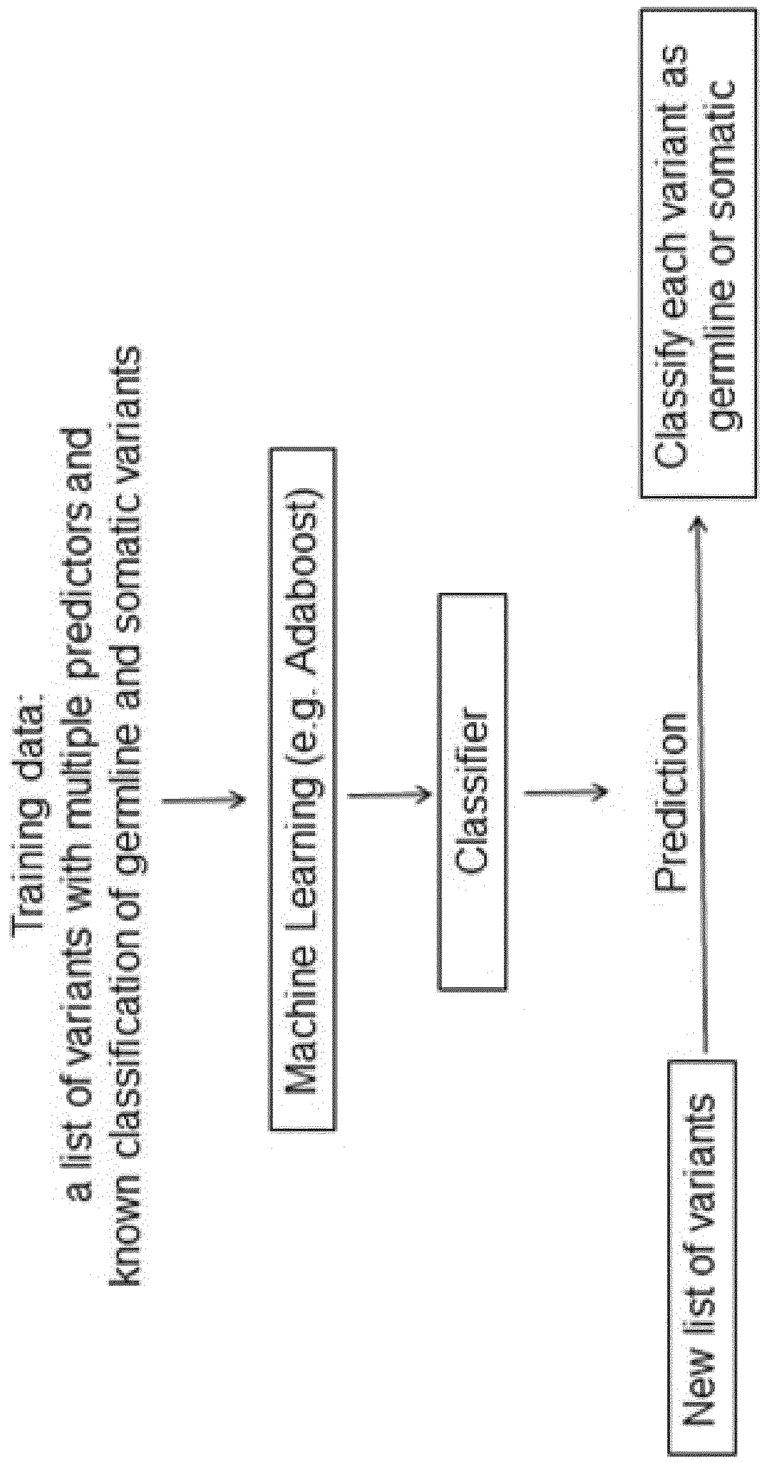

Training data:
a list of variants with multiple predictors and
known classification of germline and somatic variants
Machine Learning (e.g. Adaboost)
Classifier
Prediction
New list of variants
Classify each variant as
germline or somatic

FIG. 4

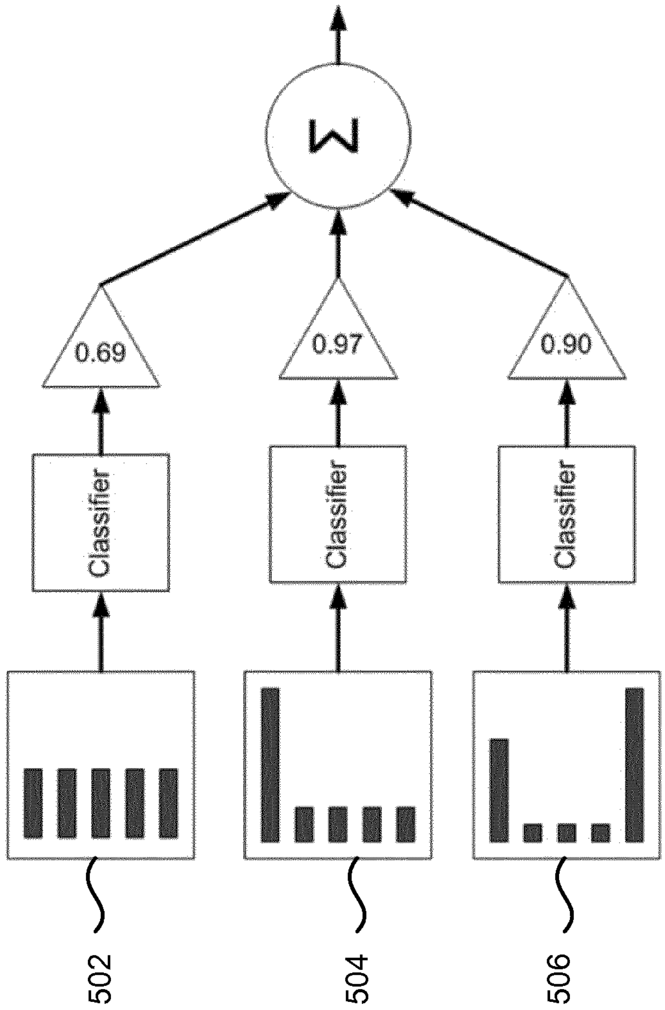


FIG. 5

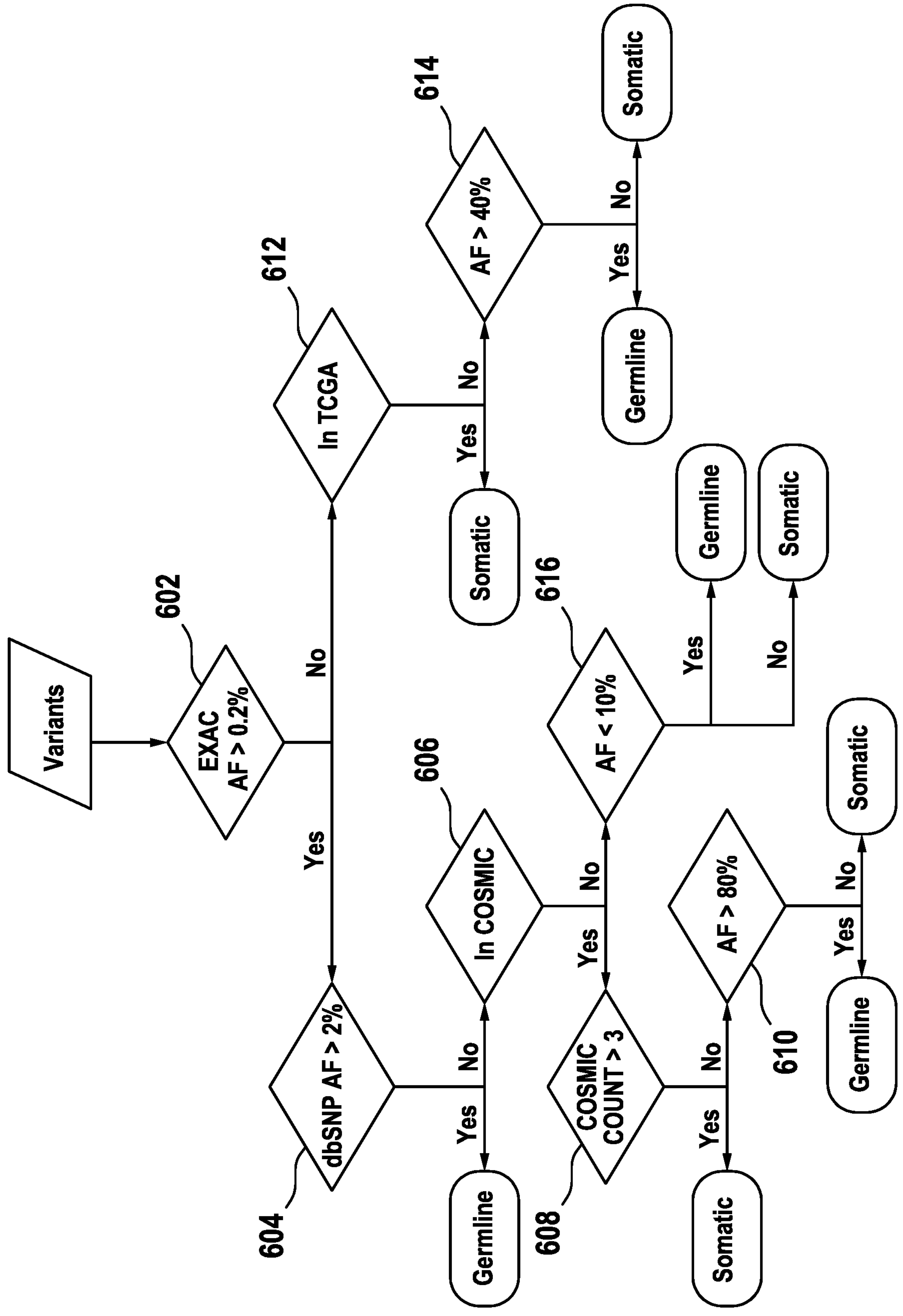

Variants
602
EXAC
AF > 0.2%
Yes
No
604
dbSNP AF > 2%
Yes
No
Germline
606
In COSMIC
Yes
No
608
COSMIC
COUNT > 3
Yes
No
Somatic
610
AF > 80%
Yes
No
Germline
Somatic
616
AF < 10%
Yes
No
Germline
Somatic
612
In TCGA
Yes
No
Somatic
614
AF > 40%
Yes
No
Germline
Somatic

Fig. 6

10-fold cross-validation in TCGA LUAD (WES)
Confusion matrix

N=62913		Predicted		
		Somatic	Germline	Classification error
Truth	Somatic	2653	86	0.0314
	Germline	75	60099	0.0012

Accuarcy: 0.9974
AUC: 0.9992

FIG. 7

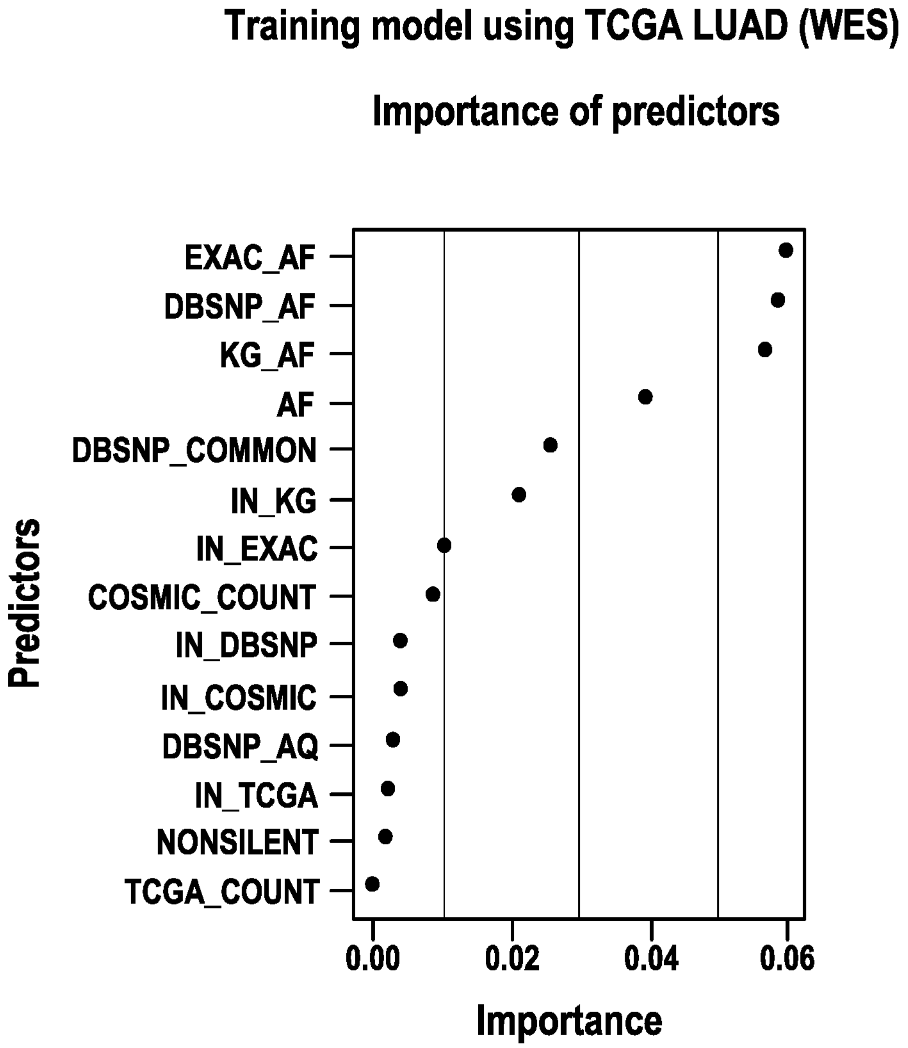

Training model using TCGA LUAD (WES)
Importance of predictors
Predictors
EXAC_AF
DBSNP_AF
KG_AF
AF
DBSNP_COMMON
IN_KG
IN_EXAC
COSMIC_COUNT
IN_DBSNP
IN_COSMIC
DBSNP_AQ
IN_TCGA
NONSILENT
TCGA_COUNT
0.00
0.02
0.04
0.06
Importance

Fig. 8

Training model: TCGA LUAD
Test on TCGA LUSC (WES)
Confusion matrix

N=53516		Predicted		
		Somatic	Germline	Classification error
Truth	Somatic	2300	104	0.0433
	Germline	69	51043	0.0013

Accuarcy: 0.9968
AUC: 0.9989

FIG. 9

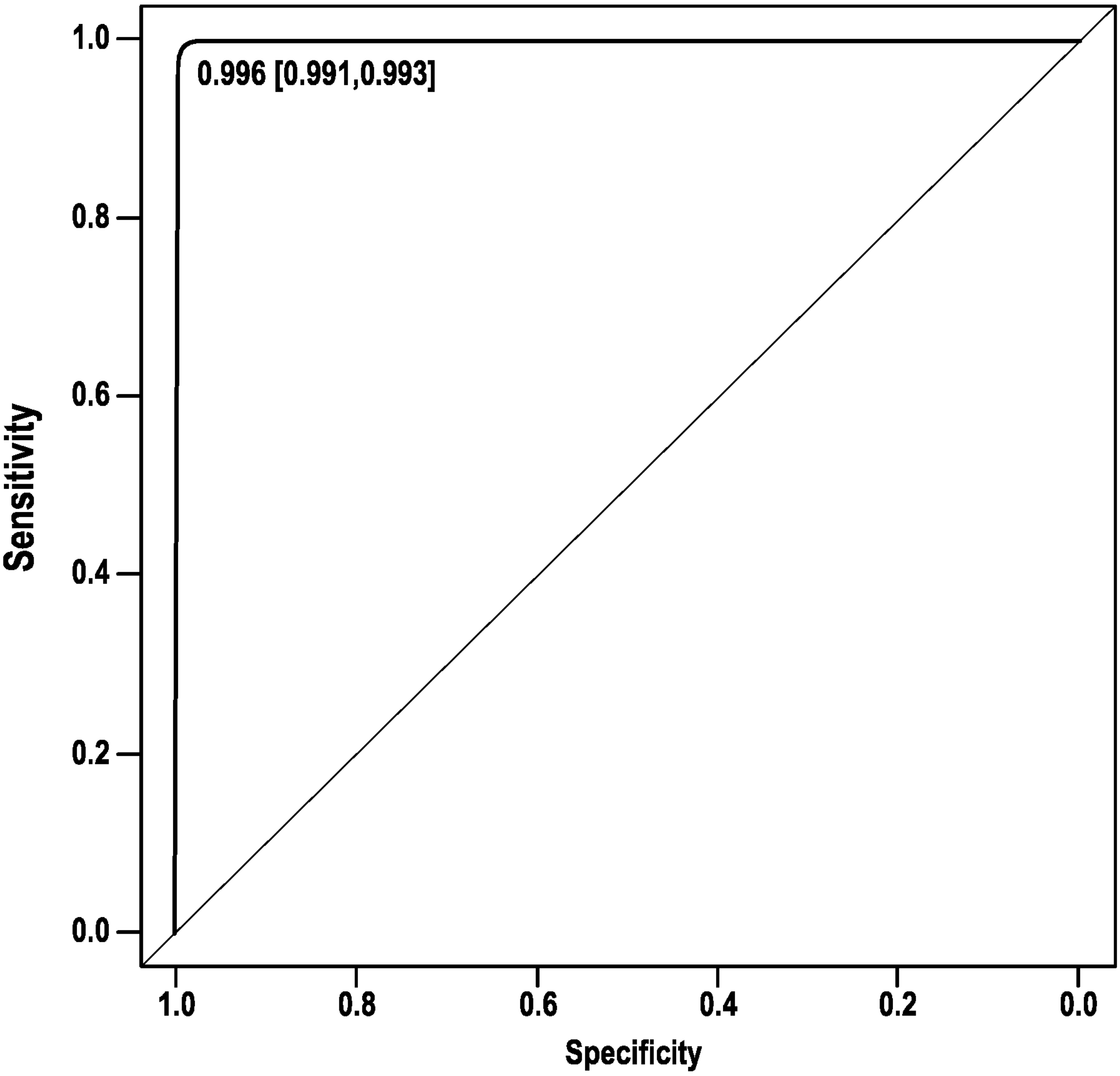
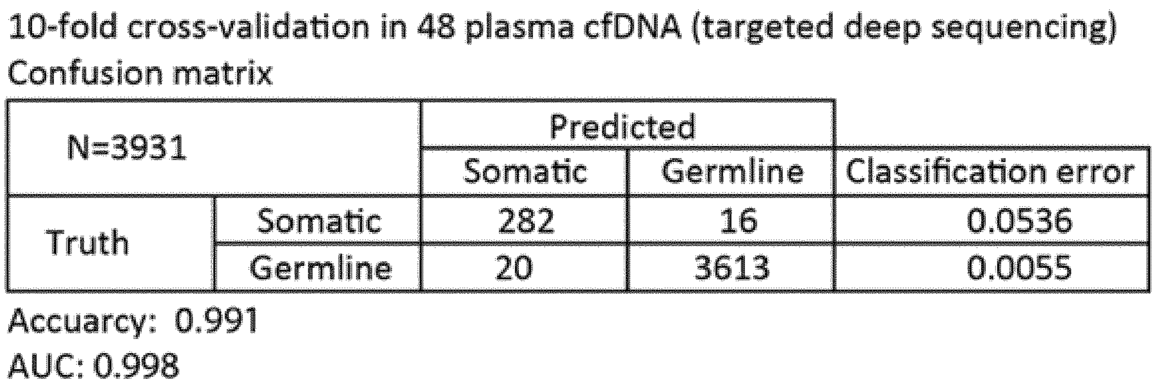

0.996 [0.991,0.993]
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 10



10-fold cross-validation in 48 plasma cfDNA (targeted deep sequencing)

Confusion matrix

N=3931		Predicted		
		Somatic	Germline	Classification error
Truth	Somatic	282	16	0.0536
	Germline	20	3613	0.0055

Accuarcy: 0.991

AUC: 0.998

FIG. 11

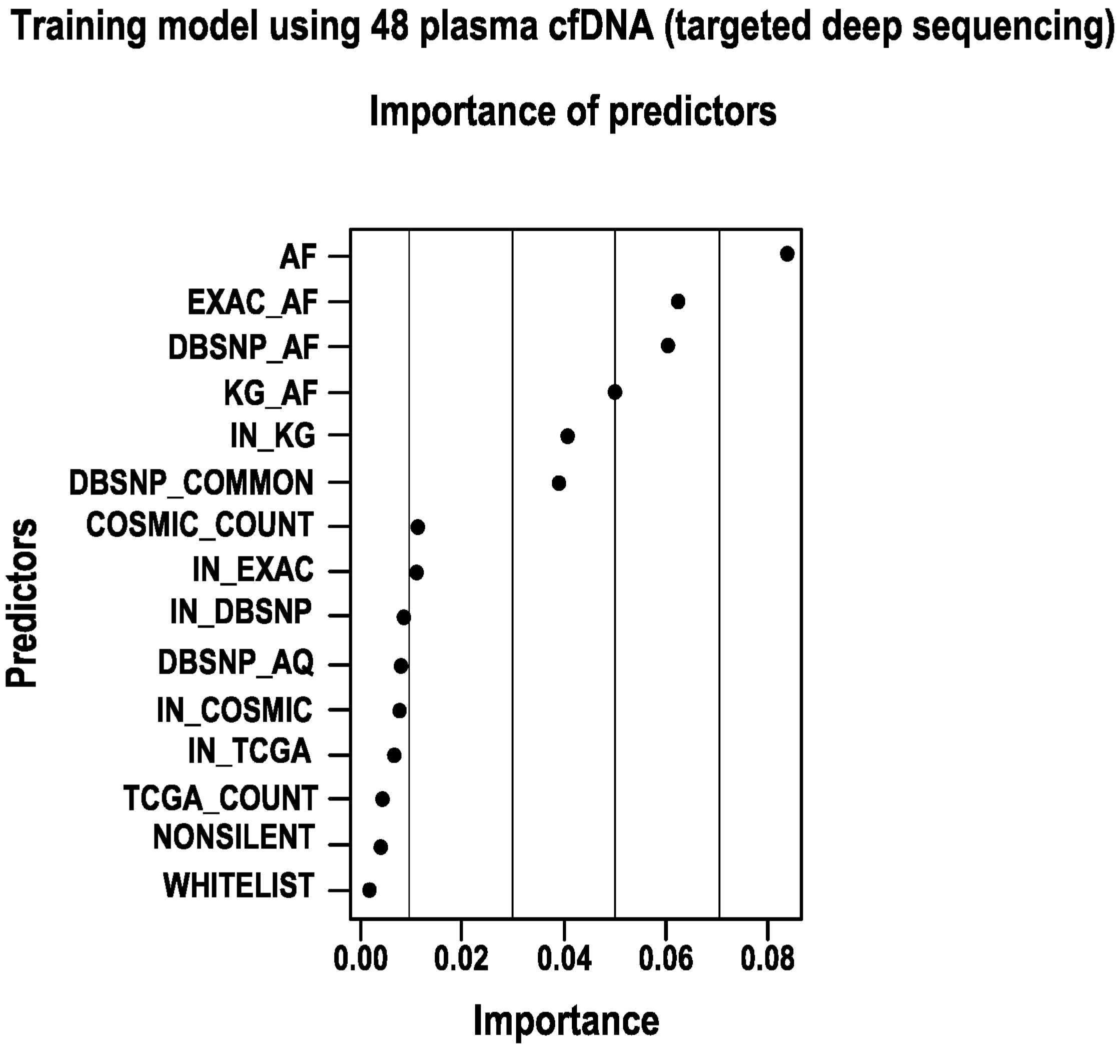

Training model using 48 plasma cfDNA (targeted deep sequencing)
Importance of predictors
AF
EXAC_AF
DBSNP_AF
KG_AF
IN_KG
DBSNP_COMMON
COSMIC_COUNT
IN_EXAC
IN_DBSNP
DBSNP_AQ
IN_COSMIC
IN_TCGA
TCGA_COUNT
NONSILENT
WHITELIST
Predictors
0.00
0.02
0.04
0.06
0.08
Importance

Fig. 12

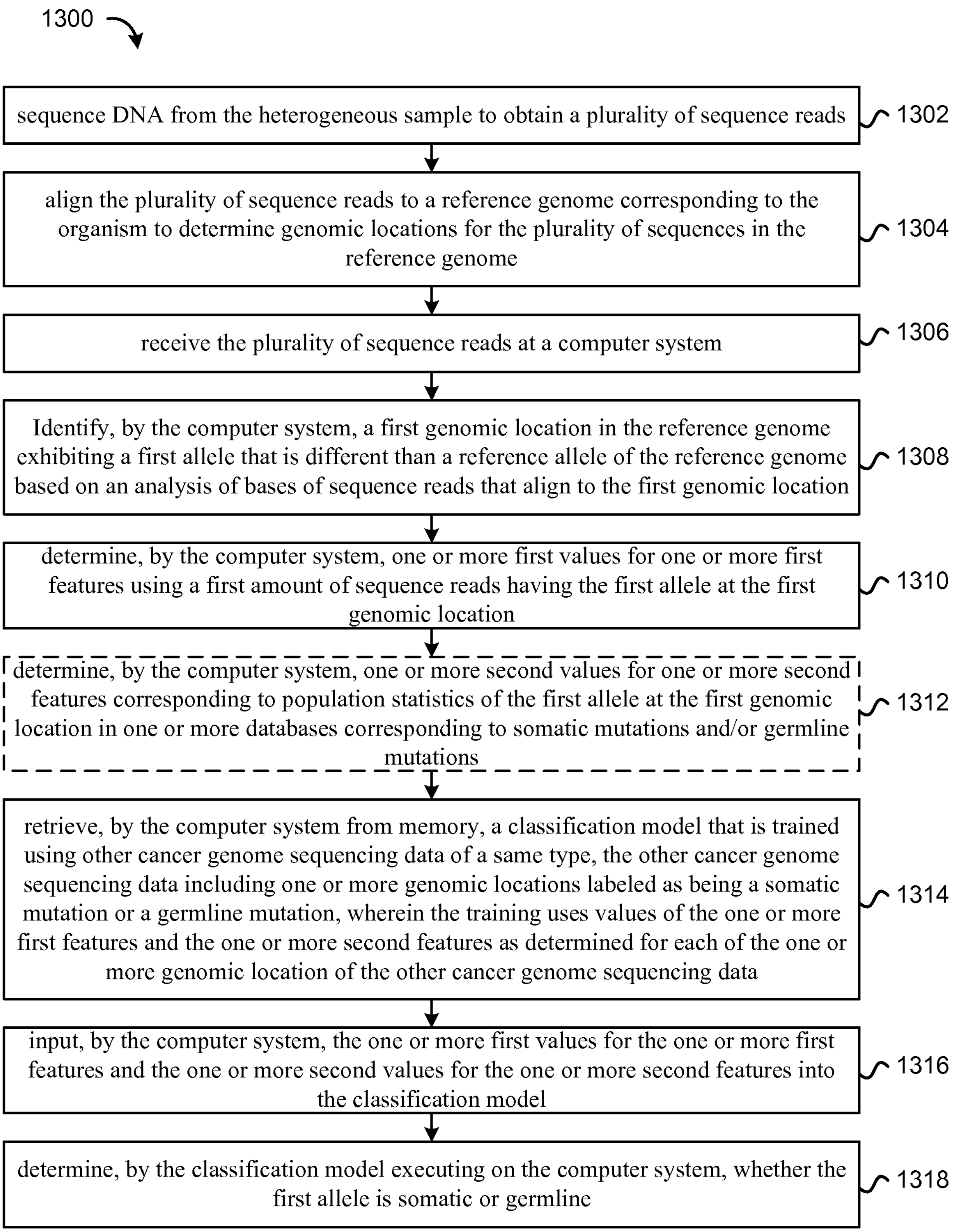

1300
sequence DNA from the heterogeneous sample to obtain a plurality of sequence reads
1302
align the plurality of sequence reads to a reference genome corresponding to the organism to determine genomic locations for the plurality of sequences in the reference genome
1304
receive the plurality of sequence reads at a computer system
1306
Identify, by the computer system, a first genomic location in the reference genome exhibiting a first allele that is different than a reference allele of the reference genome based on an analysis of bases of sequence reads that align to the first genomic location
1308
determine, by the computer system, one or more first values for one or more first features using a first amount of sequence reads having the first allele at the first genomic location
1310
determine, by the computer system, one or more second values for one or more second features corresponding to population statistics of the first allele at the first genomic location in one or more databases corresponding to somatic mutations and/or germline mutations
1312
retrieve, by the computer system from memory, a classification model that is trained using other cancer genome sequencing data of a same type, the other cancer genome sequencing data including one or more genomic locations labeled as being a somatic mutation or a germline mutation, wherein the training uses values of the one or more first features and the one or more second features as determined for each of the one or more genomic location of the other cancer genome sequencing data
1314
input, by the computer system, the one or more first values for the one or more first features and the one or more second values for the one or more second features into the classification model
1316
determine, by the classification model executing on the computer system, whether the first allele is somatic or germline
1318

FIG. 13

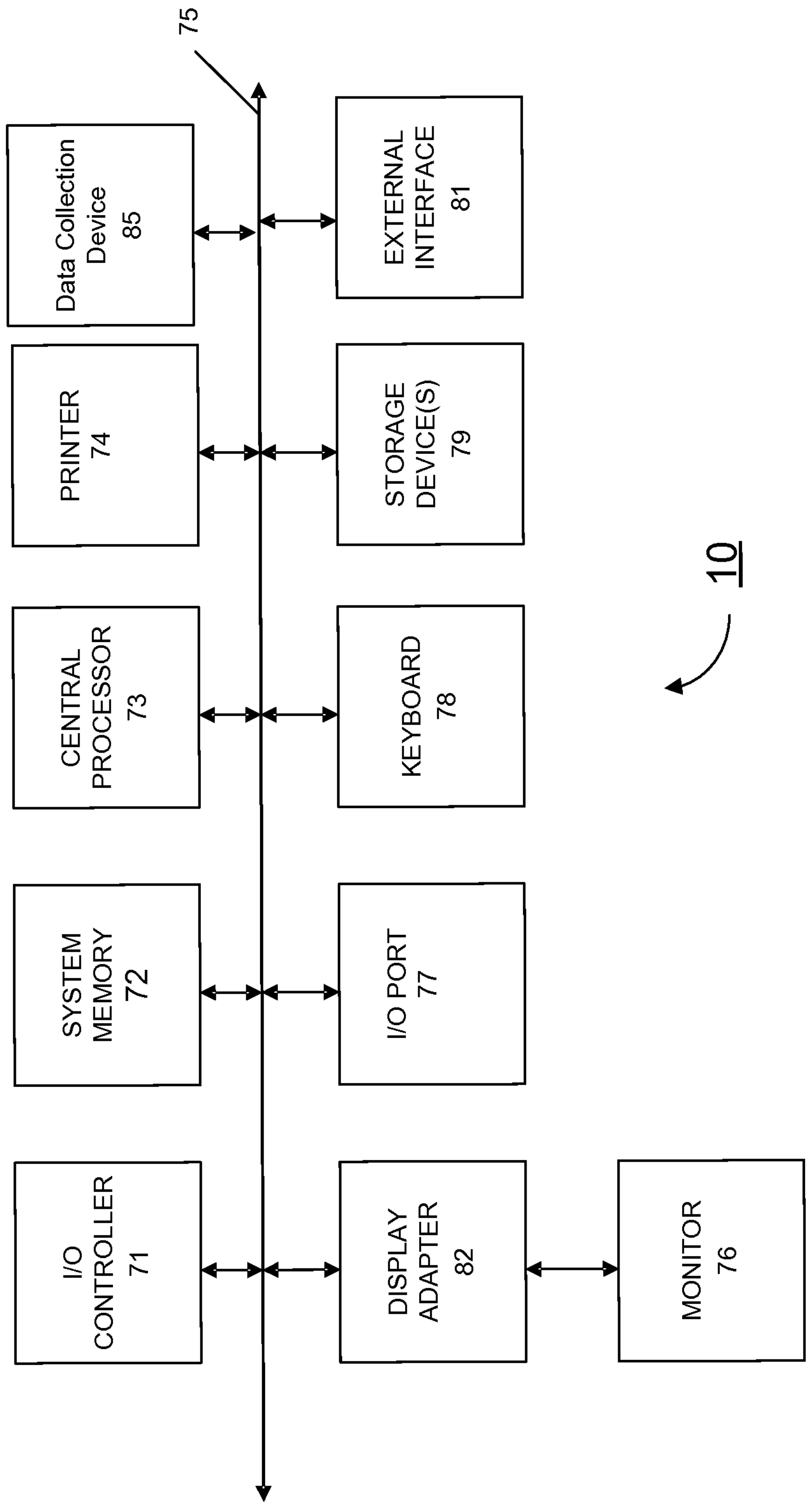

75
Data Collection Device 85
EXTERNAL INTERFACE 81
PRINTER 74
STORAGE DEVICE(S) 79
CENTRAL PROCESSOR 73
KEYBOARD 78
SYSTEM MEMORY 72
I/O PORT 77
I/O CONTROLLER 71
DISPLAY ADAPTER 82
MONITOR 76
10

FIG. 14

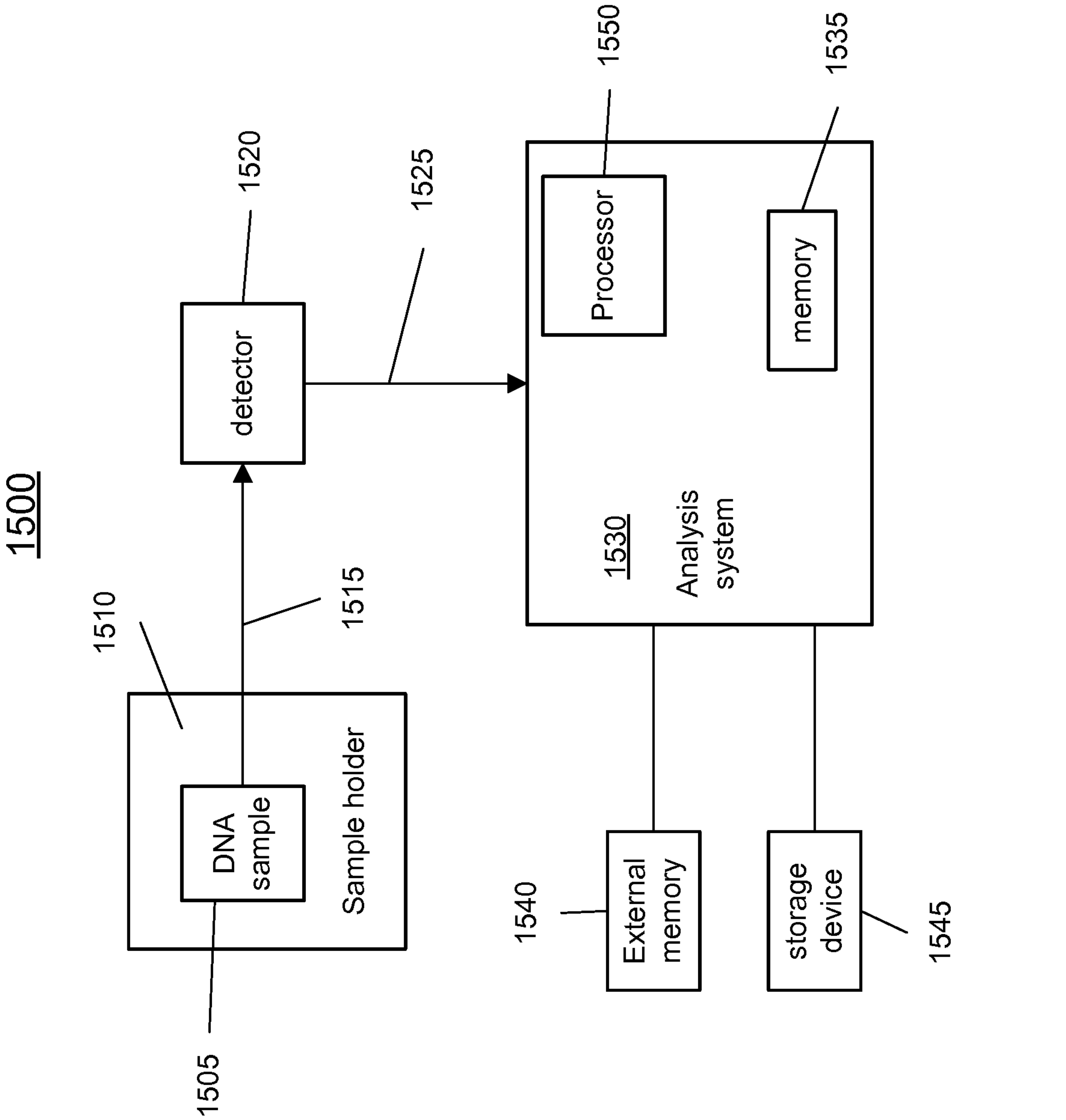

1500
1505
DNA sample
Sample holder
1510
1515
detector
1520
1525
1530
Analysis system
Processor
1550
memory
1535
External memory
1540
storage device
1545

FIG. 15

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2018/069720

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/22 C12Q1/6869
ADD.

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)
C12Q G06F

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)

EPO-Internal , WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2016/138376 Al (ASURAGEN INC) 1 September 2016 (2016-09-01) [0005] - [0008] , [0012] - [0013] , [00105] -[00116] -----	1-10
X	W0 2016/097251 Al (UNIV DANMARKS TEKNISKE [DK]) 23 June 2016 (2016-06-23) abstract; figures 1-2 p .3 1.7-32, p .11 1.5; p .26 1.23-p.28 1.5; Example 1 -----	1-10

-/- ·

☒ 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" documentwhich may throw doubts on priority claim(s) orwhich 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 12 November 2018	Date of mailing of the international search report 19/11/2018
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Authorized officer Werner, Andreas

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MARQUARD AM ET AL: "TumorTracer : a method to identify the tissue of origin from the somatic mutations of a tumor specimen", BMC MEDICAL GENOMICS, vol. 4, no. 2, 1 October 2015 (2015-10-01), page 136ra68, XP055251603, DOI: 10.1186/S12920-015-0130-0 the whole document -----	1-10
X	DING J ET AL: "Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data", BIOINFORMATICS., vol. 28, no. 2, 13 November 2011 (2011-11-13), pages 167-175, XP055257193, GB ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btr629 the whole document -----	1-10
X	KAMINKER JS ET AL: "Distinguishing cancer-associated missense mutations from common polymorphisms", CANCER RESEARCH, & 102ND ANNUAL MEETING OF THE AMERICAN-ASSOCIATION-F0R-CANCER-RESEARCH (AACR); ORLANDO, FL, USA; APRIL 02-06, 2011, vol. 67, no. 2, 1 January 2007 (2007-01-01), pages 465-473, XP002484152, ISSN: 0008-5472, DOI: 10.1158/0008-5472.CAN-06-1736 the whole document -----	1-10
X, P	WO 2018/064547 AI (UNIV COLUMBIA [US]) 5 April 2018 (2018-04-05) [005] - [0015], [0064]; claim 1; figure 1; example 1 -----	1-10
X, P	WO 2018/009723 AI (GUARDANT HEALTH INC [US]) 11 January 2018 (2018-01-11) [0005] - [0015], [00195], [00283] -----	1-10
X, P	WO 2017/139492 AI (T0MA BIOSCIENCES INC [US]; DE LA VEGA FRANCISCO M [US]) 17 August 2017 (2017-08-17) [0009], [0031], [0050]; figure 1; example 2 -----	1-10
E	WO 2018/144782 AI (THE TRANSLATIONAL GENOMICS RES INSTITUTE [US]) 9 August 2018 (2018-08-09) p.5 1.13-P. 7 1.23, p.18 1.6-p.22 1.2 -----	1-10
	-/--	

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2018/069720

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2017/061072 Al (KERMANI BAHRAM GHAFFARZADEH [US] ET AL) 2 March 2017 (2017-03-02) [0016] , [0018] , [0024] , [0042] ; claims 1-3 -----	1-10
A	US 2015/178445 Al (CIBULSKIS KRISTIAN [US] ET AL) 25 June 2015 (2015-06-25) the whole document -----	1-10
A	CIBULSKIS K ET AL: "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples" , NATURE BIOTECHNOLOGY, vol. 31, no. 3, 10 February 2013 (2013-02-10) , pages 213-219 , XP055256219 , ISSN: 1087-0156, D0I: 10.1038/nbt.2514 abstract; figure 1 -----	1-10
A	SAUNDERS CT ET AL: "Strelka: accurate somatic smal1-variant calling from sequenced tumor-normal sample pairs" , BI0INF0RMATICS. , vol. 28, no. 14, 10 May 2012 (2012-05-10) , pages 1811-1817 , XP055257165 , GB ISSN: 1367-4803 , D0I: 10.1093/bioinformatics/bts271 the whole document -----	1-10
A	ROTH A ET AL: "JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data" , BI0INF0RMATICS. , vol. 28, no. 7, 27 January 2012 (2012-01-27) , pages 907-913 , XP055257196, GB ISSN: 1367-4803 , D0I: 10.1093/bioinformatics/bts053 the whole document -----	1-10

-/--

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KOBOLDT DC ET AL: "Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection : Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection" In: "Current Protocols in Bioinformatics", 12 December 2013 (2013-12-12), John Wiley & Sons, Inc., Hoboken, NJ, USA, XP055522254, ISBN: 978-0-471-25095-1 pages 15.4.1-15.4.17, DOI: 10.1002/0471250953.bil504s44, the whole document -----	1-10
A	US 2014/143188 Al (MACKEY AARON J [US] ET AL) 22 May 2014 (2014-05-22) Using sequencing data from tumor and normal pair from a singleUsing sequencing data from tumor and normal pair from a single patient, patient, -----	1-10

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2018/069720

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2016138376 A1	01-09-2016	AU 2016222569 A1	07-09-2017
		CA 2977787 A1	01-09-2016
		CN 107614697 A	19-01-2018
		EP 3262197 A1	03-01-2018
		US 2018163261 A1	14-06-2018
		WO 2016138376 A1	01-09-2016
WO 2016097251 A1	23-06-2016	US 2017342500 A1	30-11-2017
		WO 2016097251 A1	23-06-2016
WO 2018064547 A1	05-04-2018	NONE	
WO 2018009723 A1	11-01-2018	NONE	
WO 2017139492 A1	17-08-2017	NONE	
WO 2018144782 A1	09-08-2018	NONE	
US 2017061072 A1	02-03-2017	NONE	
US 2015178445 A1	25-06-2015	EP 2891099 A1	08-07-2015
		US 2015178445 A1	25-06-2015
		WO 2014036167 A1	06-03-2014
US 2014143188 A1	22-05-2014	NONE	